Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-023-01975-6 OPEN

Systematic correspondence in co-evolving languages

Junru Wu [1]✉ & Junyuan Zhao [2]

Language co-evolution is an influential cultural force, impacting the past, present, and future of human languages. Systematic correspondence identifies corresponding features in languages evolving together, such as English "d" and German "t" in word pairs like “deed–Tat” and “deep–tief”. This study examines how social ecology influences lexical-phonological systematic correspondence using a vector-based measurement—*weighted cosine systematicity*—across two co-evolutionary lexical datasets for comparison: old to recent English-German related words, and thirty-year sliced morphemic transcriptions for Chinese dialects in Shanghai. Results show that even when related but socially independent languages evolve in different directions, they can maintain an equilibrium in systematic correspondence over centuries. In contrast, dialects can rapidly converge towards their national high variety in terms of lexical-phonological similarities, and the regional standard in terms of systematic correspondence within decades. This suggests that self-regulation of cross-linguistic systematic correspondence has its own, yet complementary, mechanism compared to the similarity-based co-evolutionary mechanism, making it a meaningful indicator and predictor for cross-linguistic lexical co-evolution.

[1] Lab of Language Cognition and Evolution, Department of Chinese Language and Literature, East China Normal University, Shanghai, China. [2] Department of Linguistics, University of Michigan, Ann Arbor, MI, USA. ✉email: jrwu@zhwx.ecnu.edu.cn

Introduction

Languages evolve in a way similar to species (Mufwene, 2001). This idea has been applied in modelling language evolution, using languages as species, words as genes, and translation equivalents as alleles of a common genetic loci. This approach, known as phylogenetic linguistics, has provided useful insights into the history of language evolution (Bowern, 2018; Zhang et al., 2019; Sagart et al., 2019).

Nevertheless, it relies on two assumptions: (1) one human individual holds one lexical form (allele) for one concept (loci), and (2) interbreeding is necessary for cross-linguistic lexical transmission. These assumptions may be valid when studying core-words (Swadesh, 1955) that are etymological cognates (Zhang et al., 2019; Sagart et al., 2019). However, to fully understand language evolution, we must consider co-evolution.

Derived from ecology, the term "co-evolution" describes the phenomenon of closely associated species influencing each other and resulting in reciprocal changes (Thompson and Rafferty, 2020). In the context of language evolution, here we adopt this term to denote the interactions between different linguistic varieties that mutually influence and bring about reciprocal changes.

As human languages evolve, they inherently influence one another due to how humans cognitively process language. This is evident in people's ability to possess multiple lexical forms from different languages for the same meaning (Kroll and Sholl, 1992; Dijkstra and Van Heuven, 1998), as well as their capacity to learn and borrow words and sounds, which allows for instantaneous transmission across related linguistic varieties (Wu et al., 2021). Consequently, words and pronunciations are not only passed down through generations, but also regularly pass over between different languages, like the spread of parasitic features (Mufwene, 2001). Thus, the change in lexical alignment between co-evolving languages likely involves mechanisms distinct from those involved in intergenerational transmission of human genomes.

In terms of cross-linguistic lexical alignment, in addition to the widely acknowledged similarity-based mechanism of language contact (Thomason, 2011), such as English "comput**e**r" being borrowed into German as "comput**e**r", a well-known evolutionary linguistic phenomenon is that across many languages there is a systematic correspondence between semantically related words and their phonemes (Dyen, 1963, p. 634; Meillet and Ford, 1967; Schleicher, 1967). For instance, English "d" typically corresponds with German "t", as demonstrated by word pairs such as "deed–Tat", "deep–tief", etc. (Grimm, 1967; Verner, 1967) The systematic correspondences between two languages can be inherited from a common ancestral source, but can also be constructed through various processes of language contact, such as lexical borrowing (Jacobson, 1971; Poplack et al., 1988; Thomason, 2011) and analogical spread of sounds (e.g., uvular rhotics across Europe, Trudgill, 1974; and superimposed sound changes in Chinese dialects, Wang, 2010). Systematic correspondence has been studied extensively and has been used to uncover the past connections between languages (Beekes and Vaan, 2011). Despite its vital importance, it remains unclear which general mechanisms are regulating the shifting relationship of systematic correspondence between co-evolving languages, especially when lexical borrowing and sound spreading are also taken into consideration. To our knowledge, few studies have made explicit predictions regarding this topic, except for Dixon's *Punctuated Equilibrium Model* (Dixon, 1997), which posits that co-evolving languages tend to converge on a prototype, until the split of peoples interrupts this process. However, it remains unclear how this mechanism applies to systematic correspondence and whether the changes in systematic correspondence are regulated by other linguistic ecological subtleties known to influence the orderly heterogeneous evolution of various linguistic varieties (Labov, 1963) and creole evolution (Mufwene, 2001).

Given the existing understanding of language evolution, three potential theories may be proposed to explain the change of systematic correspondence that occur across co-evolving languages. (1) The theory of **attrition** predicts that due to the overlaying of phonotactic constraints in neogrammarian sound changes (e.g., Grimm, 1967; Verner, 1967), the residues of lexical diffusion (Wang, 1969), and the accumulation of other exceptions (Mazaudon and Lowe, 1993), the vocabularies of co-evolving languages will drift apart and become less systematically aligned. (2) Instead, a generalisation of Martinet's (1952) **integration theory** may suggest that, due to similar external pressure, continuous mutual influences, analogical sound changes (Anttila, 1977), as well as the need to reduce the cognitive cost for maintaining two vocabularies in one mind (Bialystok, 2009), systematicity between corresponding vocabularies would increase over time. (3) Alternatively, the theory of **self-regulated adaptation** may propose that, the relationship between co-evolving languages are *restructured* (Mufwene, 2001) to adapt to the changing linguistic ecology. In some cases, a loss of systematicity in one aspect would be compensated in another aspect (Labov, 1994).

Moreover, since similarity-based phonological influences have been widely accepted as a key factor in language contact, particularly in the process of lexical borrowing (Weinreich, 1953; Poplack et al., 1988), it is necessary to ensure that mechanisms based on systematic correspondences are not simply a result of similarity-based influences in language contact.

Please note that systematic correspondence and cross-linguistic similarity in the pronunciation of related words are related concepts but have distinct meanings. To illustrate this, let's suppose we have language A and language B. In language A, words A1, A2, A3, A4… belong to the same lexical tone class and all have rising tones. In language B, the translation equivalents of these words, B1, B2, B3, B4… also belong to the same tone class, but with falling tones. In this case, we can identify a tonal systematic correspondence "rule" between language A and language B, where A1… and B1… are considered tonally corresponding words. However, it's important to note that despite their correspondence, A1… and B1… have different tonal contours—one is rising while the other is falling, indicating that they are not similar.

In the current body of international literature on language studies, there is still a lack of clarity regarding the general mechanisms that govern the fluctuating dynamics of systematic correspondence between co-evolving languages, particularly when the influence of lexical borrowing and sound spreading is taken into consideration. As a novel contribution, this paper aims to elucidate and compare these mechanisms to provide a deeper understanding of the intricate relationship between co-evolving languages. This study suggests a vector-based approach to measure systematic correspondence and evaluates the related theories using two co-evolutionary lexical datasets.

The two datasets were chosen to encompass distinct scenarios of language co-evolution. One dataset focuses on the interplay between two related national languages of equal social status, while the other dataset examines the co-evolution of non-literal local sub-dialects alongside a regional high variety and a national high variety. As mentioned earlier, the theory of attrition suggests that in both datasets, there will likely be a decrease in lexical systematic correspondence over time. Conversely, the theory of integration implies that systematic correspondence may increase in both cases. However, with the incorporation of the theory of self-regulated adaptation and Dixon's Punctuated Equilibrium

Model, contrasting patterns may emerge. Specifically, the former dataset, characterised by a split population, is expected to display an absence of consistent directions in the evolution of sound systems during changes in systematic correspondence. On the other hand, in the latter dataset, where two distinct prototypes are identifiable within the intermixed population, lexical pronunciations of the sub-dialects are expected to converge towards these prototypes. Nevertheless, existing research does not offer explicit predictions regarding the varying influences between a more distant national standard and a more similar regional high variety.

Considering the availability of data, we have selected a range of English-German related words spanning from Old English and Old High German to recent English and German pronunciations, to represent the former scenario. Furthermore, to depict the latter scenario, we have selected thirty-year sliced morphemic transcriptions documenting the Chinese dialects spoken in Shanghai. Here we offer comprehensive backgrounds for both lines of study. For more pragmatic data processing details, please consult the "methods" section.

Old to recent English-German related words. English and German, originally closely related West Germanic languages, have a shared history that is characterised by geographic and political separation, as well as complex social interruptions (Hickey, 2012).

Their common roots can be traced back to Proto-Germanic, which in turn can be linked to Proto-Indo-European. Evidence supporting this connection can be found in the systematic correspondence of cognate pronunciations, as observed by the 19th-century historical linguists (e.g., Grimm, 1967; Schleicher, 1967). This relation yield related word pairs such as "deaf-toub" in Old English and Old High German, which became "deaf-taub" in Modern English and German.

Additionally, both languages have been influenced by a higher Church Latin superstratum throughout the Middle Ages. This yield related word pairs such as "tiwesdæg-ziestag" in Old English and Old High German, which became "Tuesday-Dienstag" in Modern English and German.

These etymological factors contribute valuable data to our current investigation. Nevertheless, following the settlement of Germanic tribes such as the Angles, Saxons, and Jutes in the British Isles, who spoke Old English (which included a relatively standardised form in the late 9th century), there are few records indicating regular visits between them and their mainland relatives. On the mainland, German dialects, including documented Old High German, developed alongside some direct contact with Old French. In contrast, the linguistic ecology in the British Isles was shaped by the influences of Viking invasions in the 9th and 10th centuries, as well as the Norman Conquest since 1066. These influential events led to the transformation of Old English into Middle English. Throughout the Middle Ages, English and German likely coexisted and exerted some influence on each other through trade and cultural (primarily missionary) interactions (Gneuss, 1990; Hayden, 2017). However, the traces left on the vocabularies appear to be primarily linked to their shared influence from (Old) French. This can be seen in word pairs like "check-Scheck" and "accent-Akzent" in Modern English and German.

Closer contact between English and German likely resumed in the late Middle Ages, thanks to various factors such as the Age of Discovery, Mercantilism, Protestant Reformation, the introduction of Germanic clergies (although predominantly Dutch-speaking), and later the Enlightenment Movement. These developments played a significant role in the non-Latin national literacy advancements in both regions and ultimately formed the foundation of Modern English and Modern German (Machan, 2012; Hayden, 2017). This period left its mark on the vocabularies of both languages, as seen in word pairs like "coffee-Kaffee" and "smuggle-schmuggeln". However, specific statistics regarding individual proficiency in both languages during this period remain uncertain.

After World War II, there was a notable increase in direct influence between the vocabularies of English and German. However, English likely has a greater impact on the German-native speakers, who ranked among the top 10 in English proficiency (EF Education First, 2022), while the influence of German on English remains restricted.

Despite the extensive co-evolutionary history of British English and German vocabularies, it is crucial to emphasise that these two languages have always been spoken by distinct populations under separate regimes, each maintaining its linguistic standards. Their relatedness does not imply one language serving as a standard for the other. Therefore, the English-German dataset serves as a representative example illustrating the interaction between two related national languages of equal social status.

To the best of our knowledge, no direct statistical study has been conducted to analyse whether the previous systematic correspondence between the two vocabularies has diminished, strengthened, or undergone more complex fluctuations during the evolution of the two languages from their older versions to their modern forms, including their recent pronunciations. By contrasting this scenario with the upcoming dialectal case to be presented in the subsequent section, we can gain valuable insights for evaluating the three hypotheses that we reviewed earlier.

Related morphemes in the Chinese dialects spoken in Shanghai. The ecology of Chinese dialects exhibits certain similarities to that of European languages. The linguistic varieties within each language family can be traced back to a common ancestor, possess etymologically related vocabularies, and demonstrate a notable variation in mutual intelligibility (Tang and Van Heuven, 2009; Gooskens et al., 2018). However, throughout the shared history of Chinese dialects, two distinctive features have consistently endured, making them particularly intriguing for the current research.

First, Chinese regional dialects have almost always coexisted with a national standard (Confucius, 551BC–479BC), actively encouraged and supported by the ancient Chinese empire (GUO Pu, 276–324a; 276–324b). This multi-dialectal ecology has been compared to the diglossia observed in medieval Europe (Ferguson, 1959), where a privileged few used classical Latin alongside the vernacular languages, while the majority remained monolingual. However, evidence from Missionary documents reveal that by the 16th century, there was already widespread individual bi-dialectism involving the national standard, even among the least educated Chinese populace in certain regions (Ricci, 1552–1610). This extensive and endurant influence of the national standard has left a lasting impact. Modern dialectology studies often uncover historical superstratum of standard Chinese integrated in the lexical phonology of Chinese dialects (e.g., Wang, 2010).

The second notable characteristic of the Chinese dialectal ecology is the use of shared ideographic characters (known as "Zi") to represent related monosyllabic morphemes across dialects. Traditional Chinese rhyming books, such as those for the national standard (e.g., see Pulleyblank, 1998) and regional dialects (e.g., Li, 2019), organise their entries based on these related morphemes. Furthermore, throughout history, there are documentations showing both literate and non-literate Chinese

speakers discussing the different pronunciations of certain characters across Chinese dialects (Wang, 2023), demonstrating their understanding of the cross-dialectal relationship between these linguistic units.

These two features make the co-evolution of Chinese dialects an ideal testing ground for investigating the scenario of linguistic varieties co-evolving with identifiable common prototypes and a sizeable and stable bi-dialectal population.

These features likely apply to the language ecology in Shanghai, a thriving migration city situated in the prosperous Lower Yangtze plains. The majority of urban Shanghainese (a Wu dialect) speakers are proficient in Standard Chinese (Putonghua, a standardised common speech with pronunciation based on the Beijing dialect, according to Standing Committee of the National People's Congress, 2000). While there is no direct data on mutual intelligibility between the two, it has been established that a closely related dialect of Shanghainese, Suzhou Wu Chinese, is challenging for monolectal Beijing Mandarin speakers to comprehend (only 26% intelligibility in isolated words, according to Tang and Van Heuven, 2009). Shanghainese and Standard Chinese only partially overlap in terms of sound inventory and phonotactics. While most translation equivalents between them have etymological connections, some do not. Additionally, the range of similarities among their related morphemes can be extensive.

Chinese researchers have studied the sound changes in central urban Shanghainese over the past 160 years (Qian, 2003; Chen, 2019). They observed the influence of national standard on the reorganisation of phonological categories in Shanghainese. For example, characters originally pronounced with /ʐ/ initials (Edkins, 1853; Zhao, 1956) began to split into /ʐ/ and /dʑ/ categories in the 1960s (Jiangsu, 1960; Xu and Tang, 1988), depending on their pronunciations in Standard Chinese (Chen, 2019). This suggests a strengthening systematic correspondence between Shanghainese and Standard Chinese, which appears to support the integration hypothesis.

Although Standard Chinese and urban Shanghainese are the predominant dialects in Shanghai, there are also distinct sub-dialects within the city (You, 2013). These sub-dialects were inherited from historical prefectures like Suzhou-Fu, Songjiang-Fu, and Jiaxing-Fu. It remains unclear whether these sub-dialects, coexisting with the two dominant varieties, follow the same evolutionary trajectory as the central urban variety. Further investigation is necessary to clarify the extent to which the integration hypothesis applies and the influence of social factors on the collective changes of these sub-dialects.

Furthermore, previous studies have reported similarity-based influences of Standard Chinese observed in urban Shanghainese. For example, the pronunciation of 全 (whole) in Shanghainese has shifted from /ʐi/ (Edkins, 1853; Zhao, 1956) to /dʑyøn/ (Jiangsu, 1960; Xu and Tang, 1988), resembling the Standard Chinese pronunciation /tɕhyæn/. Then with more characters undergoing similar shifts (Chen, 2019), a new mapping rule formed between the two systems and hence influence the relation of correspondence.

Despite the existing research, it is necessary to go beyond specific examples and explore general mechanisms that can explain the two divergent scenarios and shed light on the relationship between correspondence-based and similarity-based mechanisms.

Quantified analyses on systematic correspondence. To investigate systematic correspondence in a quantitative manner that allows for statistical modelling and hypothesis testing, researchers need to define two types of units: the unit of data entries and the unit of systematic consideration. Previously, linguists commonly used mono-morphemic words, e.g., Latin "pater"~Gothic "far-an(ire)" as used to support Grim's Law (Grimm, 1967), or single morphemes, e.g., Chinese monosyllabic morphemes as used in the Chinese dialectological studies, as data entries, while levels of phonological units like consonants (e.g., Grimm, 1967; Chen, 1973) and vowels (Jespersen, 1909), as well as initials, rhymes, and tones (Karlgren, 1915–1916) were considered systematically. This prevailing method involved listing and enumerating word or morpheme pairs that correspond at a specified phonological level, following practices used by scholars like Grim, who was born in 1785. Grim enumerated Latin-Gothic examples to verify connections between phonological units (Grimm, 1967). Recent studies have continued this approach.

However, this approach does not account for the possibility that observed correspondence may occur by chance, especially when analysing small sound inventories like lexical tones. It may also overlook correspondences between small categories with limited data entries. To address similar limitations, Baxter (1992) applied Bayesian statistics to examine the rhyme categories of Old Chinese by studying the rhyming relationships between characters found in ancient Chinese poetry collections. This probabilistic strategy mitigates the impact of chance occurrences and addresses data scarcity in smaller categories.

By adopting a similar approach, we can use the Chi-square test to investigate systematic correspondence between sound inventories of languages A and B. The Chi-square test helps examine the association between categorical variables like consonant inventories. Additional Chi-square tests were conducted for this study, and the reports can be found in the supplementary information file.

While the Chi-square test typically supports the existence of systematic correspondence by rejecting the null hypothesis, it's valuable to measure correspondence on a more nuanced scale. This involves analysing data at a category-by-category or word-by-word level, providing a comprehensive understanding of the strength and patterns of correspondence between phonological units that may not be captured by the Chi-square test alone.

Previous studies have made limited attempts to address the research question at hand. One such attempt utilised Linear Mixed Effect Modelling (LME) (Bates et al., 2013) to explore the relationship between distance matrices representing tonal categories in Standard Chinese and the Euclidean distances between pitch contours in a northern Mandarin Chinese dialect. This study found that later birth years are related to increasing correlation between the matrices, and further uncovered the influences from cross-dialectal similarities in pronunciation, as well as participants' literacy education and auditory working memory (Wu et al., 2016). These results suggest alignment with the theory of self-regulated adaptation. However, the use of distance matrices as both independent and dependent variables introduced autocorrelation issues, potentially impacting the reliability of the models. Additionally, we recently became aware of an unpublished work by non-professional language enthusiasts who attempted to quantify phonological correspondence by multiplying data entries proportions for each mapping (Gu, 2023).

These endeavours reflect initial efforts to associate linguistic variation and microscopic language evolution with human lexical processing. However, all of these statistical approaches rely on the assumption that lexical retrieval events are independent, analogous to drawing coloured balls from a box (Baxter, 1992). Yet, in reality, human lexical processing involves complex simultaneous activation and competition within bilingual lexicons (Dijkstra and Van Heuven, 1998). Moreover, these studies did not account for the potential cognitive competition between mapping

rules and the weight of each sound category within the overall sound inventory. Therefore, it may be more appropriate to envision the lexical processing of related linguistic varieties as drawing magnetic balls from a complex magnetic field. Hence, careful consideration of the competition between data entries and mapping rules is crucial when exploring the co-evolutionary mechanisms of related vocabularies. Furthermore, the comparability of measurements across different phonological units warrants further investigation.

Methods

The current measurement: weighted cosine systematicity. Here, we propose a vector-based measurement called *Weighted Cosine Systematicity* (*sys_cos_w*) to quantify systematic correspondence. This method represents sound inventories as multi-dimensional vectors, with each dimension corresponding to a specific sound category's number of data entries. By employing this approach, we can effectively capture the competition among sound categories in the relationship between the two vocabularies. Moreover, the value assigned to each dimension reflects the quantity of data entries (such as morphemes) involved in this competition.

The weighted cosine systematicity accounts for three crucial factors. Firstly, a higher number of categories and/or data entries in competition with the mapping relationship indicates weaker systematic correspondence. Secondly, as the mapping under consideration involves a greater number of data entries, it implies stronger systematic correspondence. Additionally, the measurement enables the evaluation of the relative importance of the two units directly involved in the mapping by analysing their respective occurrence proportions within the total number of word pairs. This assessment provides insights into the significance of these units within their respective vocabularies.

As demonstrated in Fig. 1, for two linguistic varieties A and B, the systematic correspondence between unit *a* (from A) and unit *b* (from B), *a*~*b*, is divided into two mapping relations: *a*→*b* (indexed as "ab") and *b*→*a* (indexed as "ba"). The *mapping a*→*b* shows the sufficiency of *a* for *b* and the necessity of *b* for *a*, which depends on the alternatives of *b* given *a*, and vice versa for the mapping *b*→*a*. Therefore, sys_cos_{ab} *and* sys_cos_{ba} are each mathematically defined as the cosine similarity between a target vector $v = (0, \ldots, npair_{\text{this}}, \ldots, 0_{\text{m}})$ and a reference vector $vref = (napair_1, \ldots npair_m)$. Then sys_cos_{ab} and sys_cos_{ba} are weighted with the proportions of *a* and *b* occurrences within the total number of word pairs, $weight_a$ and $weight_b$, respectively, resulting in $sys\_cos\_w_{ab}$ *and* $sys\_cos\_w_{ba}$. Consequently, systematic distances sys_dist_{ab} and sys_dist_{ba} can be calculated by subtracting $sys\_cos\_w_{ab}$ and $sys\_cos\_w_{ba}$ from one.

Figure 1 is based on a dual-lexical system with 3079 Etymologically Related Translation Equivalent (ETE) pairs ($n = 3079$). Regarding the *mapping* between the units *a* (the rhyme /uai/ in the linguistic variety *A*) and *b* (the rhyme/u_A/ in the linguistic variety *B*), there are eleven ETE word pairs ($npairthis = 11$) involved.

(1) The unit *a* is involved in nineteen ETE word pairs ($npair_a = 19$), mapped to four categories in *B* ($m = 4$, which involves one (/uai/~/$_A$/), two (/uai/~/$_E$/), eleven (/uai/~/u_A/), and five (/uai/~/u_E/) ETE word pairs, respectively. Accordingly, for the *mapping a*→*b*, the reference vector *vref* is (1, 2, 11, 5), the target vector *v* is (0, 0, 11, 0), and the cosine similarity calculated between *vref* and *v* is taken as the systematicity quantified for *mapping a*→*b* ($sys\_cos_{ab} = 0.895$). Furthermore, by calculating the proportion of word pairs involved with unit *a* in the whole system, $weight_a = npair_a/n = 0.006$, the importance of *mapping a*→*b* can be weighted ($sys\_cos\_w_{ab} = weight_a \times sys\_cos_{ab} = 0.005$). Then the systematic distance can be calculated by subtracting the weighted systematicity from one ($sys\_dist_{ab} = 1 - sys\_cos\_w_{ab} = 0.995$).

(2) Similarly, the unit b is involved in twenty-two ETE word pairs ($npair_b = 22$), mapped to two categories in A ($m = 2$), which each involves eleven (/uai/~/u_A/ and /u_A/~/u_A/) ETE word pairs. Accordingly, for the *mapping b*→*a* (or *mapping a*←*b*), the reference vector *vref* is (11, 11), the target vector *v* is (11, 0), and the cosine similarity calculated between *vref* and *v* ($sys\_cos_{ba} = 0.707$) is taken as the systematicity quantified for *mapping b*→*a*. Also, by calculating the proportion of word pairs involving unit *b* in the whole system, $weight_b = npair_b/n = 0.007$, the importance of mapping *b*→*a* can be further weighted ($sys\_cos\_w_{ba} = weight_b \times sys\_cos_{ba} = 0.005$). Then the systematic distance can be calculated by subtracting the weighted systematicity from one ($sys\_dist_{ba} = 1 - sys\_cos\_w_{ba} = 0.995$).

Note that, different from earlier practices using crosstabs and/or Chi-square statistics (e.g., Chen, 1973), the current method provides a more fine-grained and integrated measurement: (1) an exact value is calculated for each specific ETE pair; (2) both phonemic frequency and interlingual lexical neighbourhood are incorporated, as the weight and the reference vector respectively; (3) comparisons can be further made across different linguistic levels, e.g., later we can see in the Shanghai dataset that phonemes are at higher order of magnitudes of *sys_cos* and *sys_dist* as compared to syllables.

In comparison to the Chi-square approach, our method avoids using a straw-man null hypothesis. Unlike the approach based on two distance matrices, our method directly measures systematic correspondence and does not rely on the secondary interpretation of complex statistical models. While our approach overlaps with the proportion-based approach in certain aspects, it further integrates the consideration of the number of competing mapping relations and the significance of phonemic units. In summary, the weighted cosine systematicity offers a comprehensive approach by considering the competition between sound categories, the quantity of involved data entries, and the importance of the units in the mapping relationship.

Datasets. As introduced earlier, two datasets are used in this work to represent the two scenarios of language co-evolution: (1) the interplay between two related national languages and (2) multi-dialectal interactions with two strata of high varieties. The following datasets are then chosen given the consideration of available data.

(1) The dataset *old to recent English-German related words* (Flippo, 2018; Kroch, 2020; Wiktionary, 2022) includes 1913 sets of related lexical forms from Old English (<1100 AC), Old High German (<1100 AC), Modern English (<1700 AC, but reflecting earlier pronunciations < 1400), and Modern German (<1700 AC), as well as the recent pronunciations for the Modern English and Modern German lexical forms (annotated according to Cambridge Dictionary; Collins Dictionary; Conjugator; Harper, 2022; Linas, 2022; Verbix, 2022) and additionally more recently related English and German words (e.g., /kæŋgəˈruː/-/ˈkɛŋguru/ *kangaroo-Känguru*, 1770 by Capt. Cook), yielding 5362 sets of consonant clusters (see Table 1 for examples) and 4625 sets of vowel polyphones (see Table 2 for examples).
Before the modelling, word pairs/sets with missing data were excluded from consideration. In the modelling of Modern English and Modern German, words emerged after 1700 AC were excluded. In the modelling of recent English and German pronunciations, obsolete words for which no

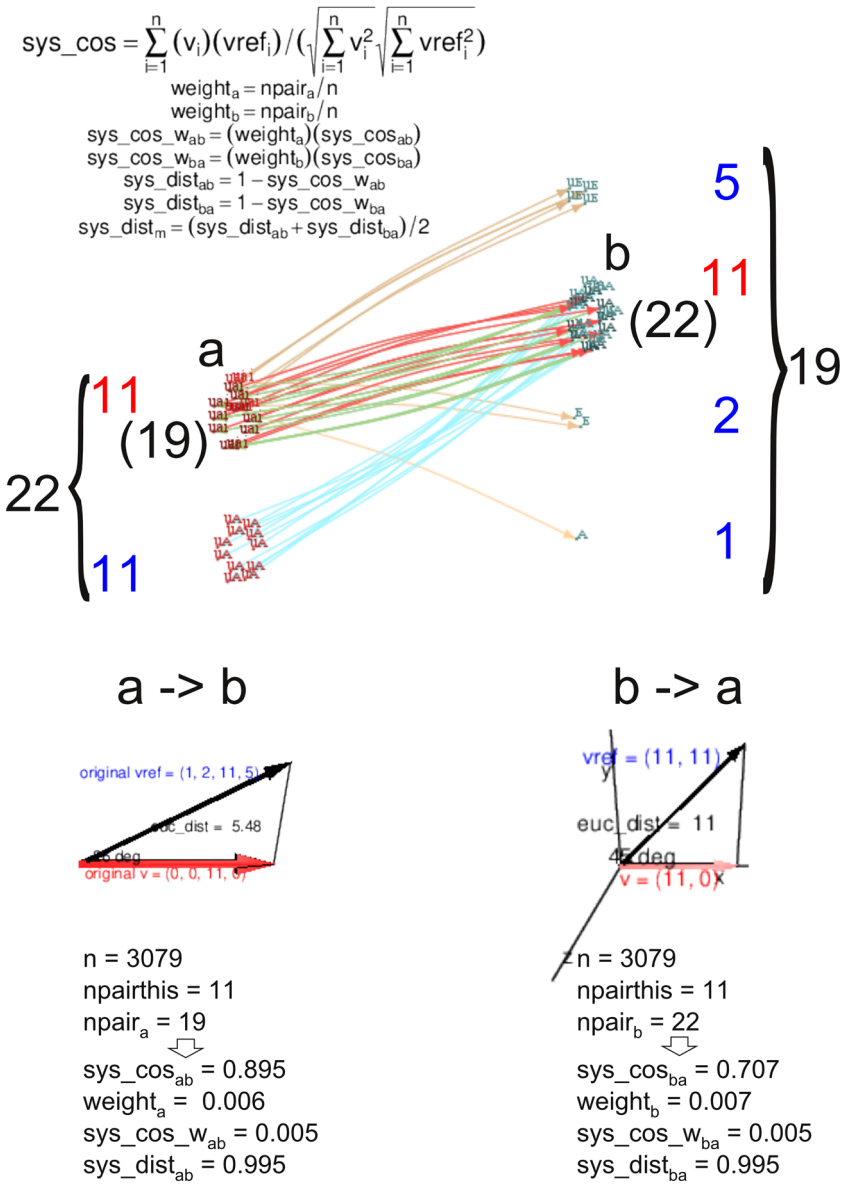


Fig. 1 Calculating weighted cosine systematicity (sys_cos_w) and systematic distance (sys_dist). Words clustered and labelled according to phonemic units (red for **a**, light-blue for **b**). Upper plot edges show pairings with direction of mapping. Lower plots show vectors *v* (red) and *vref* (black) with included angle and Euclidean distance. To calculate *sys_cos_w*, each mapping is represented as two vectors (*v* and *vref*), each of length equal to the count of this mapping's related mappings. Vector elements are pair counts. *Sys_cos_w* is cosine of included angle (v-vref) multiplied by proportion of involved pairs (*npair*) to total number of pairs (*n*).

definite pronunciations can be found were excluded. When the modelling involves Old English, Old High German, Modern English or Modern German, the positions of phonemes in words were considered according to Modern English and Modern German. When the modelling only involves recent English and German pronunciations, the positions of phonemes in words were considered according to the recent pronunciations (See supplementary materials for further word-wise details).

(2) The dataset *thirty-year sliced morphemic transcriptions for Chinese dialects in Shanghai* (You, 2013) includes IPA transcriptions for 3151 Chinese morphemes from twenty-two local Chinese dialectal varieties spoken in Shanghai collected in 2008, within which ten dialectal varieties (sub-dialects) were sampled twice, in the 1980s (o: old) and in 2008 (n: new) respectively, and the *Shiqu* (central urban) sub-dialect was additionally sampled in the 1990s (m: median). Additionally, the corresponding pronunciations in Standard Chinese (SC) were marked by the first author (PSC level 1-B). See Table 3 for examples.

Each syllabic transcription was split into its respective segmental combinations, onset consonants, rhymes, vowels, final consonants, and tone classes and then organised into tables of pairs, as exemplified in Table 4. It is important to note that the denoting numbers for tone classes hold a significant connection, as they correspond to the medieval Chinese tones. These tones consist of Yiping (1), Yangping (2), Yishang (3), Yangshang (4), Yinqu (5), Yangqu (6), Yinru (7), and Yangru (8) (Pan and Zhang, 2015). These names are widely employed to indicate related tonal categories in different modern Chinese dialects (Ho, 2015), and they also imply a similarity in tonal realisations across various sub-dialects of Shanghai.

The treatment of the datasets involves the following noteworthy practices:

Table 1 Data examples of consonant clusters (5362 sets, the column for notes not included).

Lemma	PIE_ consonant	OE_ consonant	OHG_ consonant	OE_ word	OHG_ word	ModE_ consonant	ModG_ consonant	recentE_ consonant	recentG_ consonant	ModE_ word	ModG_ word	recentE_ word	recentG_ word	position_ in_word_ Mod	position_ in_word_ Recent
after_ prefix	pt	ft	ft	æfter	aftar	ft	ft	ft	NULL	after	after-_OBS (prefix)	ˈɑːf.tər	NULL	intersyl	intersyl
after_ prefix	NULL	r	r	æfter	aftar	r	r	ɹ	NULL	after	after-_OBS (prefix)	ˈɑːf.tər	NULL	final	final
alike_same_very_similar	g	ġ	g	ġelīc	gilīh	Ø	g	Ø	g	alike	gleich	əˈlaɪk	glaiç	initial	initial
alike_same_very_similar	l	l	l	ġelīc	gilīh	l	l	l	l	alike	gleich	əˈlaɪk	glaiç	intersyl	intersyl
alike_same_very_similar	k	c	h	ġelīc	gilīh	k	ch	k	ç	alike	gleich	əˈlaɪk	glaiç	intersyl	final
…															

Table 2 Data examples of vowel polyphones (4625 sets, the column for notes not included).

Lemma	OE_ vowel	OHG_ vowel	OE_ word	OHG_ word	ModE_ vowel	ModG_ vowel	recentE_ vowel	recentG_ vowel	ModE_ word	ModG_ word	position_in_ word_Mod	position_in_ word_Recent
after_ prefix	æ	a	æfter	aftar	a	a	ɑ	a	after	after-_OBS (prefix)	initial	initial
after_ prefix	e	a	æfter	aftar	e	e	ə	ɐ	after	after-_OBS (prefix)	intersyl	intersyl
alike_same_very similar	e	i	ġelīc	gilīh	a	Ø	ə	Ø	alike	gleich	initial	initial
alike_same_very similar	ī	ī	ġelīc	gilīh	i	ei	aɪ	ai	alike	gleich	intersyl	intersyl
alike_same_very_similar	NULL	NULL	ġelīc	gilīh	e	Ø	Ø	Ø	alike	gleich	final	final
…												

Table 3 Examples of IPA transcriptions for Chinese morphemes spoken in (3151 entries).

Page	id	Morpheme	Medieval	Shi-qu_o	Shi-qu_m	Shi-qu_n	Zhe-nru	Jian-gwan	Song-jiang_o	Song-jiang_n	Sij-ing	Feng-xian_o	Feng-xian_n	Feng-cheng	Jins-han_o	Jins-han_n	Feng-jing	Qing-pu_o	Qing-pu_n	Xinz-huang	Minh-ang_n	Chua-nsha_o	Chua-nsha_n	Gao-qiao	Sanlin	Zhoupu	Nan-hui_o	Nan-hui_n	Jiad-ing_o	Jiad-ing_n	Bao-shan_o	Bao-shan_n	Chong-ming_o	Chong-ming_n	Buzhen	Liantang	SC
1	1	多	果开一平歌端	tu1	tu1	tu1	ʔdu1	ʔdu1	ʔdu1	tu1	ʔdu1	ʔdu1	ʔdu1	ʔdu1	ʔdu1	du6	tu1	ʔdu1	tu1	ʔdu1	tu1	ʔdu1	du6	ʔdu1	ʔdu1	ʔdu1	ʔdu1	tu1	tu1	tu1	tu1	tu1/tu1	tu1	tu1	tuo1
1	7	大大小	果开一平歌端	du6/d_A6	du6	du6	du6/d_A6	du6/d_A6	du6/da6	du6	du6/da6	du6/da6	du6	du6/da6	du6/d_A6		du6/d_A6	du6	du6	du6/d_A6	du2	du6/d_A6		du6/d_A6	du6/d_A6	du6/d_A6	du6	du6	du6/d_A6		du6	du6	du6/da6	du6/da6	du6/da6	t_A5
237	1887	段	山合一去换定	dø6	dø6	dø6	dø6	dø6	dø6	dø6	dø6	dø6	dø6	dø6	dø6	dø6	dø6	dø6	dø6	dø6	dø2	ʔda5	dø6	dø6	dø6	dø6	dø6	dø6	dø6	tø6	dø6	dø6	dø6	dø6	dø6	tuan5

Table 4 Examples of pair tables for syllables (Syl), segmental combinations (Seg), onset consonants (Ons), rhymes (Rhy), vowels (Vow), final consonants (Fin), and tone classes (Ton) between SC and the Shiqu_o (old central urban) Shanghainese variety.

Unit of consideration	SC unit	Shiqu unit	Mapping	Lemma
Syl	tuo1	tu1	tuo1 ~ tu1	1_多
Syl	t_A5	d_A6	t_A5 ~ d_A6	7_大大小
Syl	t_A5	du6	t_A5 ~ du6	7_大大小
Syl	tuan5	dø6	tuan5 ~ dø6	1887_段
…				
Seg	tuo	tu	tuo ~ tu	1_多
Seg	t_A	d_A	t_A ~ d_A	7_大大小
Seg	t_A	du	t_A ~ du	7_大大小
Seg	tuan	dø	tuan ~ dø	1887_段
…	…			
Ons	*t*	*t*	*t* ~ *t*	1_多
Ons	*t*	*d*	*t* ~ *d*	7_大大小
Ons	*t*	*d*	*t* ~ *d*	7_大大小
Ons	*t*	*d*	*t* ~ *d*	1887_段
…				
Rhy	uo	*u*	uo ~ *u*	1_多
Rhy	*A*	*A*	*A* ~ *A*	7_大大小
Rhy	*A*	*u*	*A* ~ *u*	7_大大小
Rhy	uan	ø	uan ~ ø	1887_段
…				
Vow	uo	*u*	uo ~ *u*	1_多
Vow	*A*	*A*	*A* ~ *A*	7_大大小
Vow	*A*	*u*	*A* ~ *u*	7_大大小
Vow	ua	ø	ua ~ ø	1887_段
…				
Fin	Ø	Ø	Ø ~ Ø	1_多
Fin	Ø	Ø	Ø ~ Ø	7_大大小
Fin	Ø	Ø	Ø ~ Ø	7_大大小
Fin	*n*	Ø	*n* ~ Ø	1887_段
…				
Ton	1	1	1 ~ 1	1_多
Ton	5	6	5 ~ 6	7_大大小
Ton	5	6	5 ~ 6	7_大大小
Ton	5	6	5 ~ 6	1887_段
…				

First, rather than confining the analysis to only core-words with established common origins, as previous studies in historical linguistics and modelling have done (Swadesh, 1955; Zhang et al., 2019; Sagart et al., 2019), the current approach considers all available lexical entries that are related. These entries may have derived from shared origins, historical or recent borrowing, or even borrowing from a third language, thus encompassing a broad spectrum of related vocabularies.

Second, our methodology diverges from classical historical practices that try to match language strata across related languages or dialects, as exemplified by previous studies (e.g., Wang, 2010; Chen, 2019). Specifically, we treat each pronunciation variant for a word as a distinct data entry and cross-reference it with the corresponding word in the other language under examination. For instance, if in language A the word "a" has two pronunciation variants, "a1" and "a2," and its counterpart "b" in language B also has two pronunciation variants, "b1" and "b2," we consider four data entries for modelling: "a1~b1," "a1~b2," "a2~b1," and "a2~b2." We employ this approach because we cannot verify or guarantee that public knowledge, such as "a1" mapping to "b1" but not to "b2," holds true in the speaker populations being examined. Conversely, anecdotal evidence from bilingual/bi-dialectal individuals seems to suggest that they frequently possess knowledge of divergent mappings.

Analyses. We utilised the weighted cosine systematicity (*sys_cos_w*) and systematic distance (*sys_dist*) measurements on both datasets. In relation to the English and German dataset, we examined vowel and consonant *sys_cos_w* and *sys_dist* overall, as well as at the beginning, middle, and end of words. As it comes to the Shanghai dataset, we tested the two measurements across various combinations of time-slices and locations at different levels including syllables (Syl), segmental combinations (Seg), onset consonants (Ons), rhymes (Rhy), vowels (Vow), final consonants (Fin), and tone classes (Ton). Moreover, for comparison purposes, we estimated pronunciation distances by calculating and analysing Optimal String Alignment (OSA) (van der Loo, 2014) among related lexical/morphemic forms.

Three sets of analyses were then applied on each dataset.

(1) The weighted cosine systematicity (*sys_cos_w*) data from both datasets regarding all the phonemic units conforms to a long-tailed Poisson distribution or a Power Law distribution, so they were multiplied by 1000 and natural log-transformed before being fed into Mixed-Linear-Effect models (nevertheless, many subsets of the corrected data may still not conform to normal distribution). The LME models (Bates et al., 2013) were fit with time-slices and the directions of mapping as the fixed predictors, as well as mappings as the random predictors.

(2) The average *sys_dist* data were calculated for each linguistic level and linguistic variety, and the average OSA distance for each linguistic variety. Subsequently, both the systematic distance and the OSA distance data were subjected to Multi-Dimensional Scaling (MDS) analysis (Venables and Ripley, 2002), followed by further analysis.
In terms of *sys_dist* MDS for the dialects in Shanghai, we initially examined the relationship between the density of the old and new sub-dialects' neighbourhoods and their distance from the statistical centroid.
To explore the co-evolution direction within this dialectal dataset, we conducted paired t-tests on the mean systematic distances between ten local sub-dialects' old and new variations in relation to (1) the statistical centroid, (2) the regional high variety *Shiqu*, and (3) the national high varieties SC. Additionally, we performed by-sub-dialect paired t-tests on the old and new mean OSA distances to the same three centres, for comparison.
Furthermore, Pearson and Spearman correlations were employed to assess the correlation between the local sub-dialects' distances to the central urban dialect and their geographic and commuting distances to the city centre.

(3) To investigate the relationship between the initial *sys_cos_w* values of individual words and the subsequent changes they undergo, we utilised a custom-built multi-line regression function (DOI 10.17605/OSF.IO/56VT8). The function aims to capture the relationship between two variables with multiple straight lines. It takes in a set of data points with x, y coordinates. Given the largest count of the lines intended for it to fit, and the number of bins for the x coordinates, it applies the k-means partition (Hartigan and Wong, 1979; R Core Team, 2022) on the y coordinates of each bin, based on a Silhouette criterion to decide the optimal number of partitions, and fits straight lines with the centroids of aligned partitions across these bins. The coefficients and intercepts of these fitted straight lines are then repeatedly adjusted after assigning each point to its nearest line until it reaches a given number of iterations.

Results

English-German lexical co-evolution. First, we applied the weighted cosine systematicity and systematic distance measurements on the dataset of *old to recent English-German related words*. Figure 2 presents the obtained *sys_cos_w* data as network diagrams.

Based on the butterfly-scatter-box-violin plots and MDS spaces presented in Fig. 3, no consistent increase or decrease in systematicity and systematic distance nor any clear, long-term pattern was observed in the MDS spaces over time. The trend varies with phoneme types and positions.

Nevertheless, regarding the OSA distances across related words, LME modelling showed that pronunciations of English and German related words are significantly diverging across generations, $t_{old} = -16.50$, $p < 0.0001$, $t_{rec} = 5.22$, $p < 0.0001$ (with modern OSA as the baseline). See Fig. 4c for its OSA MDS space.

Dialectal co-evolution in Shanghai. Then we applied the weighted cosine systematicity and systematic distance measurements to the dataset of *thirty-year sliced morphemic transcriptions for Chinese dialects in Shanghai*, which represents a set of non-literal local sub-dialects co-evolving with a regional high variety (*Shiqu*, central urban Shanghainese) and a national high variety (*SC*, Standard Chinese).

As shown in Fig. 5, in the MDS spaces, the dialectal varieties distribute in a similar way as magnets in magnet fields, with a sub-dialect's neighbourhood density decreasing with the increase of its distance to the statistical centroid. The regional and national high varieties (*Shiqu* & *SC*) are located at or closely to the centre (except for SC regarding the whole-tonal syllables).

The local sub-dialects' old versus new mean systematic distances and mean OSA distances to the (1) statistical centroid, (2) the regional high variety *Shiqu*, and the (3) national high varieties *SC*, as well as the corresponding t-statistics are illustrated by clustered heatmaps in Fig. 6. Within thirty years, all of the sub-dialects experienced a decrease in systematic distances to the *Shiqu* variety (Fig. 6, b1, except for onsets and tones). Conversely, some sub-dialects' diverged from the *SC* variety, while others converged (Fig. 6, c1). Nonetheless, the OSA distances of the local sub-dialects to the *SC* variety decreased consistently (Fig. 6, c2, except for finals and tones), whereas changes in their OSA distances from the *Shiqu* variety varied (Fig. 6, b2).

Systematic distances analysis also showed a decreased distance of the *Shiqu* variety to SC (Fig. 6, c1, 7th row, except for tones), in contrast to the divergence revealed by OSA distances (Fig. 6, c2, 1st row). See also the modelling of weighted cosine systematicity data in Figs. 7 and 8.

Furthermore, Fig. 9 reveals that the systematic distances between the sub-dialects and the *Shiqu* high variety are correlated with geographical distances and commuting times: for the old pronunciations, geographic distance and biking time are stronger correlative factors, whereas public-transportation-time is more prominent for the new pronunciations. In general, the correlation coefficients increased.

Comparing changes in systematic (Figs. 5 and 9) and OSA distances (Fig. 4a, b) in MDS spaces shows that new local varieties are mostly on or near the connecting lines between SC (national high variety, which locates much farther away) and old versions in the OSA space, but not in the *sys_dist* space.

Regression effects on the change of systematicity. Regarding individual word pairs, it appears that they have drastically different directions of co-evolutionary changes. Nonetheless, a

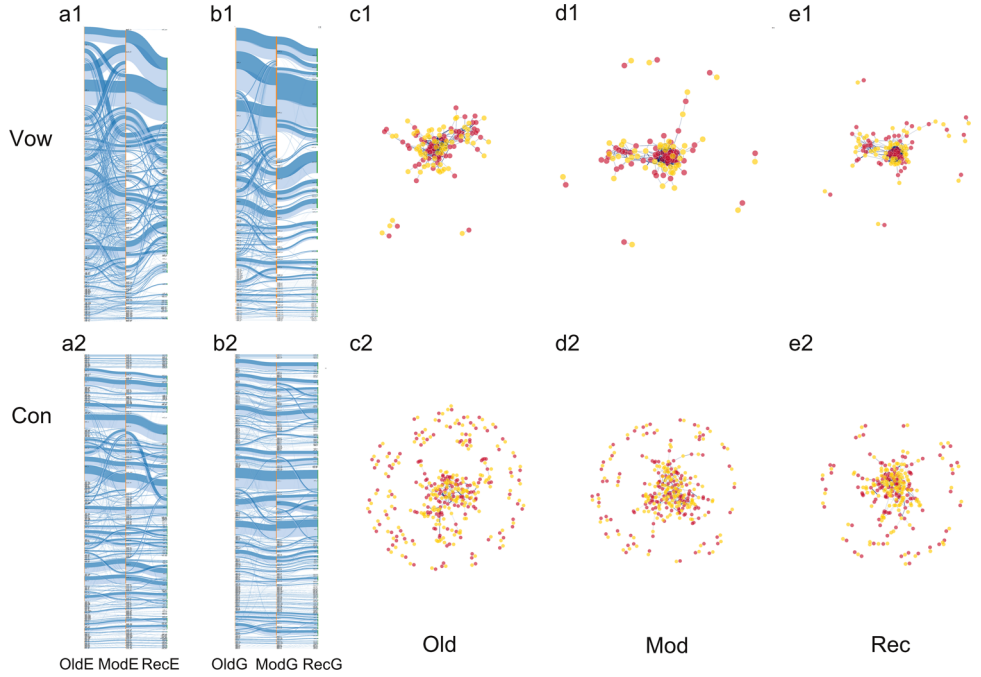


Fig. 2 Network diagrams of systematic correspondences. a1–b2 Sanky flow diagrams showing the relationship between Old English (OldE), Modern English (ModE) and Recent English (RecE), and Old (High) German (OldG), Modern German (ModG) and Recent German (RecG) for vowels (**a1**, **b1**) and consonants (**a2**, **b2**). **c1–e2** The weighted cosine systematicity between English and German units for vowel (**c1**, **d1**, **e1**) and consonant (**c2**, **d2**, **e2**) correspondence networks, in their Old (Old, **c1**, **c2**), Modern (Mod, **d1**, **d2**), and Recent (Rec, **e1**, **e2**) relationships, with English units represented in red and German units in yellow. The width of flow in **a1–b2** and the width of links in **c1–e2** represents *Weighted Cosine Systematicity (sys_cos_w)*, with dark blue indicating links from left to right (*ab*) and light-blue indicating links from right to left (*ba*) in **a1–b2** (Allaire et al. 2017).

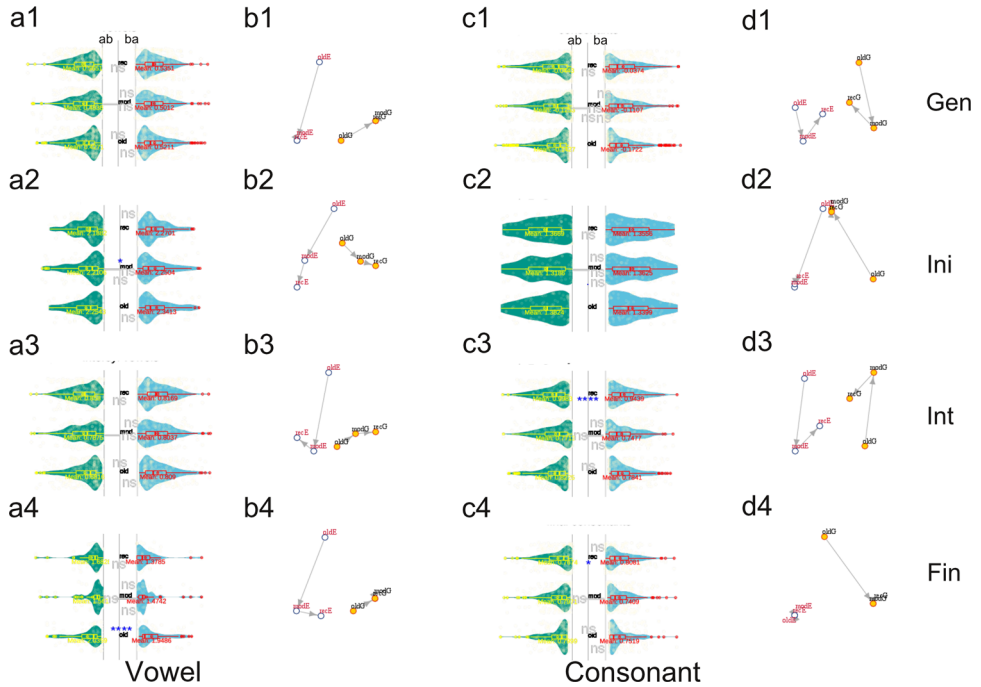


Fig. 3 Weighted cosine systematicity and systematic distances between English and German. a1–a4, c1–c4 Butterfly-scatter-violin-box plots show the weighted cosine systematicity multiplied by 1000 and natural log-transformed, across different time-slices and mapping directions. LME results are marked on conditions with positive effects, with main effects annotated in the centre of each plot. Left: *sys_cos_ab_w* for the mapping English→German; right: *sys_cos_ba_w* for the mapping German→English. From bottom to top: old, modern, recent. **b1–b4, d1–d4** MDS plots based on mean *sys_dist* measurements illustrate the co-evolutionary trajectories of English and German varieties. Arrows are plotted from old varieties to modern varieties and from modern varieties to recent varieties. Labels: "oldE" for old English, "modE" for Modern English, "recE" for recent English, "oldG" for Old High German, "modG" for Modern German, "recE" for recent German. **a1–b4** For vowels**. c1–d4** For consonants**. a1**, **b1**, **c1**, **d1** In general**. a2**, **b2**, **c2**, **d2** At word initial positions**. a3**, **b3**, **c3**, **d3** At inter-syllabic positions within words. **a4**, **b4**, **c4**, **d4** At word final positions.

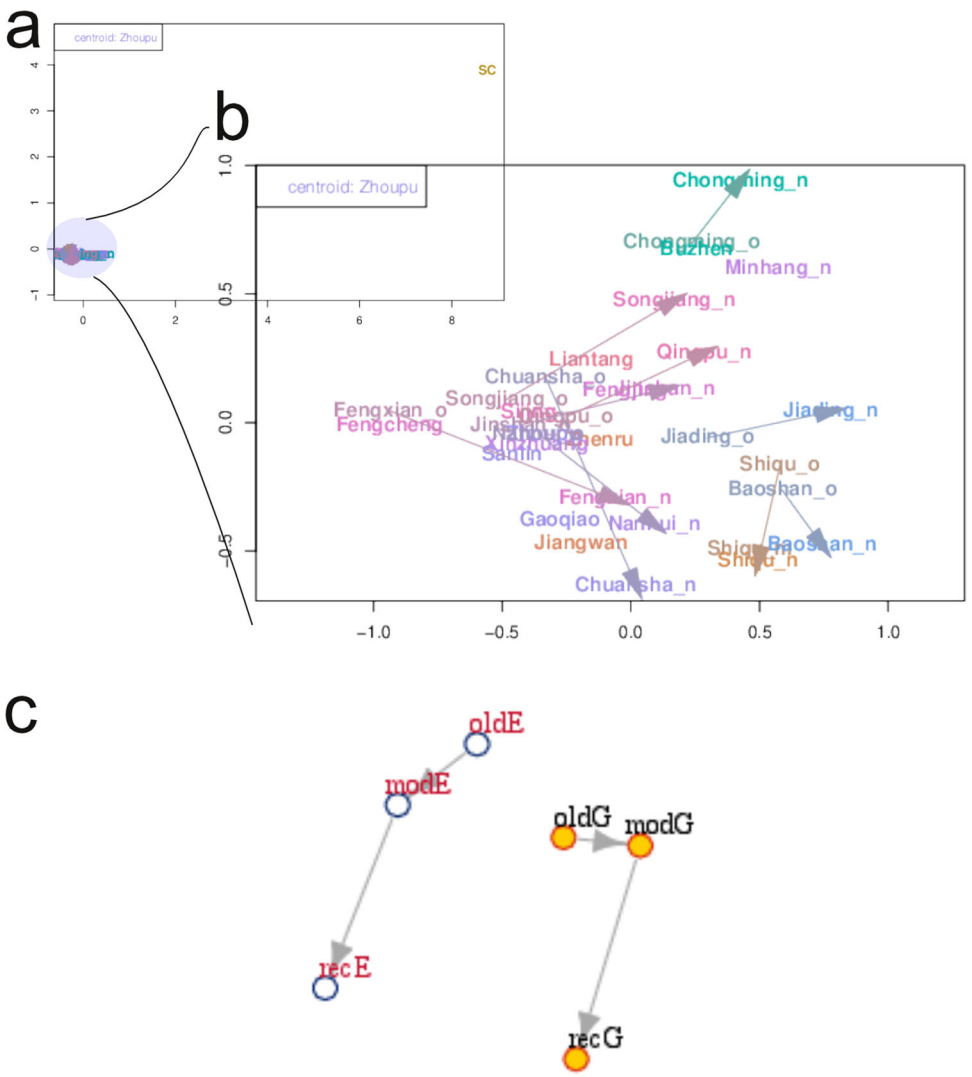


Fig. 4 Multi-dimensional scaling (MDS) visualisation of mean optimal string alignment (OSA) distances. a, **b** MDS plot for OSA distances between syllables across the dialectal varieties spoken in Shanghai, with Standard Chinese (SC) included (**a**) and excluded (**b**). An arrow is drawn from each sub-dialect's old version to its new version, and labels and points are colour-coded according to dialectal varieties. **c** MDS plot for OSA distances between lexical forms across old (*oldE* and *oldG*), modern (*modE* and *modG*), and recent (*recE* and *recG*) English and German varieties. An arrow is drawn from each language's old variety to modern vary and from modern variety to recent variety.

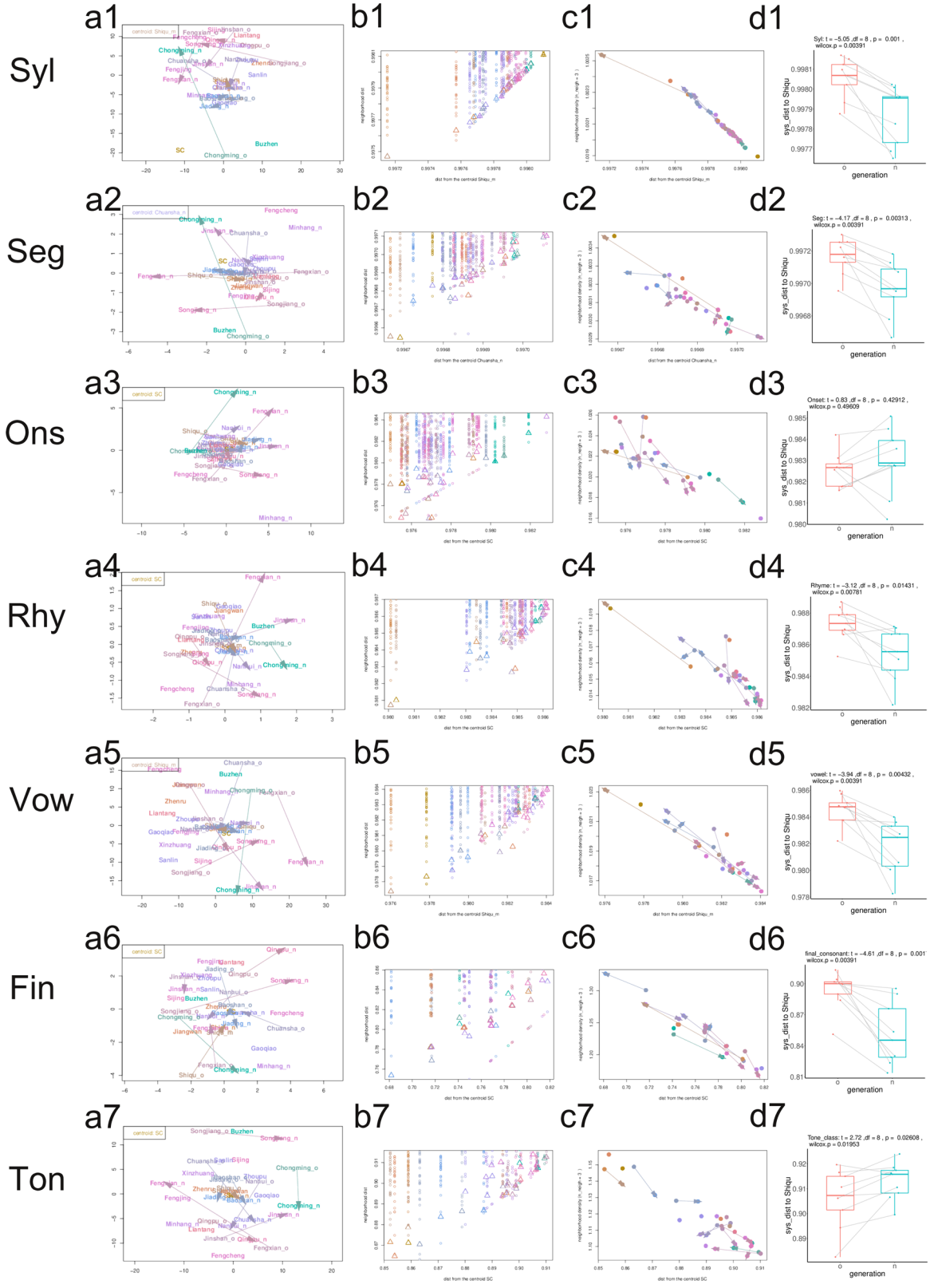
word-wise underlying pattern can be seen in both datasets: the Regression Effect over time (Galton, 1879; Senn, 2011) . As shown in Fig. 10, word pairs with high *sys_cos_w* scores in an earlier time slice tend to have lower or less increased scores in subsequent time-slices. Conversely, those with low *sys_cos_w* scores earlier tend to have higher or less reduced systematicity later on. This underlying pattern is evident despite the overall converging trend seen in the Shanghai dataset and testified in both datasets.

Discussion

Hypotheses tested. This study investigates how social ecology affects systematic correspondence in co-evolving languages, providing insights into general sound change mechanisms derived from a pool of synchronic variability (Ohala, 1989). On the one hand, we examined two socially independent yet related languages, English and German, which have maintained a balanced relation in systematic correspondence over centuries of diverging sound changes. On the other hand, the local dialectal varieties in Shanghai have systematically converged toward the regional and national high variety within decades, while still maintaining their distinct lexical pronunciations. The findings suggest that the systematicity of lexical relations between co-evolving languages is not doomed to increase or decrease. Therefore, neither the attrition hypothesis nor the integration hypothesis is adequately supported.

Instead, evidence exists to support the self-regulated adaptation theory. Given the variations in linguistic ecologies between these two cases, these findings indicate that having a stable structure with standardised higher linguistic varieties as "prototypes" (Dixon, 1997) or "superstrata" may have a significant impact on aligning lower forms of language with their corresponding higher forms. On the other hand, when two populations with their own separate standards come into contact without a shared prototype, this contact alone may not be enough to produce a similar effect. These findings support the notion that linguistic ecology plays a crucial role in shaping the development of languages co-evolution.

Furthermore, it should be noted that there is evidence of correspondence-based (Fig. 6, b1, c1) and similarity-based (Fig. 6,



b2, c2) convergence at work. These two types of convergence work together in a complementary manner. Additionally, there is the phenomenon of segmental convergence and tonal divergence counteracting each other (Fig. 6, b1). Moreover, there is the Regression Effect (Galton, 1879; Senn, 2011), whereby word pairs that had a higher degree of correspondence in the past tend to have lower or less increased correspondence in the present (and vice versa for word pairs with a lower degree of systematicity). All these findings support the self-regulated adaptation hypothesis and the more general self-organisation theory (Green et al., 2008).

Additionally, the change in systematic correspondence is influenced by external factors such as the strength of contact,

Fig. 5 Modelling of the systematic distances (sys_dist$_m$) in the Shanghai dataset. a1–a7 Multi-dimensional scaling (MDS) plots of the mean *sys_dist* for each dialectal variety spoken in Shanghai, with arrows connecting each sub-dialect's old version to its corresponding new version and statistical centroids annotated at the top-left of each plot. **b1–b7** Triangles representing the relationship between one variety's mean *sys_dist* to the statistical centroid (the horizontal coordinate) and its mean neighbuorhood *sys_dist* (the vertical coordinate, n_neigh = 3, taking three closest neighbours into consideration); Points representing the other varieties' mean *sys_dist* (the vertical coordinates) to this variety. **c1–c7** Points representing the relationship between one variety's mean *sys_dist* to the statistical centroid (the horizontal coordinate) and its mean neighbourhood density (= 1 – mean neighbourhood *sys_dist*, the vertical coordinate), again with arrows connecting each sub-dialect's old version to its corresponding new version. **d1–d7** Paired box plots for the range, quartiles, and median between old (o, in the left) and new (n, on the right) varieties' mean sys_dist$_m$ to the Shiqu variety, with results of paired t-tests and Wilcox-tests annotated at the top of each plot. These plots are vertically arranged according to the phonemic units: syllables (*Syl*: **a1**, **b1**, **c1**, **d1**), segmental combinations (*Seg*: **a2**, **b2**, **c2**, **d2**), onset consonants (*Ons*: **a3**, **b3**, **c3**, **d3**), rhymes (*Rhy*: **a4**, **b4**, **c4**, **d4**), vowels (*Vow*: **a5**, **b5**, **c5**, **d5**), final consonants (*Fin*: **a6**, **b6**, **c6**, **d6**), and tone classes (*Ton*: **a7**, **b7**, **c7**, **d7**). The labels and points are colour-coded consistently.

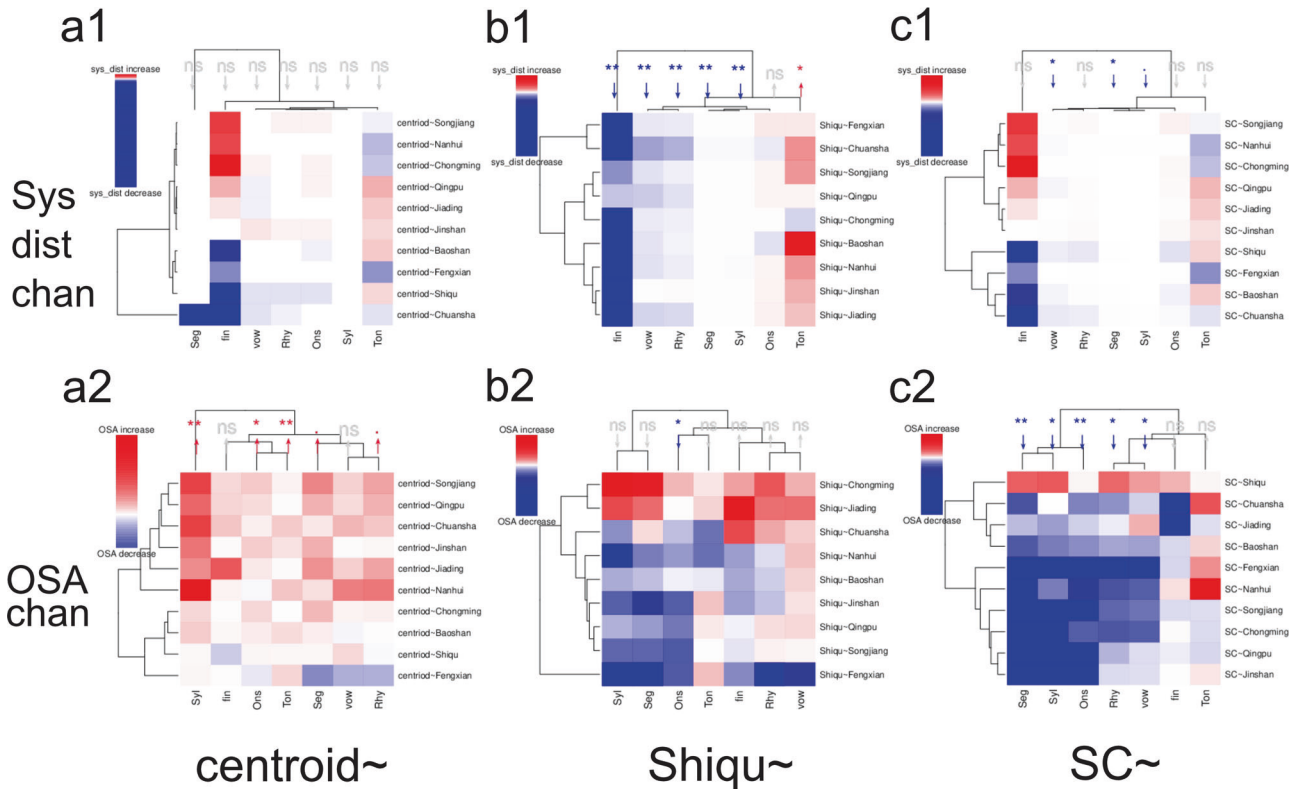


Fig. 6 Clustered heatmaps displaying the changes in systematic distance (sys_dist$_m$) and Optimal String Alignment (OSA) distances in the Shanghai dataset. a1, **a2** Changes in mean *sys_dist* (**a1**) and OSA (**a2**) between statistical centroid in the Shanghai local dialectal varieties (labelled on the right). **b1**, **b2** Changes in mean *sys_dist* (**b1**) and OSA (**b2**) between the Shanghai *Shiqu* (central urban) dialect and the other Shanghai sub-dialects (labelled on the right). **c1**, **c2** Changes in mean *sys_dist* (**c1**) and OSA (**c2**) between Standard Chinese (SC) and Shanghai local dialectal varieties (labelled on the right).The phonemic units of syllables (*Syl*), segmental combinations (*Seg*), onset consonants (*Ons*), rhymes (*Rhy*), vowels (Vow), final consonants (Fin), and tone classes (Ton) are labelled at the bottom of each plot, ordered according to the result of clustering. The heatmaps are colour-coded so that blue indicates decreasing distances and red indicates increasing distances, with higher saturation indicating greater changes. Significance codes and directions of paired t-tests comparing by-sub-dialect mean old and new *sys_dist/OSA* are annotated at the top of each plot (red↑for increasing, blue ↓for decreasing).

language status, and geographical distances, consistent with Labov's (1963) theory of social motivations for sound change, Mufwene's (2001) ecological theory for language contact, and Dixon's (1997) punctuated equilibrium model.

Unveiling the historical transformations of specific languages. The current findings align with previous studies in historical linguistics and dialectology, providing further insight into the specific languages under investigation: German, English, and Chinese dialects.

Historical linguists can uncover significant historical linguistic events by examining the systematic distances in the MDS spaces. For example, the divergence of English and German vowels, as well as the larger systematic distances in vowels between Old English and Modern English compared to Old High German and Modern German (Fig. 3, b1–b4) indicate that English vowel evolution is less regular than German vowel evolution during this stage. This finding may represent the influence of the Great Vowel Shift (1350–1700) in English (Jespersen, 1909). This shift involved conditional sound changes and lexical diffusion (Wang, 1969), contributing to the misalignment of pronunciations we observe in modern English and German spellings.

Conversely, when examining the evolution of final consonants (Fig. 3, d4), we find that the systematic distances between Old High German and Modern German are larger than those between Old English and Modern English. This suggests that English final

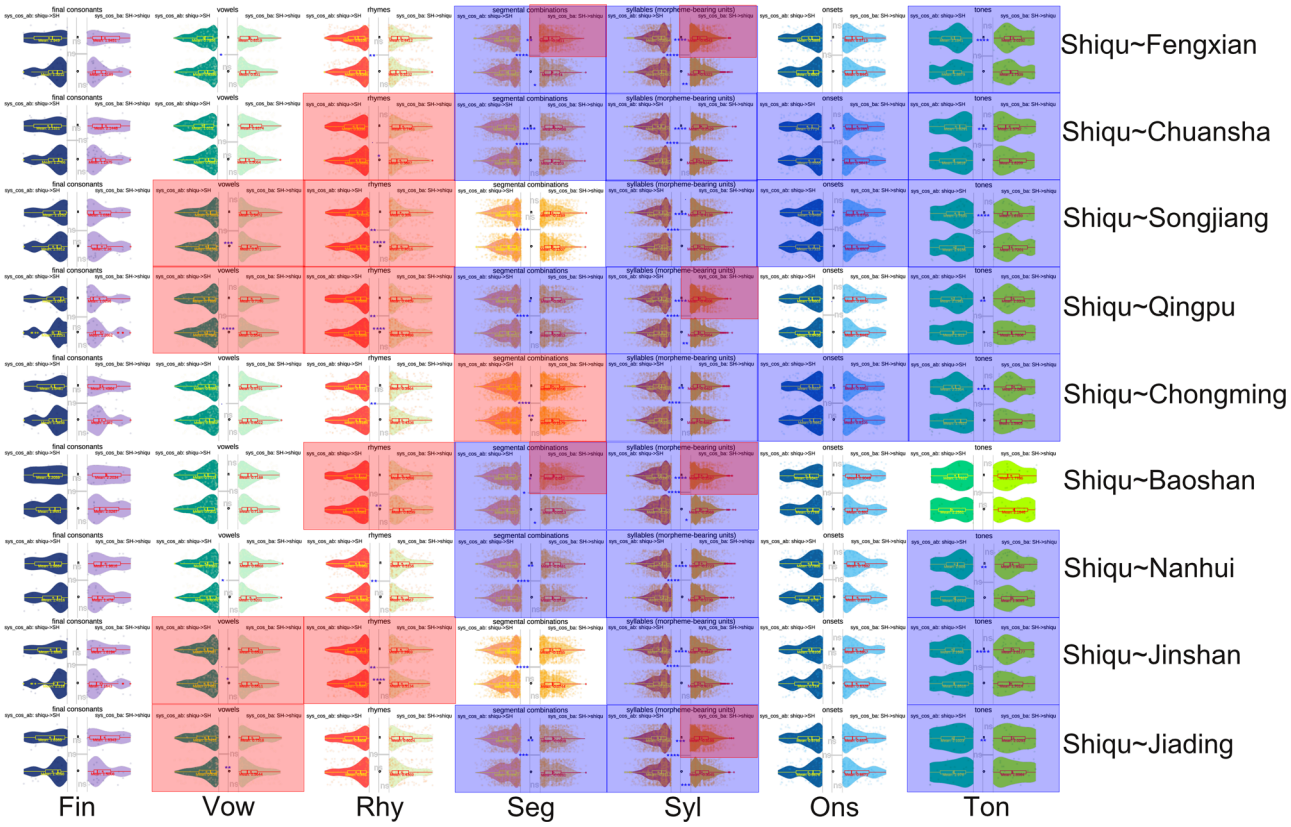


Fig. 7 Modelling of the weighted cosine systematicity (*sys_cos_w*) between Shanghai *Shiqu* and the other local dialectal varieties. Butterfly-scatter-violin-box plots are used to show the *sys_cos_w* values (multiplied by 1000 and natural log-transformed) for different time-slices (*old* and *new*) and mapping directions (left: *Shiqu→SH local*, right: *SH local→Shiqu*). Each translucent point on the scatter plots represents a single mapping, with its horizontal coordinate representing the *sys_cos_w* value (the vertical coordinate is jittered for visualisation purposes). The coloured violin shapes indicate the probability density of the sys_cos_w values, while the box plots represent the range, quartiles, median and odd values, annotated with means (yellow for *Shiqu→SH loca* and red for *SH local→Shiqu*). Based on the LME results, significance codes are marked on conditions with positive effects, with main effects annotated on the grey cross lines at the centre, and interaction effects in the panel related to the interaction term. Blue shades are added to cells representing positive main effects for the new varieties, and red shades are added to cells representing positive main effects for the new varieties. In addition, blue shades on the right corner of cells represent positive interaction effects for the mapping *SH local→Shiqu*, and red shades represent negative interaction effects. The subplots are vertically arranged according to the pairs of linguistic varieties, and horizontally according to the phonemic units.

consonants evolve more regularly than German final consonants at this stage, possibly influenced by the drop of English infinitive verb suffices (although see Szmrecsanyi, 2012) and the complex changes of German verb conjunctions, such as the alternation between strong and weak inflections (Bailey, 1997).

On the other hand, the observation that the local dialectal varieties in Shanghai have systematically converged toward the national standard aligns with previous examples in dialectology, which demonstrate that certain sound categories of the central urban Shanghainese varieties have become more in sync with Standard Chinese (Qian, 2003; Chen, 2019). Interestingly, we not only statistically consolidated the previous findings on the sub-dialects in Shanghai, but also observed a previously ignored phenomenon, namely that the standardised national high variety (*SC*) and the regional high variety (*Shiqu*, not standardised but strongly associated with higher social status) influence the low varieties with different biases on two complementary mechanisms: one based on pronunciation similarity, the other based on systematic correspondence. Furthermore, it is worth emphasising that lexical tones in these Chinese dialects, being suprasegmental phonemes, exhibit distinct co-evolution patterns as opposed to segmental phonemes. Moreover, certain local linguistic varieties further highlight this distinctive nature of tones, as they appear to be accompanied by onsets or finals. This correlation between onsets/finals and the historical shift of tones in Chinese has been identified in previous renown studies (Pan and Zhang, 2015; Karlgren, 1915–1916).

In addition, the present findings provide additional illustrations for two commonly applied rules that are relevant across different areas. Firstly, the value of *sys_cos_w* for each cross-linguistic mapping rule is found to be inversely proportional to its ranking, aligning with Zipf's Rule (Chao and Zipf, 1949). Secondly, word pairs that had stronger correspondence in the past tend to exhibit lower or less significant increase in correspondence in the present (and vice versa for pairs with lower systematicity), which aligns with the Regression Effect (Senn, 2011).

Limitations. It is important to consider several limitations of our method.

Firstly, due to the unavailability of aligned lexical frequency information and lexical stress of Old English and Old High German, as well as the lack of certainty in the syllabic boundaries of English and German, we were unable to incorporate them into our method, which may have implications for the accuracy and completeness of our analysis.

Secondly, although our sample size was the largest to our knowledge, the possibility of enlarging it should be explored in

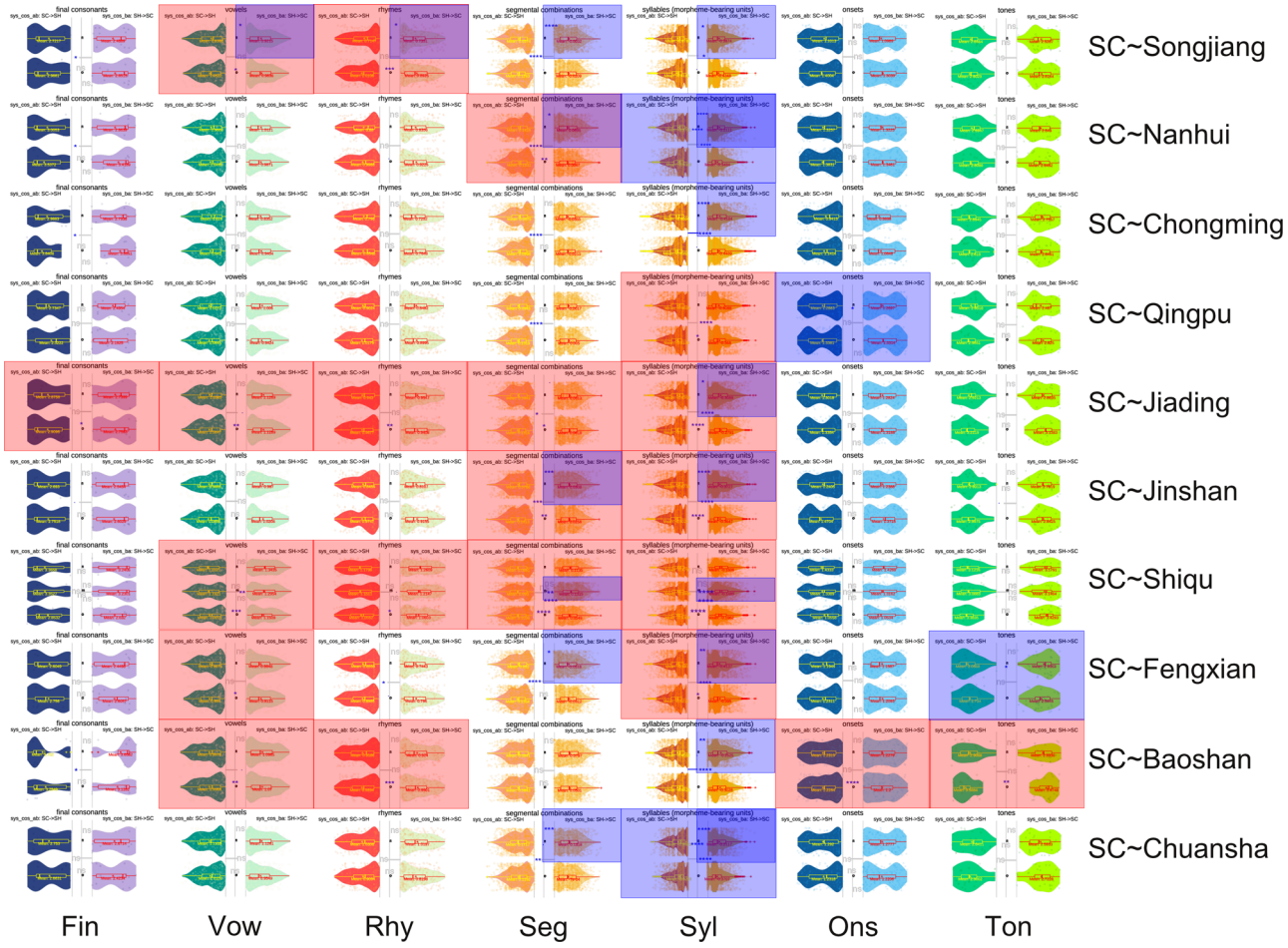


Fig. 8 Modelling of the weighted cosine systematicity (*sys_cos_w*) between Standard Chinese (SC) and the Shanghai local dialectal varieties. The butterfly-scatter-violin-box plots show the value of *sys_cos_w* (multiplied by 1000 and natural log-transformed) for each time slice (*old*, *middle*, and *new*) and mapping direction (*SC to SH local*, and *SH local to SC*). Each translucent point on the scatter plot represents one mapping, with its horizontal coordinate representing the *sys_cos_w* value (the vertical coordinate is jittered). The coloured violin shapes indicate the probability density of the *sys_cos_w* values, while the box plots represent the range, quartiles, median and odd values, with means indicated in yellow (for *SC to SH local*) and red (for *SH local to SC*). Based on the Linear Mixed Effects results, significance codes are marked on conditions with positive effects, main effects annotated on the grey cross lines at the centre, and interaction effects in the panel relating to the interaction term. Blue shades are added to cells representing positive main effects for the new varieties, and red shades for positive main effects for the old varieties. Additionally, blue shades in the right corner of cells signify positive interaction effects for the mapping *SH local to SC*, and red shades for negative interaction effects. The subplots are arranged vertically according to the pairs of linguistic varieties, and horizontally according to the phonemic units.

future studies to enhance the statistical power and reliability of our findings.

Moreover, as we only utilised the two available datasets, questions regarding their representativeness naturally arise. While we made efforts to include two diverse and large datasets, we acknowledge that further datasets could provide a more comprehensive perspective.

To ensure the applicability of our theories to a broader range of historical linguistics and dialectology, it would be valuable to apply the proposed measurement, weighted cosine systematicity, to datasets from a broader range of linguistic ecology and to include a larger variety of time-slices. For instance, datasets from Sprachbund, where systematic correspondences are formed purely through contact rather than shared origins, e.g., the Balkan languages, can provide valuable insights. Similarly, datasets from adstratum languages, such as two standard languages are used in parallel in a country, e.g., modern French and Dutch in Belgium, can offer significant contributions. Additionally, exploring dialects within one country that do not explicitly denote related vocabularies with the same set of ideographic symbols, e.g., the Dutch, Norwegian, or English dialect (Trudgill, 1986), can also provide valuable information. By studying these diverse datasets, we may enhance our understanding of historical linguistics and dialectology in a more comprehensive manner.

Lastly, it is crucial to consider the potential biases that could arise from our approach to handling pronunciation variants. While we made a conscious effort to minimise assumptions about speakers' knowledge, variations in the decision to include certain lexical variants during the original data collection process may still have influenced the results. It is important to acknowledge that these variations could introduce biases and potential limitations that might impact the overall findings and interpretations of our study.

By examining and acknowledging these limitations, we may ensure a more accurate and robust evaluation of our methodology.

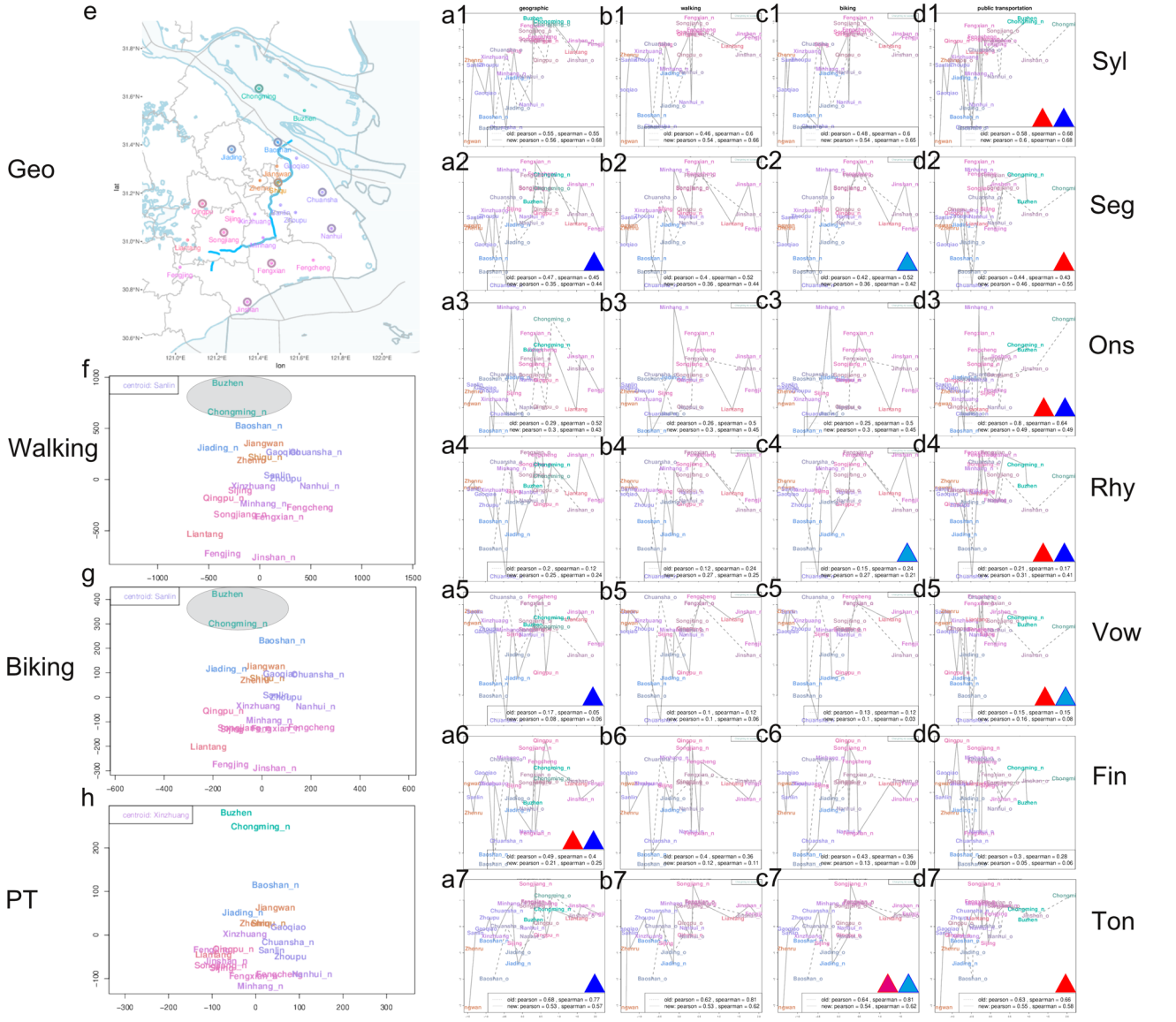


Fig. 9 The Correlation between real-world and systematic distances in the Shanghai dataset. a1–d7 The vertical coordinates represent scaled mean *sys_dist* between old (dashed lines) and new (solid lines) Shanghai *Shiqu* (urban centre) varieties and the other Shanghai local dialectal varieties; the horizontal coordinates represent scaled geographic distance (**a1–a7**), scaled walking time (in minutes, **b1–b7**), scaled biking time (in minutes, **c1–c7**), or scaled public-transportation time (in minutes, **d1–d7**) between city centre and these locations. The panels are arranged vertically according to phonemic units: syllables (*Syl*: **a1**, **b1**, **c1**, **d1**), segmental combinations (*Seg*: **a2**, **b2**, **c2**, **d2**), onset consonants (*Ons*: **a3**, **b3**, **c3**, **d3**), rhymes (*Rhy*: **a4**, **b4**, **c4**, **d4**), vowels (*Vow*: **a5**, **b5**, **c5**, **d5**), final consonants (*Fin*: **a6**, **b6**, **c6**, **d6**), and tone classes (*Ton*: **a7**, **b7**, **c7**, **d7**). Both Pearson and Spearman correlations are annotated in right bottom legends. Triangles on the plots indicate the real-world distance measurements tested that show the highest correlation with the old varieties (blue for Pearson, light-blue for Spearman), and with the new varieties (red for Pearson, fuchsia for Spearman). **e** Geographic locations where the Shanghai dialectal data were collected, the new varieties are marked with coloured points, the old varieties marked with coloured circles, administrative boundaries marked in grey, and hydrographic objects marked in light-blue (according to You 2013; Baidu Maps 2022a; National Geomatics Center of China 2022). **f**, **g**, **h** MDS plots created using walking time (*Walking*, **f**), biking time (*Biking*, **g**), and public-transportation time (*PT*, **h**) across the sampling sites (according to Baidu Maps 2022b), labelled with the corresponding dialectal varieties. *Chongming* and *Buzhen* cannot be accessed exclusively by walking or cycling.

Conclusion

This study has focused on two important scenarios. Firstly, we have examined the co-evolution of two closely related national languages with equal social status by analysing the English-German dataset. Secondly, we have investigated the co-evolution of non-literal local sub-dialects alongside a regional and national high variety using the Shanghai dataset. By utilising weighted cosine systematicity, a vector-based measurement, we have been able to explore the quantitative impact of linguistic ecology on language co-evolution and have tested various co-evolutionary theories. This study provides valuable quantitative insights into the historical transformations of specific languages, which can be generalised to a broader scope of historical linguistics and dialectology. Overall, it sheds light on the mechanisms and patterns of language evolution, contributing to a deeper understanding of the complex and dynamic nature of languages over time.

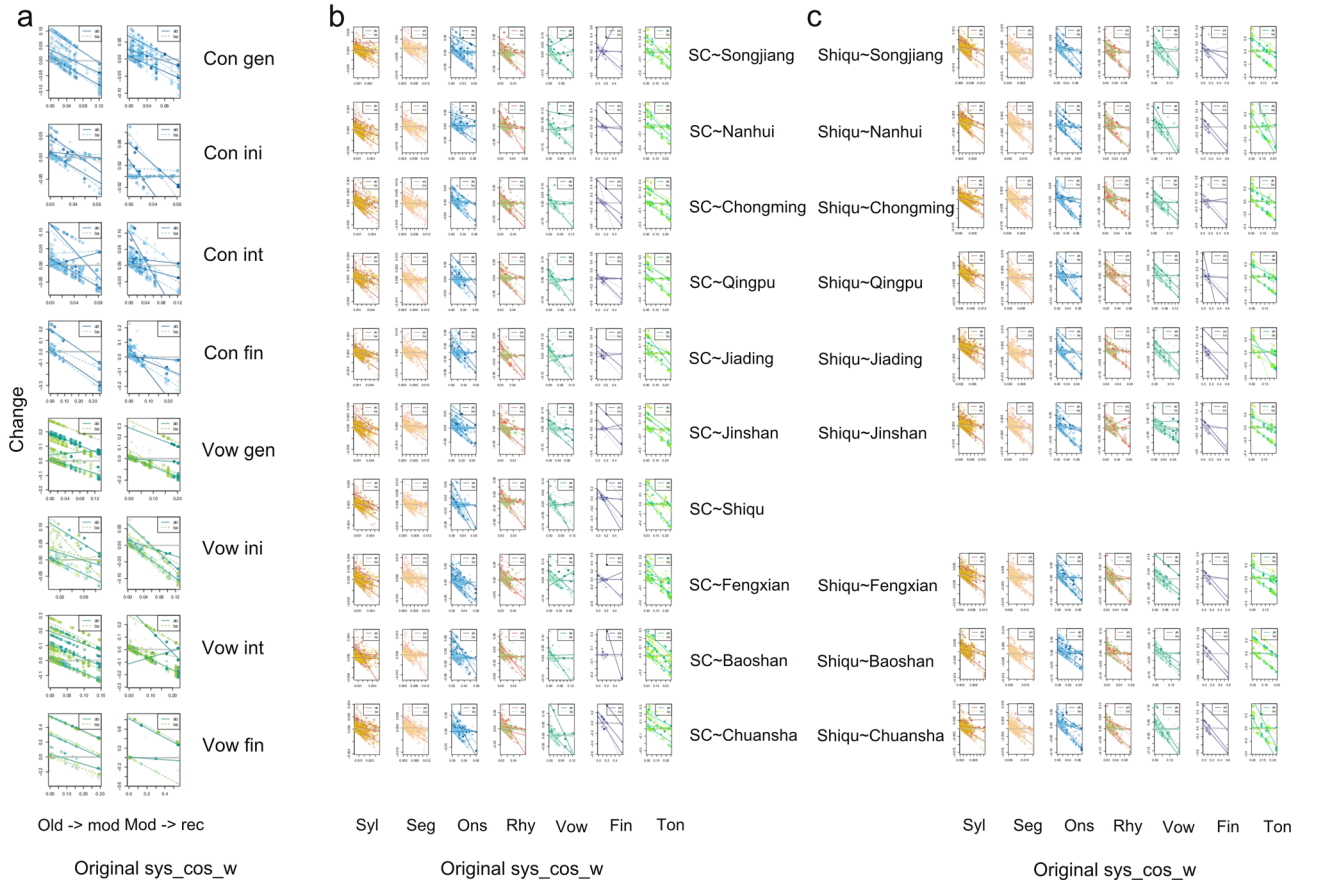


Fig. 10 Weighted cosine systematicity (sys_cos_w) and its word-wise changes. The horizontal coordinate of each point in each subplot represents the lemma's original *sys_cos_w*, and the vertical coordinate represents its subsequent change, as calculated by subtracting the original *sys_cos_w* from the latter *sys_cos_w*. The zero line of change (grey solid line) divides each plot into two parts. Mappings considered from different directions are colour-coded, as annotated with legends within the subplots. Fitting lines are plotted within each subplot (solid lines for the mappings left variety→right variety, dashed lines for the mappings right variety→left variety). **a** For changes of English-German *sys_cos_w*, vertically arranged according to phonemic units—consonants in general (*Con gen*), initial consonants (*Con ini*), inter-syllabic consonants (Con int), final consonant (Con fin), vowels in general (*Vow gen*), initial vowels (*Vow ini*), inter-syllabic vowel (*Vow int*), and final vowels (*Vow fin*), horizontally arranged according to the original and later time-slices (left: *old to modern*, right: *modern to recent*). **b** For changes of *sys_cos_w* between SC and Shanghai local dialectal varieties, vertically arranged according to pairs of linguistic varieties, horizontally arranged according to phonemic units—syllables (*Syl*), segmental combinations (*Seg*), onset consonants (*Ons*), rhymes (*Rhy*), vowels (*Vow*), final consonants (*Fin*), and tone classes (*Ton*). **c** For changes of *sys_cos_w* between the Shanghai *Shiqu* (urban centre) variety and the other Shanghai local dialectal varieties, arranged in a similar way as in **b**.

Data availability

The data that support the findings of this study are available from the following third parties. The word entries in the dataset of old to recent English-German related words were from Wiktionary (Flippo, 2018; Kroch, 2020; Wiktionary, 2022), which were proof-read and annotated according to Cambridge Dictionary (2022), Collins Dictionary, Conjugator (2022), Harper (2022), Linas (2022), and Verbix (2022), mainly using British pronunciations and modern German pronunciations. The dataset of *thirty-year sliced morphemic transcriptions for Chinese dialects in Shanghai* was published in print (You, 2013), which was digitised and cross-checked by the first author. However, restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are, however, available from the authors upon reasonable request and with permission of the above-cited third parties. The LME modelling data (including model estimates, standard errors, degree of freedom, *t*-values, *p*-values, *R* squares) that support the findings of this study are available in Open Science Framework [https://osf.io/xgu2y/][https://doi.org/10.17605/OSF.IO/XGU2Y]. In addition, traditional Chi-square analyses and Cramer's V statistics (not reported in the paper) are also provided in the same repository.

Code availability

The code that support the findings of this study are available in Open Science Framework with the identifier(s) [https://doi.org/10.17605/OSF.IO/56VT8].

Received: 23 March 2023; Accepted: 24 July 2023;
Published online: 02 August 2023

References

Allaire JJ, Gandrud C, Russell K, Yetman CJ (2017) networkD3: D3 JavaScript Network Graphs from R
Anttila R (1977) Analogy. Mouton, The Hague
Baidu Maps (2022a) Baidu Coordinate Pickup System. https://api.map.baidu.com/lbsapi/getpoint/. Accessed 21 Nov 2022

Baidu Maps (2022b) Baidu Maps. https://map.baidu.com/. Accessed 23 Nov 2022
Bailey CG (1997) The etymology of the old high German weak verb. PhD Thesis, Newcastle University
Bates D, Maechler M, Bolker B, Walker S (2013) lme4: linear mixed-effects models using Eigen and S4. R package version 1.0-4
Baxter WH (1992) A handbook of Old Chinese phonology. Walter de Gruyter
Beekes RSP, Vaan MACde (2011) Comparative Indo-European linguistics: an introduction, 2nd edn. John Benjamins Pub. Co, Amsterdam; Philadelphia
Bialystok E (2009) Bilingualism: The good, the bad, and the indifferent. Biling Lang Cogn 12:3–11. https://doi.org/10.1017/S1366728908003477
Bowern C (2018) Computational phylogenetics. Annu Rev Linguist 4:281–296. https://doi.org/10.1146/annurev-linguistics-011516-034142
Cambridge Dictionary (2022). https://dictionary.cambridge.org/. Accessed 8 Aug 2022
Chao YR, Zipf GK (1949) Human behavior and the principle of least effort: An introduction to human ecology. Language 26:394
Chen M (1973) Cross-dialectal comparison: a case study and some theoretical considerations. J Chin Linguist 1:38–63
Chen Z (2019) Kāibù yǐlái Shànghǎi chéngshì fāngyán yǔyīn yǎnbiàn (开埠以来上海城市方言语音演变. The evolution of phonetics in Shanghai urban dialect since its opening). Yuyan Yanjiu Jikan (语言研究集刊 J Lang Res) 280:313–433
Collins Dictionary (2022) https://www.collinsdictionary.com/dictionary/german-english. Accessed 8 Aug 2022
Confucius (551BC–479BC) Shù'ér piān · Dì shí-bā zhāng: Zǐ suǒ yǎyán, 《Shī》、《Shū》、zhí lǐ, jiē yǎyán yě (述而篇·第十八章: 子所雅言, 《诗》、《书》、执礼, 皆雅言也. Chapter 18 of the "Shu'Er Pian" section: The Master spoke in Elegant Speech. The 'Book of Songs', the 'Book of History', and the practices of ritual propriety were all carried out in Elegant Speech). In: Lun-Yu (论语. Analects)
Conjugator German verb conjugation (2022) https://conjugator.reverso.net/conjugation-german.html. Accessed 8 Aug 2022
Dijkstra T, Van Heuven WJB (1998) The BIA model and bilingual word recognition. In: Grainger J, Jacobs AM (eds) Localist connectionist approaches to human cognition. Lawrence Erlbaum Associates Publishers, Mahwah, New Jersey, pp. 189–225
Dixon RMW (1997) The rise and fall of languages. Cambridge University Press, Cambridge, U.K
Dyen I (1963) Why phonetic change is regular. Language 39:631. https://doi.org/10.2307/411958
Edkins J (1853) A grammar of colloquial Chinese as exhibited in the Shanghai dialect, 1st edn. Presbyterian Mission Press, Shanghai
EF Education First (2022) EF English proficiency index. https://www.ef.com/assetscdn/WIBIwq6RdJvcD9bc8RMd/cefcom-epi-site/reports/2022/ef-epi-2022-english.pdf. Accessed 20 Jun 2023
Ferguson CA (1959) Diglossia. WORD 15:325–340. https://doi.org/10.1080/00437956.1959.11659702
Flippo H (2018) Common English-German cognates. https://www.thoughtco.com/common-english-german-cognates-4077037. Accessed 8 Aug 2022
Galton F (1879) Psychometric experiments. In: Brain
Gneuss H (1990) The study of language in Anglo-Saxon England. Bull John Rylands Univ Libr Manchester 72:3–32
Gooskens C, van Heuven VJ, Golubović J et al. (2018) Mutual intelligibility between closely related languages in Europe. Int J Multiling 15:169–193. https://doi.org/10.1080/14790718.2017.1350185
Green DG, Sadedin S, Leishman TG (2008) Self-organization. In: Encyclopedia of Ecology. Elsevier, pp. 3195–3203
Grimm J (1967) Germanic grammar. In: Lehmann WP (ed) A reader in nineteenth-century historical Indo-European Linguistics. Indiana University Press, Bloomington and London
Gu G (2023) Hànyǔ zú yǔyīn jùlí: Yīnlè fǎ-Yījù "yǔyīn duìyì zhěngqí chéngdù" de Hànyǔ lìshǐ guānchá gōngjù (汉语族语音距离: 音类法——依据"语音对应整齐程度"的汉语历史观察工具. Chinese language family phonetic distance: phonemic method—a Chinese historical observation tool based on the 'degree of phonetic correspondence'). http://sino.kaom.net/si_net_yin_note.php
GUO Pu (276–324a) Gài wén 《Fāngyán》 zhī zuò, chū hū yóuxuān zhī shǐ, suǒ yǐ xúnyóu wàn guó, cǎi lǎn yì yán, chē guǐ zhī suǒ jiāo, rénjì zhī suǒ dǎo, mǐ bù bì zài, yǐwéi còují (盖闻《方言》之作, 出乎轓轩之使, 所以巡游万国, 采览异言, 车轨之所交, 人迹之所蹈, 靡不毕载, 以为奏籍. It is said that the compilation of "Fangyan" originated from the imperial envoy's chariot journeys, undertaking voyages across countless territories to collect and explore various dialects. Wherever the wheels of the chariot rolled, wherever people traveled, every minute detail was meticulously recorded and compiled for documentation). In: Fangyan-Xu (方言·序. Preface to the Study of Dialects)
GUO Pu (276–324b) Lèi lí cí zhī zhǐ yùn, yòng guāi tú ér tóng zhì (类离辞之指韵, 用乖途而同致. Analyze the sound patterns across distinct dialects to draw analogies; trace diverging paths that ultimately converge to a common source). In: Fangyan-Xu (方言·序. Preface to the Study of Dialects)
Harper D Etymonline (2022) https://www.etymonline.com/. Accessed 8 Aug 2022
Hartigan JA, Wong MA (1979) Algorithm AS 136: a K-means clustering algorithm. J R Stat Soc Ser C (Appl Stat) 28:100–108. https://doi.org/10.2307/2346830
Hayden D (2017) Language and linguistics in medieval Europe. In: Oxford research encyclopedia of linguistics. Oxford University Press, Online
Hickey R (2012) Assessing the role of contact in the history of English. In: Nevalainen T, Traugott EC (eds) The Oxford handbook of the history of English, 1st edn. Oxford University Press, pp. 485–496
Ho D-A (2015) Chinese dialects. In: Wang WS-Y (ed.) The Oxford handbook of Chinese linguistics. Oxford University Press, New York, pp. 149–159
Jacobson R (1971) Über die phonologischen Sprachbünde (On the phonological Sprachbunds). In: Selected Writings I. Mouton, The Hague-Paris
Jespersen O (1909) A modern English grammar on historical principles (Part I: Sounds and Spellings). Winter, Heidelberg
Jiangsu Sheng he Shanghai shi fangyan diaocha zhidao zu (江苏省和上海市方言调查指导组. Jiangsu Province and Shanghai City Dialect Investigation Guidance Group) (1960) Jiāngsū shěng hé Shànghǎi shì fāngyán gàikuàng (江苏省和上海市方言概况. Overview of Dialects in Jiangsu Province and Shanghai Municipality). Jiangsu Renmin Chubanshe (江苏人民出版社, Jiangsu People's Publishing House), Jiangsu
Karlgren B (1915–1916) Études sur la phonologie chinoise. Leyde et Stockholm, Stockholm. Chinese edition: Zhao, YR et al (1994) Zhōngguó yīnyùnxué yánjiū (中国音韵学研究. Studies on Chinese phonology. trans: Zhao,YR, Luo, C, Li, F-K). Shangwu Yinshuguan (商务印书馆. The Commercial Press), Beijing
Kroch A (2020) Examples of Grimm's Law (not exhaustive!). https://www.ling.upenn.edu/~kroch/courses/lx411/handouts/GRIMM.pdf. Accessed 8 Aug 2022
Kroll JF, Sholl A (1992) Lexical and conceptual memory in fluent and nonfluent bilinguals. Adv Psychol 83:191–204
Labov W (1963) The social motivation of sound change. Word. 19(3): 273–309
Labov W (1994) Principles of language change, Vol.1: internal factors. Blackwell, Oxford & Cambridge
Li Z (2019) Jindài Hànyǔ guānhuà fāngyán yùnshū yùn-tú wénxiàn jíchéng (近代汉语官话方言韵书韵图文献集成. The compilation of rhyme books and rhyme charts for modern Mandarin and dialects in Late Imperial China), 1st edn. Shangwu Yinshuguan (商务印书馆. The Commercial Press), Beijing
Linas Etymologeek (2022) https://etymologeek.com/. Accessed 27 Nov 2022
Machan TW (2012) Language contact and linguistic attitudes in the Later Middle Ages. In: Nevalainen T, Traugott EC (eds) The Oxford handbook of the history of English, 1st edn. Oxford University Press, pp. 518–527
Martinet A (1952) Function, structure, and sound change. Word 8:1–32. https://doi.org/10.1080/00437956.1952.11659416
Mazaudon M, Lowe JB (1993) Regularity and exceptions in sound change. In: Domenici M, Demolin D (eds) Annual Conference of the Linguistic Society of Belgium. Bruxelles, Belgium
Meillet A, Ford GB (1967) The comparative method in historical linguistics. H. Champion, Paris
Mufwene SS (2001) The ecology of language evolution, 1st edn. Cambridge University Press
National Geomatics Center of China (2022) National Fundamental Geo-information metadata. https://www.ngcc.cn/ngcc/html/1/391/392/16114.html. Accessed 1 Aug 2022
Ohala JJ (1989) Sound change is drawn from a pool of synchronic variation. In: Language change: contributions to the study of its causes. pp. 173–198
Pan W, Zhang H (2015) Middle Chinese phonology and Qieyun. In: Wang WS-Y (ed.) The Oxford handbook of Chinese linguistics. Oxford University Press, New York, pp. 80–90
Poplack S, Sankoff D, Miller C (1988) The social correlates and linguistic processes of lexical borrowing and assimilation. Linguistics 26:47–104
Pulleyblank EG (1998) Qieyun and Yunjing: the essential foundation for Chinese historical linguistics. J Am Orient Soc 118:200–216
Qian N (2003) Shànghǎi yǔyán fāzhǎnshǐ (上海语言发展史. The history of language development in Shanghai), 1st edn. Shanghai Renmin Chubanshe: Shiji Chuban Jituan (上海人民出版社: 世纪出版集团. Shanghai People's Publishing House: Century Publishing Group), Shanghai
R Core Team (2022) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
Ricci M (1552–1610) Xuéhuìle guānhuà, kěyǐ zài gè shěng shǐyòng, jiù lián fùrú dōu néng yòng guānhuà gēn wàishěng rén jiāotán (学会了官话, 可以在各省使用, 就连妇孺都能用官话跟外省人交谈. Once one learns Mandarin Chinese, they can use it in various provinces, and even women and children can communicate with people from other provinces in Mandarin). In: Della Entrata Della Compagnia Di Gesù e Christianita Nella China. Chinese edition: LIU J, Wang Y (1986) Li Ma-tou chuan chi (利瑪竇全集: 利瑪窦中國傳教史, Complete works of Fr. Matteo Ricci, S.J: Matteo Ricci's History of

Missionary Work in China. trans: LIU J, WANG Y). Guangqi Chubanshe (光啟出版社. Gong-Chi Publishing Co., Ltd)
Sagart L, Jacques G, Lai Y et al. (2019) Dated language phylogenies shed light on the ancestry of Sino-Tibetan. Proc Natl Acad Sci USA 116:10317–10322. https://doi.org/10.1073/pnas.1817972116
Schleicher A (1967) Introduction to a compendium of the comparative grammar of the Indo-European, Sanskrit, Greek and Latin languages. In: Lehmann WP (ed) A reader in nineteenth-century historical Indo-European Linguistics. Indiana University Press, Bloomington and London
Senn S (2011) Francis Galton and regression to the mean. Significance 8:124–126. https://doi.org/10.1111/j.1740-9713.2011.00509.x
Standing Committee of the National People's Congress (2000) Zhōnghuá Rénmín Gònghéguó guójiā tōngyòng yǔyán wénzì fǎ (中华人民共和国国家通用语言文字法. Law on the standard spoken and written Chinese language of the People's Republic of China)
Swadesh M (1955) Towards greater accuracy in lexicostatistic dating. Int J Am Linguist 21:121–137. https://doi.org/10.1086/464321
Szmrecsanyi B (2012) Analyticity and syntheticity in the history of English. In: Nevalainen T, Traugott EC (eds.) The Oxford handbook of the history of English, 1st edn. Oxford University Press, pp. 654–665
Tang C, Van Heuven VJ (2009) Mutual intelligibility of Chinese dialects experimentally tested. Lingua 119:709–732. https://doi.org/10.1016/j.lingua.2008.10.001
Thomason SG (2011) Language contact: an introduction. Repr. Edinburgh Univ. Press, Edinburgh
Thompson JN, Rafferty JP (2020) Coevolution. Encyclopedia Britannica
Trudgill P (1986) Dialects in contact. B. Blackwell, Oxford, UK; New York, NY, USA
Trudgill P (1974) Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography. Lang Soc 215–246. https://doi.org/10.1017/S0047404500004358
van der Loo MPJ (2014) The stringdist package for approximate string matching. R J 6:111–122. https://doi.org/10.32614/RJ-2014-011
Venables WN, Ripley BD (2002) Modern applied statistics with S, Fourth. Springer, New York
Verbix (2022) Verbix verb conjugator. https://verbix.com/. Accessed 8 Aug 2022
Verner K (1967) An exception to the first found shift. In: Lehmann WP (ed) A reader in nineteenth-century historical Indo-European Linguistics. Indiana University Press, Bloomington and London
Wang WS-Y (1969) Competing changes as a cause of residue. Language 45:9. https://doi.org/10.2307/411748
Wang H (2010) Céngcì yǔ duànjiē-Diézhìshì yīnbiàn yǔ kuòsànshì yīnbiàn de jiāochā yǔ qūbié (层次与断阶——叠置式音变与扩散式音变的交叉与区别. Strata and steps-broken: The overlap and differences between external phonological superposition and internal phonological diffusion). Zhongguo Yuwen (中国语文 Studies of the Chinese Language) 314-320+383-384
Wang Y (2023) Nánběi Cháo Suí Táng Sòng fāngyánxué shǐliào kǎolùn (南北朝隋唐宋方言学史料考论. Research on historical materials of dialectology during the Southern and Northern dynasties and Sui, Tang, and Song dynasties). Kexue Chubanshe (科学出版社. Science Press)
Weinreich U (1953) Languages in contact: finding and problems. Mouton, The Hague
Wiktionary (2022) Appendix: German cognates with English. In: Wiktionary. https://en.wiktionary.org/wiki/Appendix:German_cognates_with_English. Accessed 8 Aug 2022
Wu J, Chen Y, van Heuven VJJP, Schiller NO (2016) Predicting tonal realizations in one Chinese dialect from another. Speech Commun 76:1–27. https://doi.org/10.1016/j.specom.2015.10.006
Wu J, Zheng W, Han M, Schiller NO (2021) Cross-dialectal novel word learning and borrowing. Front Psychol 12:734527. https://doi.org/10.3389/fpsyg.2021.734527
Xu B, Tang Z (1988) Shànghǎi shìqū fāngyán zhì (上海市区方言志. Gazetteer of dialects in Shanghai urban area). Shanghai Jiaoyu Chubanshe (上海教育出版社. Shanghai Education Press, Shanghai
You R (ed.) (2013) Shànghǎi dìqū fāngyán diàochá yánjiū (上海地区方言调查研究. A linguistic survey of Shanghai dialects), 1st edn. Fudan Daxue Chubanshe (复旦大学出版社, Fudan University Press), Shanghai
Zhang M, Yan S, Pan W, Jin L (2019) Phylogenetic evidence for Sino-Tibetan origin in northern China in the late neolithic. Nature 569:112–115
Zhao Y (1956) Xiàndài Wúyǔ de yánjiū (现代吴语的研究. A study of modern Wu dialect, 1928), 2nd edn. Kexue Chubanshe (科学出版社.Science Press), Beijing

Acknowledgements

This work was supported by Chinese Fundamental Research Funds for the Central Universities (2017ECNU-YYJ017), and by Shanghai Philosophy and Social Sciences Fund (2017BYY001, finished in 2019). We would like to thank LIU Haidong from Sun Yat-Sen University for a discussion about the mathematical modelling of the evolutionary trajectories in the MDS space, and GU Qianping from Tokyo University for information about the English-German word list. We would also like to express our gratitude to the China Social Science Foundation for reviewing our research proposal, even though they repeatedly concluded that it did not warrant funding.

Author contributions

JW raised the research question, designed the study, prepared the datasets, wrote the codes, carried out the analyses, as well as conceptualised and wrote the article. JZ raised the idea to quantify systematic correspondence by measuring the angle between two vectors and proposed a preliminary mathematical model in a course project supervised by JW.

Competing interests

The author(s) declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-023-01975-6.

Correspondence and requests for materials should be addressed to Junru Wu.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.