Available online at www.sciencedirect.com

ScienceDirect

Lingua 287 (2023) 103512

Lingua

www.elsevier.com/locate/lingua

Comparing Germanic, Romance and Slavic: Relationships among linguistic distances

Wilbert Heeringa [a,c], Charlotte Gooskens [b,c,*], Vincent J. van Heuven [a,c,d,e]

[a] Fryske Akademy, P.O. Box 54, 8900 AB Leeuwarden, The Netherlands
[b] School of Humanities, Arts, and Social Sciences, University of New England, Australia
[c] Center for Language and Cognition, Groningen University, P.O. Box 716, 9700 AS Groningen, The Netherlands
[d] Multilingualism Doctoral School, University of Pannonia, Egyetem u. 10, 8200 Veszprém, Hungary
[e] Leiden University Centre for Linguistics, P.O. Box 9515, 2300 RA Leiden, The Netherlands

Received 13 September 2022; revised 20 February 2023; accepted in revised form 20 February 2023;

Abstract

Languages differ along multiple dimensions (lexis, phonology, morphology, syntax). Related languages descend from a common ancestor language but have diverged over time. This paper asks whether languages diverge equally along all dimensions, and, to the extent that they do not, which dimension reflects the traditional language family tree best. We computed measures of (i) lexical distance (ii) phonetic distance, and (iii) syntactic distance. The measures were computed on all words and sentences extracted from a corpus of translations of four relatively short English texts into another four Germanic languages (Danish, Dutch, German, Swedish), five Romance languages (French, Italian, Portuguese, Romanian, Spanish) and six Slavic languages (Bulgarian, Croatian, Czech, Polish, Slovakian, Slovenian). We examined the correlation structure of the distances for all pairs of Germanic (10), Romance (10) and Slavic (15) languages (i.e., within-family comparisons only). The results indicate that the linguistic dimensions are generally correlated (weakly but significantly), and that the correlations are stronger for pairs within families than when all 35 pairs are examined together. Cladistic family trees correlate best with the lexical distance ($0.851 < r < 0.887$). This confirms that the genealogical language trees are predominantly based on lexical rather than phonetic or syntactic considerations.

Keywords: Linguistic distance; Lexical distance; Levenshtein distance; Syntactic distance; Genealogical distance; Language phylogeny

* Corresponding author.
E-mail addresses: wheeringa@fryske-akademy.nl (W. Heeringa), c.s.gooskens@rug.nl (C. Gooskens), V.J.J.P.van.Heuven@hum.leidenuniv.nl (V.J. van Heuven).

https://doi.org/10.1016/j.lingua.2023.103512
0016-7037/© 2023 The Author(s). Published by Elsevier B.V.

1. INTRODUCTION

1.1. Linguistic distance and language change

Languages differ from one another, by definition. They do not differ along a single dimension but along many. It is customary to divide the description of language structure into a limited number of grammatical components or levels of analysis, viz. sound structure (phonology), morphology, lexis, syntax, and semantics, see, e.g., Crystal (2008).

Moreover, the extent to which languages differ is highly variable. We may have the intuition that Dutch and German differ less from each other than either differs from Mandarin. The traditional approach taken by linguists to differentiate between languages is to try and establish family relationships between pairs of languages, and trace the origin of existing languages back to a common ancestor language. Some comparative linguists even argue that all languages have the same origin and that it may be possible to reconstruct an even older common ancestor that might lie at the root of Dutch, German and, for instance, Mandarin (e.g., Ceolin et al., 2021; Greenberg, 1987; Ruhlen, 1994). However, according to Trask (1996:207) there is little support among linguists for most of the higher order groupings of human languages.

Our current knowledge about genealogical relations among languages is based on the classical comparative method (for an overview see, e.g., Trask, 1996:Chapter 8). The division of related languages into subgroups is accomplished by finding shared linguistic innovations that differentiate them from the parent language. The family trees that have been proposed by comparative linguists are largely based on phonological correspondences between languages. Affiliation trees can be constructed from the degree of correspondences in the phonological forms of cognate words between pairs of languages, as long as the sound shapes used by different languages for the same concept are sufficiently numerous, regular, and systematic, and can thus be assumed to bear a non-accidental relationship. It is possible for languages to have different degrees of relatedness depending on the number of shared linguistic innovations. Given, however, that languages do not only differ in the sound shapes of their vocabulary but in all the grammatical components we mentioned, is it reasonable to expect languages to differ from one another to the same degree in all components? Is it true that if languages A and B differ less from each other than from language C in terms of their sound structure, they will also differ less in terms of their morphology, syntax and semantics? If there were a good deal of parallelism between the components, then the traditional comparative method is safe: study one component, and the degree of similarity in the other components is predictable. In the present study we will correlate linguistic distances at the level of phonology, vocabulary and syntax to cladistic distances, i.e., the distances that can be calculated between language pairs in a language tree.

At first sight, it seems reasonable to expect languages to diverge over time from the ancestor language, and from each other, along all dimensions discussed above. Language is a function of the history of its speakers and the varied situations in which speakers may find themselves and therefore the rate of change of languages may vary. It is also known that linguistic change occurs at different speeds at different linguistic levels and that there are differences in the rapidity of change within linguistic levels. Phonetic changes come about faster and tend to be more pervasive than changes in, for example, syntax or in the high-frequency part of the vocabulary (e.g., irregular verbs and function words tend to remain stable over centuries), see Trudgill (2020). Such differences in rate of change can be taken into account, and will be abstracted away from if parallelism in the extent of change in grammatical components between language pairs is expressed in correlation coefficients. It is also well documented that changes in one grammatical domain may trigger changes in other domains (Trips, 2022). For instance, if a language loses its case system, as happened in the development of classical Latin into the more recent Romance languages, word order becomes more fixed, probably because thematic roles have to be expressed in language one way or another. Such causal trade-offs are less obvious between phonological and syntactic developments but even if there is no communicative pressure for changes in one domain to trigger changes in other domains, there is no reason why language change should be restricted to just one or two grammatical domains. This leads to the question we would like to answer in our study, viz. to what extent is linguistic distance between (pairs of) genealogically related languages correlated across the linguistic domains of phonology, vocabulary and syntax?

The present study is an attempt to come to grips with this problem of parallelism in grammatical components when comparing genealogically related language families (Germanic, Romance and Slavic). In the course of a large-scale study of the cross-lingual intelligibility, both in the spoken and written modality, of a number of related European languages, we collected distance measures between 70 pairs of spoken languages in each of three linguistic domains, viz. lexis (in terms of cognate vocabulary), phonology (segmental differences between cognate word pairs) and syntax (correlation between trigram frequencies of parts-of-speech), see Gooskens and Van Heuven, 2017, 2020; Gooskens et al., 2018; Van Heuven et al., 2015.

1.2. Earlier work on the relationship between linguistic domains

Several studies exist that quantify relationships among linguistic levels. Gooskens and Heeringa (2006) measured lexical, pronunciation and prosodic distances among 15 local Norwegian dialects. They found a correlation of $r = 0.49$ between lexical distances and pronunciation distances, $r = 0.43$ between pronunciation and prosodic distances, and $r = 0.18$ between lexical and prosodic distances.

Spruit et al. (2009) measured the degrees of association among aggregate phonetic, lexical and syntactic differences in 70 Dutch dialect varieties. They found that pronunciation is marginally more strongly associated with syntax ($r = 0.65$) than with lexis ($r = 0.62$) and that syntax and lexis are more loosely associated ($r = 0.50$). Grieve (2013) compared common patterns of regional phonetic and lexical variation in American English dialects and found that the two levels follow similar patterns. We are not aware of attempts to compare the correlations between linguistic domains across language families. Therefore, the question remains whether relationships among linguistic levels are universal, or whether they differ per language group or even per language.

Heeringa and Hinskens (2014) computed lexical, phonetic and morphological distances among 86 local Dutch dialects. The data were collected in the period from 2008 until 2011. Two older male speakers and two younger female speakers were recorded in each of the 86 locations. The authors focus on dialect change and do not provide the reader with the correlations of the measurements among the three levels. Fortunately, it is possible to calculate the correlations from their data. The correlation between the lexical distances and the distances in the phonology ("sound components" in Heeringa and Hinskens, 2014) is $r = 0.46$ (older male speakers) and $r = 0.47$ (younger female speakers). The correlation between the lexical distances and the morphological distances is $r = 0.37$ (older male speakers) and 0.46 (younger female speakers). The highest correlations are found between the distances in the sound components and the morphological distances: $r = 0.49$ (older male speakers) and $r = 0.59$ (younger female speakers). All these correlations are significant at the 0.01 level.

The three studies focus on a few linguistic levels and on small areas within the Germanic language area. They include lexis and phonology. Gooskens and Heeringa (2006) also include prosody, while Spruit et al. (2009) include syntax and Heeringa and Hinskens (2014) include morphology.

In the present investigation we measure linguistic distances among standard languages spoken in sixteen European countries. The distances are assessed at the lexical, phonetic, and syntactic level. On the basis of the measurements, we can classify modern European languages and compare the classifications to traditional classifications of the three language families. Traditional language classifications mostly reflect the historical language situation and are based on lexical and phonological characteristics of the languages with the specific purpose of showing the genealogic relationships among the languages. We, in contrast to this, used a random selection of words from modern texts in 16 European languages and based our measurements on objective dialectometric measurements that reflect the present-day relationships among the languages. Computing distances at various linguistic levels allow us to correlate distances at different levels with each other (intercorrelations) and to examine the extent to which the distances at the various levels express the same relations among the languages.

We focus on three Indo-European language groups: Germanic, Romance and Slavic and investigate to what extent the relationships among linguistic levels are the same across those three groups. To summarize, the following research questions are addressed in the present paper:

1. To what extent do the distances at various linguistic levels correlate with each other?
2. Are the relationships among linguistic levels the same for Germanic, Romance and Slavic?
3. How well do distances among modern languages reflect traditional classifications?

Our paper is organized as follows. The remainder of the present section reviews literature on the genealogical relationships among the languages we selected. In Section 2 we discuss our data set and the way distances among languages were measured at the three linguistic levels, viz. lexis, phonology, and syntax. In Section 3, we present the results for each of the language groups, and per language groups for each of the three linguistic levels. The results are compared with traditional classifications in Section 4. Additionally, the relationships among the linguistic levels are compared across the three language groups. We finish the paper with conclusions and discussion in Section 5.

1.3. Classifications of the languages in the investigation

We will now give a brief overview of classifications known as cladistic trees or family trees of the 15 modern languages in the investigation (see Figs. 1 and 2).

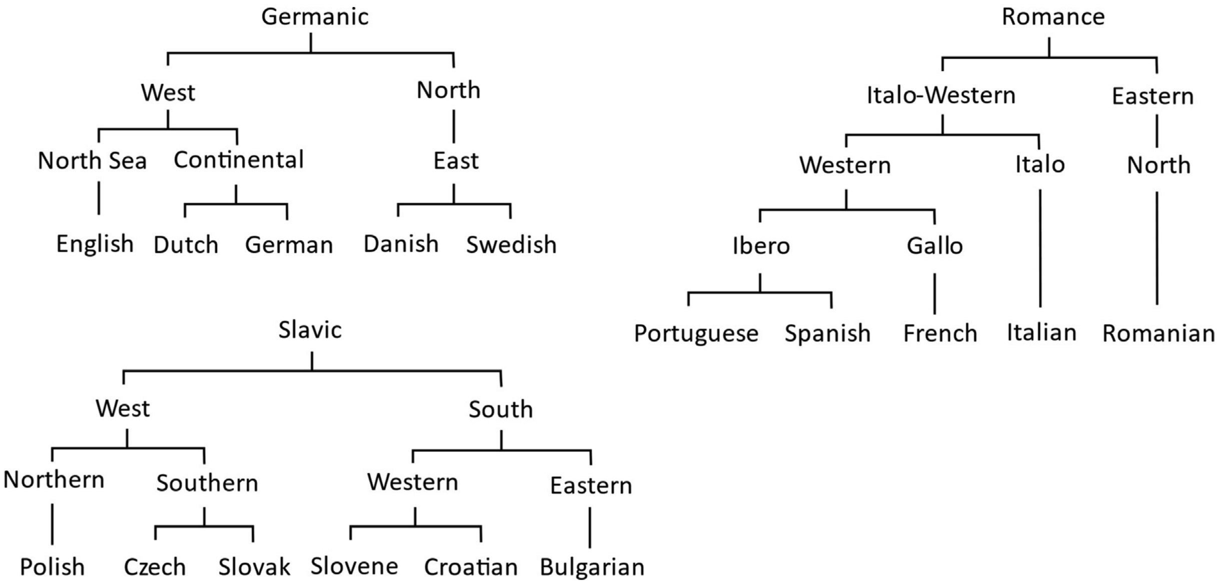


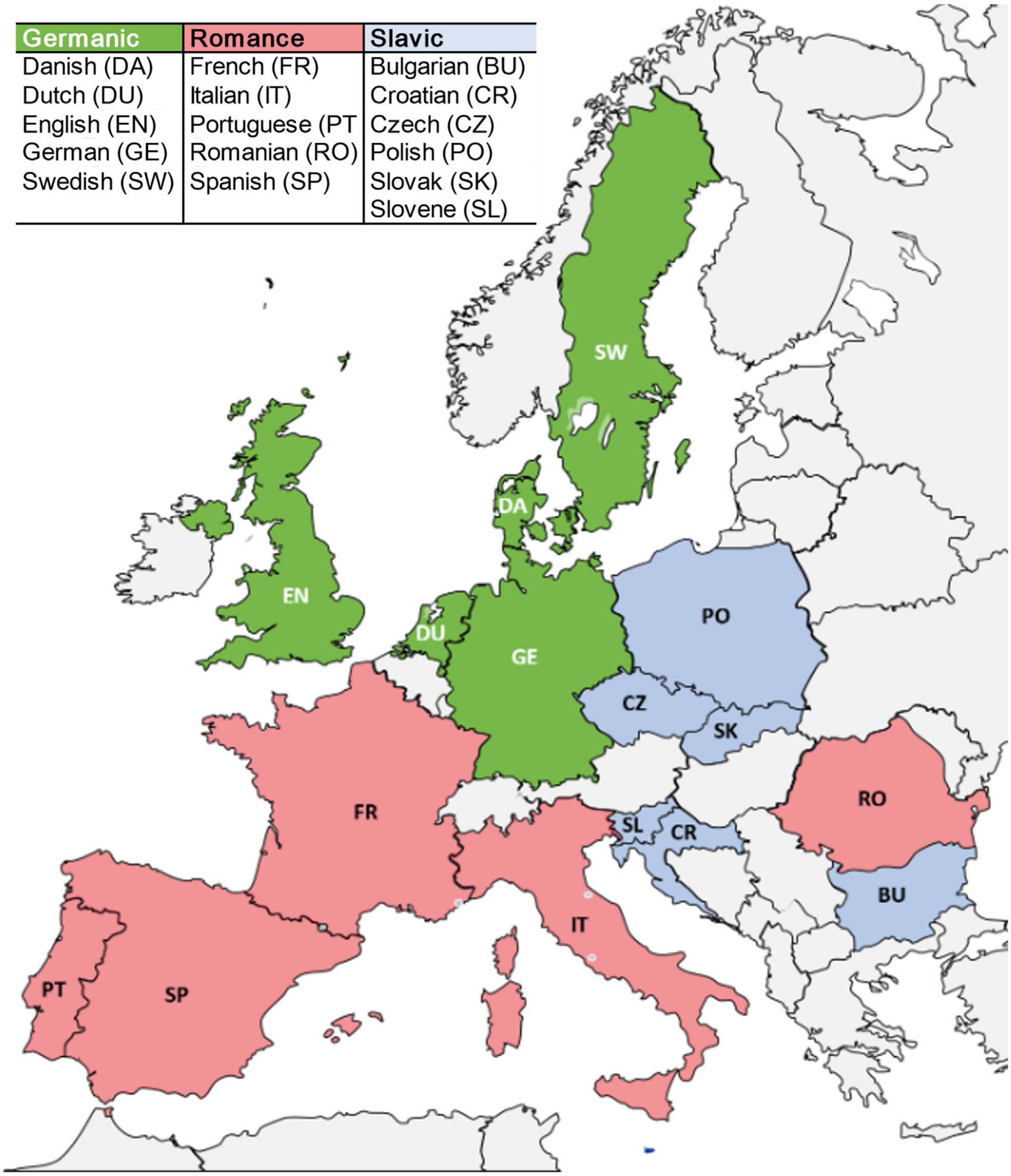
Fig. 1. Cladistic trees for Germanic, Romance and Slavic languages in sample, as proposed by Harbert (2007), Hall (1974), and Sussex and Cubberley (2006), respectively. Adapted from Gooskens and Van Heuven (2020).

1.3.1. Germanic

The five Germanic languages in the investigation are traditionally divided into two branches of the Germanic language family tree, the North Germanic branch (Danish and Swedish) and the West Germanic branch (Dutch, English and German), see e.g., Hendriksen and Van der Auwera (1994:4); Kufner (1972).

The North Germanic languages originate from Old Norse, a language that was spoken during the Viking Age (800–1050). At the end of the Viking Age, Old Norse split up into East (spoken in Denmark and Sweden) and West Scandinavian (Norway, Iceland and the Faroe Islands). During the Middle Ages, Danish grew rather distinct from Swedish due to influences from the south (Torp, 2002). Especially the Danish phonology changed considerably. The most salient differences between Danish and Swedish are the fact that Danish has a glottal stop (*stød*) where Swedish has a lexical tone contrast, and that unvoiced plosives changed in voiced plosives after a vowel in Danish (Danish *hedde* vs. Swedish *heta* ‘be called’). Furthermore, unstressed suffixes changed to schwa in Danish but not in Swedish (Danish *skole* /sgo:lə/ vs. Swedish *skola* /sku:la/ ‘school’, Braunmüller, 1999; Vikør, 2002). In spite of the phonological differences between Swedish and Danish, the two languages are still so similar that they are only considered independent languages because of their position as standardized languages spoken within the limits of a state. Differences at other linguistic levels are small. Danish and Swedish share over 90 percent of their lexicons (Bergman, 1979; Gooskens, 2007) and, apart from small differences in the ending of plurals and nouns, the Danish and Swedish morpho-syntactic systems are very similar.

The parentage of the languages in the West Germanic branch is less clear than in the case of the North Germanic languages. It has been suggested that this branch could originally be split into three dialect groups, Ingwaeonic, Istwaeonic and Herminonic. However, the three languages no longer reflect this division in a straight-forward fashion, due to geographical discontinuity and interference from North Germanic and French in the case of English, and mutual influence between Dutch and German (Harbert, 2007:9; Hendriksen and Van der Auwera, 1994:8). Nowadays, the West Germanic branch is typically split into two sub-branches, which are referred to as Continental Germanic (including German and Dutch) and North Sea Germanic (English), see Hendriksen and Van der Auwera (1994); Ruhlen (1987:327). Dutch and German are closely related but less similar than Swedish is to Danish. The two languages share a large part of their vocabularies but there are a number of differences between their phonological and morpho-syntactic systems. German underwent the High German Consonant shift between the third and ninth century while Dutch (and English) did not. Its effect can be seen by comparing modern German words with their Dutch and English counterparts (in that order, target phones underlined), e.g., [pfunt–pɔnt–pɑund] ‘pound’, [apfəl–ɑpəl–æpəl] ‘apple’, [katsə–kɑt–kæt] ‘cat’, [hɛrts–hɑrt–hɑ:t] ‘heart’ and [maxən–ma:kə–meik] ‘make’. In contrast to Dutch (and English), case is marked on nouns, adjectives and determiners in German.



Germanic	Romance	Slavic
Danish (DA)	French (FR)	Bulgarian (BU)
Dutch (DU)	Italian (IT)	Croatian (CR)
English (EN)	Portuguese (PT	Czech (CZ)
German (GE)	Romanian (RO)	Polish (PO)
Swedish (SW)	Spanish (SP)	Slovak (SK)
		Slovene (SL)

Fig. 2. The languages and their geographic locations included in the investigation. Map adapted from https://commons.wikimedia.org/wiki/User:Kolja21/Maps_of_the_European_Union. Re-use permitted under the GNU Free Documentation License, Version 1.2, published by the Free Software Foundation.

There has been substantial contact between English and other languages throughout history, in particular with French but also with Scandinavian languages. For this reason, English is more distant at all linguistic levels from German and Dutch than these two languages are from each other. Gooskens and Heeringa (2004) measured phonetic distances among a number of Germanic languages, including the five languages in our investigation, and found that English is also phonetically more distant to the other languages than any of the other Germanic languages in the investigation. English is usually classified as a West-Germanic language. It is assumed that English originates from Old English or Anglo-Saxon and that its closest relatives are Old Frisian and Old Saxon. The large number of words with a Scandinavian origin in Middle English is attributed to language contact. However, Emonds and Faarlund (2014) pointed out that modern English syntax is more similar to Scandinavian syntax than to that of the West-Germanic languages and this leads to their conclusion that English is in fact a North Germanic language that was strongly influenced by Anglo-Saxon. However, a number of authors (e.g., Van Gelderen, 2016; Van Kemenade, 2016; Kortmann, 2016; Trudgill, 2016) presented arguments against this conclusion and the classification of English is still an open issue.

1.3.2. Romance

There is considerable disagreement about the subgrouping of the languages in the Romance language family tree and as noted by Bossong (2016) the subdivision is strongly dependent on the criteria used. We will present three different classifications that have been proposed in the literature. The classifications are based on different criteria and therefore it can be expected that they reflect distances on each of the linguistic levels to various degrees.

Von Wartburg's (1936) classification is based mainly on phonetic (and partly orthographic) criteria. It divides the various languages along the La Spezia-Rimini Line, which runs across north-central Italy just to the north of the city of Florence (whose speech forms the basis of standard Italian). The line coincides with a number of important isoglosses that distinguish Romance languages south and east of the line from Romance languages north and west of it. Generally speaking, the West Romance languages show common innovations that are typically absent from East Romance language varieties. Romanian belongs to the East branch along with the languages of central and southern Italy, while the West branch includes Portuguese, Spanish, French and Italian. Within most of the West Romance language area there are traditional dialects between contiguous languages that make further subgrouping somewhat arbitrary.

Rohlfs (1971) looked at the degree of lexical differentiation from Latin, and made a sub-division into Inner and Outer Romance languages (see Dietrich and Geckeler, 2000:17). Italian is least different from Latin and is clustered together with French in the Inner Romance branch, while Portuguese, Spanish and Romanian deviate more and all belong to the Outer Romance branch. Bossong (2016:70) notes that the position of French is ambiguous. On the one hand, it is part of the western continuum and on the other hand it often occupies a special more extreme position.

The classification that is now most widely accepted seems to be one that is based primarily on phonological criteria (Hall, 1974; Lewis, Simons and Fennig, 2015; Ruhlen, 1987). The Eastern branch includes Romanian only, while the other four Romance languages in our sample are subsumed under the Italo-Western branch. Within the latter, Italian forms its own sub-branch, while French, Portuguese and Spanish form a Western Romance cluster. This cluster splits into Gallo-Romance (French) and Ibero-Romance (Portuguese and Spanish). We will use this classification in our analysis of the relationship between linguistic distances and phylogenetic trees.

1.3.3. Slavic

Slavic languages have traditionally been divided into three main branches on the basis of genealogical and geographical principles. This division is still maintained for the modern languages (Jakobson, 1955:1–3; Sussex and Cubberley, 2006:42; Trask, 1996:188). The six Slavic languages in our sample belong to either West Slavic (Czech, Polish, Slovak) or to South Slavic (Bulgarian, Croatian, Slovene). These two branches have developed rather independently throughout history and we expect this to be reflected in the linguistic distances. Sussex and Cubberley (2006:3–58) provide an overview of phonological and morphological features that differentiate the branches and individual languages.

Within the West Slavic branch, Czech and Slovak both belong to the same southern branch of West Slavic and are known to be similar to an extent that the speakers can read and understand a fair amount of each other's languages. Polish belongs to the southern branch and is less similar but still closely related to the other languages.

In the South Slavic branch, Bulgarian entered the Balkan Peninsula via another route than Croatian and Slovene and therefore had contact with other languages. The three South Slavic languages therefore belong to two different sub-branches. Bulgarian is an outlier in the Slavic language family, since it lost case and the infinitive, which also brought about changes in syntax. The word order is generally more constrained in Bulgarian than in the other Slavic languages.

1.3.4. Family trees summarized

On the basis of the proposals summarized in Sections 1.3.1–2–3 we arrive at the cladistic trees in Fig. 1 for the modern languages in each of our three groups.

Distances in these cladistic trees can be computed between each language pair within a family by counting the number of steps one has to go up from one member of the pair to the lowest common node shared between the members of the pair and then down again to the other member. The smallest cladistic distance between any two family members in our data is 2 (as between Portuguese and Spanish). The largest distance in our set is found between Romanian and either Portuguese or Spanish, which is 7. The cladistic distances are listed in the Appendix, Tables B.3 for Germanic, C.3 for Romance, and D.3 for Slavic. In Section 4.2 we will correlate the cladistic tree distances with linguistic distances to investigate how well the three linguistic levels predict the cladistic distances.

2. DATA SOURCE, MEASUREMENT TECHNIQUES AND VISUALIZATION

The material used for the linguistic distance measurements presented in this paper was originally collected with the aim of testing mutual intelligibility among closely related languages in Europe (Golubović, 2016; Gooskens and Van Heuven, 2017, 2020; Gooskens et al., 2018; Swarte, 2016). Fig. 2 shows a map of Europe with an overview of the 16 target languages.

All of the official national languages were included from the three largest (in terms of number of members and speakers) language families in the EU member states, i.e., Germanic (5 languages), Romance (5 languages) and Slavic (6 languages). If the same language is an official language in multiple member states, only the variety from the country with the largest number of speakers was included. For example, German is an official language in both Austria and Germany, but since the largest number of speakers live in Germany, only Northern Standard German was included. We only measured the distances among the languages within each of the three language families.

In Section 2.1 we describe our data set and in Section 2.2 we explain the methods for measuring distances at the various linguistic levels. Once the linguistic distances are calculated, the grouping of the languages is visualized by hierarchical cluster trees, which we explain in Section 2.3.

2.1. Data set

To measure linguistic distances that could be compared across language families we needed recordings of the same texts in the 16 languages. To achieve this end, we selected four short English texts from a set of exercises used at the University of Cambridge to prepare students for the Preliminary English Test (PET).[1] These are culturally neutral texts at the intermediate (B1) level as formulated by the Common European Framework of Reference for Languages (Council of Europe, 2001). We made slight adjustments to the texts, so all four English texts contained ca. 200 words each, with 16 or 17 sentences per text (66 sentences in all).

Native speakers of the target language with some translation training, translated each of the English texts in to the fifteen 15 other test languages (see Fig. 2). A text was first translated by one native speaker and then checked by at least two others, such that the final version was agreed upon by all designated translators. Translators and checkers were instructed to avoid overly literal or unnatural constructions, while still trying to stick to the original English texts (in terms of vocabulary and word order) as much as possible within the syntactic constraints of the target language. In this way we hoped to get texts that were as comparable as possible across languages. The four texts (in English only) can be found in Appendix A.

2.2. Measuring linguistic distances

Within each of the three language families, linguistic distances were measured for all pairs of the five (Germanic, Romance) or six (Slavic) stimulus languages used in the project (35 language combinations in all). Distances were computed at three linguistic levels, viz. lexical, phonetic and syntactic. We will now explain each of the measures.

2.2.1. Lexical distance

To calculate lexical distances, we aligned the words of the 16 texts. The linguistic relationship between two languages may be asymmetrical. A word in language A may have a cognate in language B but a word in language B need not have a cognate synonym in language A. For instance, the Dutch word *plek* ‘place, spot, location’ has no cognate in German. The nearest equivalent for *plek* in German would be *Ort*, which is cognate to Dutch *oord*. We modelled this asymmetry creating 16 tables, one for each language. The first column is the list of words in the source language and in the other four (for the Germanic and Romance languages) or five (for the Slavic languages) columns we put the cognate words (if any) for each target language. Cells in the table remain empty when there is no cognate. As an example, Table 1 shows a selection from the Dutch stimulus list with the cognates in the other four Germanic languages.

Following Séguy (1973), lexical distance was computed for all pairs of languages within families as the percentage of non-cognates (i.e., historically unrelated words). We counted the empty cells in the aligned tables and divided this num-

[1] https://www.cambridgeenglish.org/exams/preliminary.

Table 1
Example of the compilation of the cognate lists. The leftmost column represents the Dutch source list. For each word, cognates in each of the target languages is noted. An empty cell means there is no cognate. Number and percentage of non-cognates is shown in the bottom row.

Dutch source	Danish	English	German	Swedish
kracht	kraft	-	Kraft	kraft
toestemming	-	-	Zustimmung	-
vorm	form	form	Form	form
adem	ånde	-	Adem	anda
kinderen	-	children	Kinder	-
Non-cognates	2 (40%)	3 (60%)	0 (0%)	2 (40%)

ber by the total number of words of the stimulus language and then converted this to percentages. In our example in Table 1 the number and percentages of non-cognates are given in the bottom row.

2.2.2. Phonetic distance

To measure phonetic distances, for practical reasons we only used the texts that were translated from English (and the English text itself). We aligned the words of the 16 texts and transcribed them automatically using speak-ng, a multilingual speech synthesizer (Dunn and Vitolins, 2019). Then the transcriptions were manually corrected by linguists who were familiar with the languages and/or native speakers of those languages.

For each pair of languages, their respective word lists were aligned, and the phonetic distance was calculated as the mean distance in the cognate word pairs. Since not all cognates appear in every language, the set of cognate pairs differs per language. For example, English *children* is translated as *kinderen* in Dutch, *Kinder* in German, *børn* in Danish and *barn* in Swedish. Here English, Dutch and German share cognates with each other, and the two Scandinavian languages share cognates with each other, but Danish and Swedish do not share cognates with the other three languages.

For the Germanic language sample, the number of cognate pairs varies between 203 and 405, for Romance between 272 and 500, and for Slavic between 194 and 559. For each set we calculated Cronbach's α, which is a function of the number of items and the average inter-item correlation. In our case, it is a function of the number of words and the average correlation among the respective distance matrices (which contain the phonetic distances between all pairs of languages) – at the word level. Its values range between zero and one, higher values indicating greater reliability. As a rule of thumb, values higher than 0.7 are considered sufficient in social sciences to obtain consistent results (Nunnally, 1978). For Germanic, Romance and Slavic we found Cronbach's α values of 0.997, 0.994 and 0.999, respectively.

The phonetic distance between cognates was computed by the Levenshtein algorithm. This distance is defined as the penalty incurred by the least costly set of edit operations needed to convert the string of phonetic symbols in language A to its cognate transcription in B. Three string edit operations are possible: insertion, deletion and substitution of a symbol. Ignoring diacritics to simplify the example, Table 2 illustrates the algorithm, converting the English word *interest* to its Swedish cognate *intresse*.

In slot 4 /ə/ is deleted, in slot 6 /ɛ/ is substituted for /e/, in slot 8 /ə/ is inserted, while /t/ is deleted in slot 9. The length-normalised Levenshtein distance is then computed by dividing the total costs (i.e., 4 edits) by the alignment length (number of alignment slots, i.e., 9). The word *interest* can be mapped onto *intresse* in many different ways, but the Levenshtein distance always gives the cost of the cheapest mapping. The minimum cost, however, must be based on an alignment in which vowels match with vowels and consonants match with consonants; /j, i, w, u/ can be aligned with vowels and consonants, while [ə] and [ɐ] can be aligned with any vowel or sonorant consonant. As there are four operations and the alignment has nine slots, the normalised Levenshtein distance in the example is (4 / 9) × 100 = 44%.

Table 2
Alignment showing the mapping of the pronunciation of English 'interest' onto the pronunciation of Swedish 'intresse' according to the Levenshtein distance.

	1	2	3	4	5	6	7	8	9
English	ɪ	n	t	ə	r	e	s		t
Swedish	ɪ	n	t		r	ɛ	s	ə	
edits				del.		sub.		ins.	del.
cost	0	0	0	1	0	1	0	1	1

In the example in Table 2, substitutions, insertions and deletions have a cost of 1, and matches have a cost of 0. This approach does not reflect that, for example, the pair [i, o] is seen as being more different than the pair [i, ɪ]. Following Wieling et al. (2009) and Wieling (2012), we used the more sophisticated pointwise mutual information (PMI) Levenshtein solution, which employs automatically obtained, linguistically sensitive sound segment distances based on how frequently sound segments co-occur in the alignments.[2] The overall phonetic distance from language A to language B is the arithmetic mean of the length-normalized PMI-weighed distances for all cognate word pairs.

When calculating phonetic distances, primary stress is processed. Normally, a primary stress character is found immediately before the first segment of a stressed syllable. We move the stress character to before the first vowel in the syllable. Subsequently, the stress character is processed as an insertion or deletion.

We process length 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].

However, PMI Levenshtein stumbles over the multiplication of the segments, probably since this prevents a proper calculation of the frequencies of co-occurrences of segments in the alignments. We solved this by first running Levenshtein distance while ignoring length marks and saving the PMI segment distances that were obtained. Subsequently we ran “normal” Levenshtein distance that used the previously calculated PMI segment distances as operation costs.

For practical reasons, only a restricted 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 distance between [t] and [s] is calculated and the distance between [h] and [s] is calculated, and subsequently the average of the two distances is calculated as the final distance. For the other diacritics the procedure is *mutatis mutandis* the same.

2.2.3. Syntactic distance

As mentioned in Section 2.1, four texts were translated from English to all 15 other test languages. The syntax measures were based on the 66 sentences in the four texts (Appendix A). In order to measure syntactic differences, we adopted the method introduced by Nerbonne and Wiersma (2006), which has also been used by Swarte (2016) for the Germanic language sample, and by Golubović (2016) for the corresponding Slavic sample.

The method is based on transitional probabilities between parts-of-speech (PoS). First, the words in the 66 sentences were manually labeled with PoS-tags. We distinguished 16 PoS-tags: abbreviation, adjective, adjectival verb (Slavic languages only), adverb, conjunction, determiner, interjection, modal verb, noun, numeral, particle, preposition, pronoun, verb, a special tag for *to* before infinitive, and sentence boundary.

Subsequently, an inventory of trigrams of PoS-tags across the texts of the different languages is made. Then the number of occurrences for each trigram per language is counted. Thus, a vector of trigram counts for each language is obtained. The syntactic distance between any two languages is then calculated by comparing their respective frequency vectors. Nerbonne and Wiersma (2006:85) write that the “choice of vector difference measure (...) does not affect the proposed technique greatly, and alternative measures can be used straightforwardly.” We followed Swarte (2016) and Heeringa et al. (2018) by calculating the syntactic distance between any two languages as 1 minus the Pearson’s product-moment correlation coefficient (r) found between the two vectors associated with the language pair concerned. We chose this difference measure since it is both easy to implement and easy to interpret.

We illustrate the method with a two-sentence corpus: *It will become easier* and *This is not true*. In Table 4 the PoS-tag trigrams are shown that can be derived from the two sentences. For each sentence four trigrams are found. In Table 5 the inventory of the trigrams on the basis of the two sentences is given. We find seven different types of trigrams. The last column is the frequency vector of trigrams of PoS-tags. Most trigrams occur once but one, i.e., $-pron-mod, appears twice.

Originally, Swarte (2016) computed the syntactic distances between the members of a language pair bi-directionally, i.e., from the original text in language A to its most literal (but still grammatically correct) translation in B, and from the original text B into its most literal translation in A. This means that, like for the lexical distances, two different distances per language pair were obtained, which made it possible to express asymmetric relations. However, we found a very high correlation of $r = 0.97$ ($p < .001$) between the non-directional and the bi-directional syntactic distances. For this rea-

[2] The PMI Levenshtein distances were calculated with LED-A (freely available at: https://www.led-a.org/).

Table 3
Diacritics processed when calculating phonetic distances. For details see text.

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

Table 4
Trigrams of PoS-tags in two sample sentences.

	It	*will*	*become*	*easier*			*This*	*is*	*not*	*true*	
$	pron	mod				$	pron	mod			
	pron	mod	verb				pron	mod	part		
		mod	verb	adj				mod	part	adj	
			verb	adj	$				part	adj	$

Table 5
Frequencies of PoS-tag trigrams in a corpus of two sentences. The last column is a frequency vector of trigrams of part-of-speech tags.

Trigram			*It will become easier*	*This is not true*	Frequency
$	pron	mod	1	1	2
pron	mod	verb	1	0	1
mod	verb	adj	1	0	1
verb	adj	$	1	0	1
pron	mod	part	0	1	1
mod	part	adj	0	1	1
part	adj	$	0	1	1

son, we decided to base all further analyses on the simpler symmetrical (i.e., non-directional) trigram measure of syntactic distance, also for the language pairs in the Romance and Slavic families.

The trigram method is easy to implement and does eventually even not require parallel corpora. But the results of this method may be more difficult to interpret linguistically. Here the work of Swarte (2016) is helpful again. In addition to the trigram method, she used two other syntactic measures.

The first method is the movement measure. Swarte (2016:127) describes this measure as “the number of words that are moved when translating a sentence from language A into language B” where a “movement consists of a deletion of a word in one position in the sentence and an insertion of that word at another position in the sentence.”

The second method is called the indel measure. This method simply counts the mean number of whole-word insertions and deletions needed to convert the 66 written sentences in language A to their closest counterparts in language B (or vice versa).

We exemplify the two measures in Table 6, with one single sample sentence in English (top row) and its most literal translation to Dutch (second row). The pronoun *it* has to be deleted before the modal verb and its equivalent *het* must be inserted after the modal. Similarly, to invert the order of verb and adjective at the end of the sentence, *become* must be deleted before and inserted immediately after the adjective. The total number of indels equals 4. When the edit distance is expressed in terms of movements, there are two movements, each spanning a distance of two words, so that, in this example, the movement distance equals 4.

In the Germanic language sample, Swarte (2016:134, Table 5.16) correlated both the movement measure and the indel measure with the trigram measure. Trigram measures correlated strongly with the movement distance ($r = 0.969$, $p < .001$), and moderately with the indel measure ($r = 0.568$, $p < .01$).[3] This means that the variance in the trigram dis-

[3] Swarte (2016) erroneously mentions a correlation of $r = 0.941$. The correct correlation is mentioned by Heeringa et al. (2018) in Table 4 where three movement measures are mentioned. The method described by Swarte (2016) is referred to as ‘movement binary’, but the one she likely used is the one referred to as “movement linear”.

Table 6
Illustration of indel and movement measures of syntactic distance for an English sample sentence and its nearest equivalent in Dutch.

English	*After*	*a*	*while*	*it*	*will*		*become*	*easier*		
Dutch	*Na*	*een*	*tijdje*		*zal*	*het*		*makkelijker*	*worden*	
	0	0	0	1	0	1	1	0	1	Indel = 4
				>	>	2	>	>	2	Move = 4

tances is explained for 94% by the movement measures and for 32% by the indel measures. Therefore, trigram measures represent mainly the number of words that have moved in a sentence, but partially the number of words that were added and omitted as well.

2.3. Cluster analysis

On the basis of the distance measurements, we visualized each of the three linguistic distance measures by means of a hierarchical cluster analysis. The results are binary tree structures (dendrograms) in which the languages are the leaves and the length of the branches reflect the distance between the leaves (Jain and Dubes, 1988). We computed one cluster tree for each linguistic distance type for each language family in our investigation, i.e., nine trees in all. Several alternatives exist. We used Ward's method, which minimizes the total within-cluster variance. At each step in the procedure, the pair of clusters is found that leads to minimum increase in the total within-cluster variance after merging (Jain and Dubes, 1988; Ward, 1963).

3. RESULTS

In this section we present the distances along the three linguistic dimensions we obtained for pairs of languages in the Germanic (Section 3.1), Romance (Section 3.2) and Slavic (Section 3.3) family. The distance matrices can be consulted in the Appendix, Tables B.1-2 for Germanic, C.1–2 for Romance, and D.1–2 for Slavic. In the main text we highlight and discuss qualitative parallels that can be observed between the traditional family trees and the three distance dimensions, as graphically illustrated in Fig. 3. In Section 4, we present a quantitative comparison of the linguistic and cladistic trees.

3.1. Distances in Germanic language pairs

3.1.1. Lexical distances (Appendix, Table B.1)

Looking at the visualizations of the lexical distances in Fig. 3 (top row), we note that the traditional distinction between North Germanic and West Germanic is well reflected. The lexical distance between Swedish and Danish is very small (mean distance over AB and BA order of the languages is 5.2%). Also, Dutch and German share a rather large part of their vocabulary (mean distance 20.7%). English is clustered with the other two West Germanic languages but the distances are large (between 38.1 and 51.4%) as a result of the large number of loan words from Romance languages.

3.1.2. Phonetic distances (Appendix, Table B.2, upper triangle)

As in the visualization of the lexical distances, the traditional distinction between North Germanic and West Germanic is well reflected. We found a distance of 29.6% between Dutch and German, and a distance of 34.5% between Danish and Swedish. It has often been noted that the Danish pronunciation has changed very rapidly during the past century (Brink and Lund, 1975; Grønnum, 1998), but nevertheless Danish and Swedish are clustered together. English is clustered with the other two West Germanic languages.

3.1.3. Syntactic distances (Appendix, Table B.2, lower triangle)

Like the other linguistic distances, the syntactic distances result in a North-West division of the Germanic language area, see Fig. 1. However, an important difference is the clustering of English with the Scandinavian languages. As mentioned above, English is usually classified as a West-Germanic language while Emonds and Faarlund (2014) argued that English syntax is of a Scandinavian rather than a West Germanic type and therefore question the traditional placement of Middle English as West Germanic. The authors argue that Middle English developed as a form of Anglicized Norse and that English therefore should be relocated at the North Germanic branch of the language family tree. This view seems to be supported by our syntactic distance measurements.

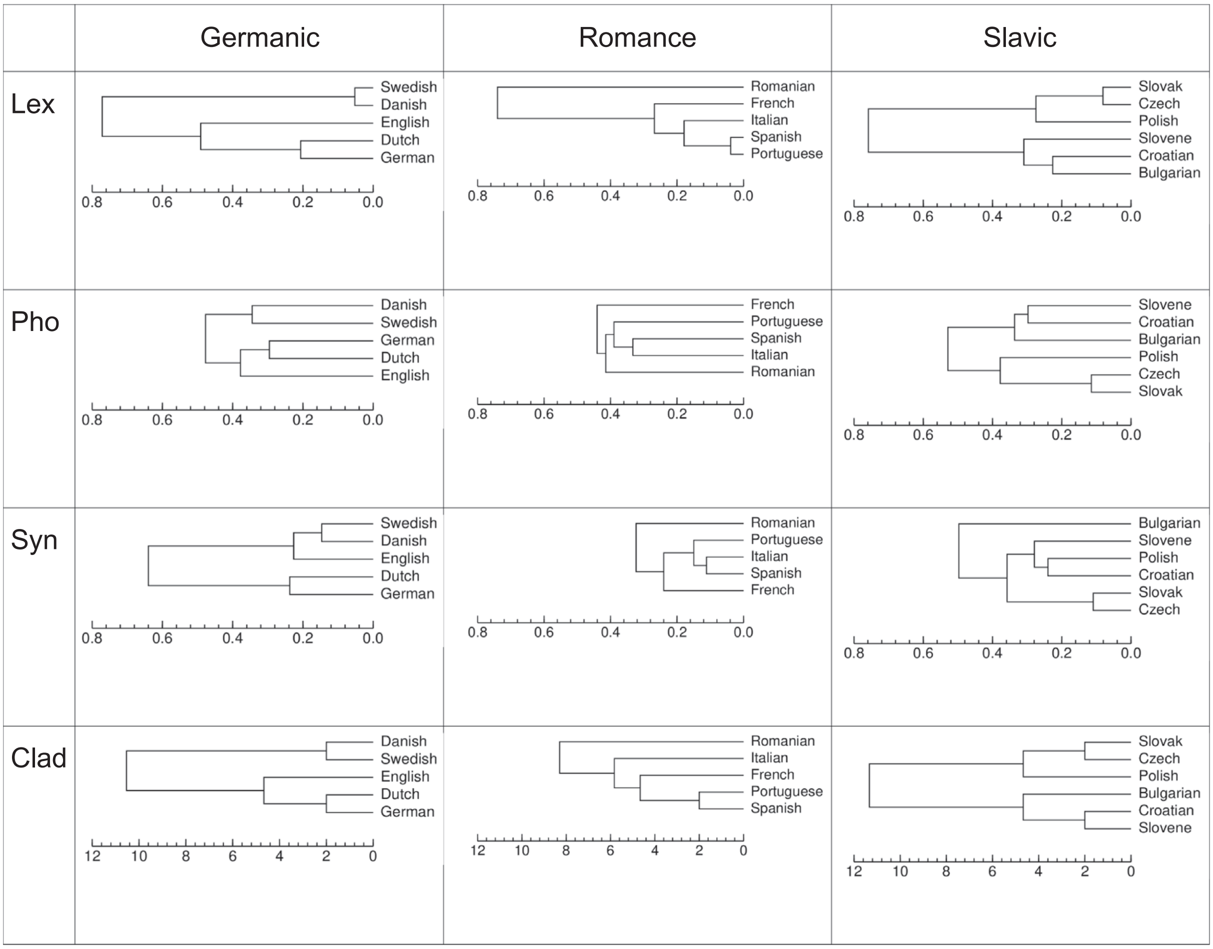


Fig. 3. Dendrograms per language group for each linguistic level. The fourth row shows dendrograms that are obtained from cladistic distances.

3.2. Distances in Romance language pairs

3.2.1. Lexical distances (Appendix, Table C.1)

As became clear from the discussion above, the characterization of the Romance language family is not straightforward and various language trees have been suggested in the literature. Rolfs' (1971) division could be expected to reflect our lexical measurements better than that of other scholars, since his division was mainly based on the degree of lexical differentiation from Latin while other divisions are based on phonetic/phonological criteria. However, the division of the Romance language area resulting from the lexical distance measurements presents Romanian as an outlier and has the greatest resemblance with the family tree suggested by Von Wartburg (1936), Hall (1974) and others. A difference from Hall's characterization is that Italian has a smaller distance to Spanish and Portuguese than to French, while French is clustered with Spanish and Portuguese in Hall's characterization. With the exception of Romanian, the lexical distances within Romance are generally rather small, between 3.9% for Spanish-Portuguese and 23.0% for French-Italian. Romanian borrowed a large part of its vocabulary from surrounding languages, especially Slavic, and only half of the Romanian words in our investigation have a cognate in the other four languages.

3.2.2. Phonetic distances (Appendix, Table C.2, upper triangle)

The tree structure does not show a clear grouping into clusters. The smallest distance is found between Italian and Spanish (33.3%), forming a kind of core varieties. Then Portuguese, Romanian and French are added sequentially. French is an outlier, as it is in the characterization by Rohlfs (1971). Italian is closest to Romanian (31.4%), which confirms the clustering of these two languages into East Romance by Von Wartburg (1936) on the basis of the isoglosses that run along the La Spezia-Rimini Line. In our data, Spanish is closer to Italian than to Portuguese. This can probably be explained by innovations in the Portuguese phonological system which have resulted in a rather complicated vowel system with nasalized vowels and a high incidence of assimilations (Mateus and d'Andrade, 2000) while the Spanish vowel system consists of only five vowels (Cressey, 1978).

3.2.3. Syntactic distances (Appendix, Table C.2, lower triangle)

The Romance dendrogram shows the largest syntactic similarities among Portuguese, Spanish and Italian. Romanian is again the outlier, which confirms Von Wartburg's (1936) representation of the language area. French shows the largest similarities to Spanish and Italian, which is more in line with the characterization by Rohlfs (1971).

3.3. Distances in Slavic language pairs

3.3.1. Lexical distances (Appendix, Table D.1)

The traditional division into a West and a South branch of the Slavic languages is well reflected by the lexical distances in our investigation. The only exception is the large lexical distance between Bulgarian and Slovene. Varieties of Croatian and Slovene are part of the South Slavic dialect continuum. In the case of Bulgarian, there are many dialects of Serbian spoken in between, and whereas these do share some similarities to Croatian, they are quite different from Slovene. Consequently, standard Slovene and standard Bulgarian are quite distant from each other. The distance between Slovak and Czech is smallest (8.1%). The remaining languages have distances between 21.6% (for Polish-Czech) and 39.2% (for Polish-Bulgarian).

3.3.2. Phonetic distances (Appendix, Table D.2, upper triangle)

As in the visualization of the lexical distances, the cladistic division into a West and a South branch of the Slavic languages is well reflected. However, the classification within the South branch is different. In the lexical domain we found the smallest distance between Croatian and Bulgarian, but in the phonetic domain we found the smallest distance between Croatian and Slovene, which seems more logical because Croatia and Slovenia are neighboring countries while Bulgaria does not border on either of these countries.

3.3.3. Syntactic distances (Appendix, Table D.2, lower triangle)

The West Slavic languages are syntactically clustered as they are for the other linguistic levels, but the syntactic relationships among the South Slavic languages deviate from the other levels. Croatian is clustered with West Slavic while Polish is syntactically closer to Slovene and Croatian than the South Slavic languages are to each other.

4. RELATIONSHIPS AMONG LINGUISTIC LEVELS COMPARED ACROSS THE LANGUAGE GROUPS

Section 3 shows that distance measurements at different linguistic levels may yield different clusterings of the languages within each language family. This means that the relationships among the languages differ across linguistic levels. In this section we will have a closer look at the relations among the three linguistic levels. In Appendices B, C and D, the distance matrices for respectively Germanic, Romance and Slavic are given for each of the three linguistic levels.

Except for the lexical level, distances are symmetric, i.e., the distances from language A to B is equal to the distance from language B to A. For the lexical level, distances A vs. B and B vs. A are averaged. Subsequently, in the analyses below only half matrices (after averaging and excluding cells along the main diagonal) will be used.

4.1. Comparison of linguistic levels

Product-moment correlations among the three linguistic levels are presented in Table 7. In classical correlation tests the assumption is made that the objects that are correlated are independent. However, values in distance matrices are usually not independent (Bonnet and Van de Peer, 2002). To account for distance correlations, Mantel (1967) developed an asymptotic test, in which the null hypothesis is that distances in the one matrix are independent of the corresponding distances in the other. We used the function mantel from the R package ecodist (Goslee and Urban, 2007). The significance of each of the correlations was evaluated with rows and columns jointly randomized per matrix and with 1000 permutations per test.

There is no pair of levels with a significant correlation in all three language groups. There are two pairs of levels that have significant correlations in two groups, namely lexis versus phonetics (in Germanic and Slavic) and phonetics ver-

Table 7
Product-moment correlations r between linguistic levels and family tree structure.[10.]

Family	Distance parameter	Lex	Pho	Syn	Cladistic
Germanic	Lexis		0.734*	0.482	0.851*
(*N* = 10)	Phonetics			0.120	0.832*
	Syntax				0.416
Romance	Lexis		0.482	0.818*	0.887*
(*N* = 10)	Phonetics			0.578	0.440
	Syntax				0.635
Slavic	Lexis		0.816**	0.420*	0.872*
(*N* = 15)	Phonetics			0.686*	0.706*
	Syntax				0.353*

* $p \leq 0.05$, ** $p \leq 0.001$ (two-tailed).

[10] The correlations with cladistic distance will be used in Section 4.2. We assume that cladistic distances as listed in Appendices B3, C3 and D3 are measures at the interval level, and can be analyzed with the product-moment correlation method. When correlations with cladistic distance are computed with ordinal methods such as Kendall's tau, the correlations drop considerably due to the large number of ties in the cladistic distances.

sus syntax (Romance and Slavic). A significant correlation between phonetics and syntax was only found for the Slavic language group.

In Table 7, the strength of the correlations between the distances in the three domains differs unsystematically from one language family to the next. There are three pairs of distances, i.e., Lex-Pho, Lex-Syn, Pho-Syn, in each of three language families. When we correlate the strength of these correlations for each of the three family pairs, we find a significant and negative correlation for the pair Romance-Slavic (r = –0.999, p <.05). We do not find a significant correlation for the pairs Germanic-Romance (r = 0.223, n.s.) and Germanic-Slavic (r = –0.177, n.s.). The absence of any positive significant correlation shows that there are no fixed relationships among the distances in the three domains across our language families.

4.2. Comparison of linguistic levels with cladistic distances

4.2.1. Correlations

The correlation coefficients between the three linguistic levels and the cladistic distances are shown in the rightmost column of Table 7 (above).

For all three language groups, the lexical distances correlate significantly with the cladistic distances. The phonetic distances have a significant correlation only for the Germanic and Slavic language groups. The syntactic distances have a significant correlation only for the Slavic language group.

4.2.2. Multiple regression

In this section we study how well the three linguistic levels predict the cladistic distances by means of a regression analysis. Given the dependencies within the distance matrices we do not perform classical linear regression analyses, but use the function MRM from the R package ecodist (Goslee and Urban, 2007). The function MRM (short for Multiple Regression on distance Matrices) is implemented according to the methodology of Legendre et al. (1994). A permutation test uses a pseudo-t test to assess significance, rather than the regression coefficients directly.

The results are shown Table 8. Only for Germanic did one of the predictors, i.e., Phonetics, reach significance. For Romance and Slavic, we found weak evidence that the lexical level is a predictor of cladistic distance (p values are 0.064 and 0.054 respectively).[4]

5. CONCLUSIONS AND DISCUSSION

In this paper we established symmetrical distances between members of 10 language pairs in the Germanic family, 10 Romance pairs, and 15 Slavic pairs. The distances were measured in each of three grammatical domains, i.e.,

[4] From the view that p-values are a continuum and provide a relative measure of strength of evidence, a p-value between 0.05 and 0.10 can be considered as indicating "weak evidence" (Ganesh and Cave, 2018).

Table 8
Results of multiple regression analyses where the three linguistic levels predict cladistic distance.[11] For ease of comparison with Table 7, both R^2 and R are specified.

	Coefficients				*F*	R^2	*R*	*p*
	Intercept	Lexis	Phonetics	Syntax				
Germanic	–6.484	0.045	23.329*	0.028	10.251	0.837	0.915	n.s.
Romance	2.425	0.092	4.017*	–0.079	8.921	0.817	0.904	n.s.
Slavic	1.001	0.126	–0.495*	–0.002	12.548	0.774	0.880	< 0.05

[11] A higher R^2 in the MRM results is not always accompanied by a lower *p*-value, due to the dependencies within each of the distance matrices that are included in the regression model. The *p*-value is the result of a permutation test that takes into account the symmetric structure of a dissimilarity matrix (or set of matrices), and those dependencies mean that (as we found) R^2 is not necessarily related to *p*-value in the way that parametric statistics may lead one to expect (Sarah Goslee, personal communication).

vocabulary, phonology, and syntax of the languages at issue. In our introduction we formulated three questions, which we will now try to answer on the basis of the results we obtained from our computational exercises.

Our first question was: *To what extent do the distances at various linguistic levels correlate with each other?*

Inspection of Table 6 reveals that for each of the three language groups, there is always at least one pair of correlated linguistic levels. In the Germanic and Romance groups, only one pair of levels is significantly correlated, while all pairs of levels are significantly correlated in the Slavic group. Since the number of language pairs in each group is small ($N = 10$ for Germanic and for Romance, $N = 15$ for Slavic), the correlation coefficients have to be substantial to reach significance. For the social sciences, the following guidelines may be considered adequate for the characterization of the strength of a correlation coefficient (Cohen, 1988). The effect size should be interpreted as small when *r* equals 0.1, medium for $r = 0.3$ and large for $r = 0.5$. Accordingly, we bracket effect sizes of *r* into three categories, i.e., low/poor correlation when $r < 0.2$, medium strength correlation for $0.2 < r < 0.4$, and high/strong correlation for any $r \geq 0.4$.

Following these guidelines, we may observe that the three linguistic domains, i.e., lexis, phonetics and syntax, generally do not change independently of one another in the historical development of the languages we targeted.[5] We found a majority of significant correlations (5 out of 9) that are both significant and strong. Three more correlations between domains were strong ($r \geq 0.4$) but did not reach statistical significance because of the small number of language pairs (in the Germanic and Romance sample; in the Slavic sample, such *r*-values would have been significant). Only once do we find a correlation that should be considered low, i.e., between phonetics and syntax in the Germanic group: $r = 0.12$, n.s.).

In light of the evidence, we should conclude that, as a general rule, languages evolve over time along parallel pathways. If two languages grow apart from one another in their historical development, changes will not take place in just one domain but innovations tend to occur in all three domains we studied. This tendency can be shown most clearly for the languages in the Slavic group, but the same trend can be observed in Germanic and Romance. Moreover, even though some of the correlations are low, we find positive correlations only; none of the nine correlations between linguistic domains in our data are negative. Apparently, it does not happen that sister languages grow apart in one domain and compensate the divergence by becoming more alike in some other domain.

This parallel development is reminiscent of what has happened in the development of sound systems in the languages of the world. It was suggested, at first, that complex word prosodic systems evolved to keep the segmental phoneme inventory of the language small. This predicts a negative correlation between, for instance, the number of lexical tones and the number of segmentally different syllables in languages. For instance, Mandarin has four lexical tones, which is above average for tone languages. This is offset by a rather limited set of ca. 400 segmentally different syllables, which is much smaller than average.[6] More recent investigations, however, have shown that segmental and suprasegmental complexity at the word level run parallel, i.e., are positively correlated (see e.g., Maddieson, 2011, 2013). Using the *World Atlas of Linguistic Systems* (Dryer and Haspelmath, 2009) as the source of information, Maddieson shows that both the sizes of the consonant and of the vowel inventories of a language are positively correlated

[5] English would appear to be an exception to this general pattern. The lexical distances suggest that English groups with the West-Germanic languages. At the syntactic level, however, we find English clustered with the North-Germanic languages. At the phonetic level, we find English apart from all of the other Germanic languages, likely as the result of Romance and Celtic influences.

[6] In the WALS sample of 527 languages, 218 (= 41%) have lexical tone. The majority of the tone languages (60%) have two tones (high vs. low).

with the number of tones, and that more generally “Greater complexity in one aspect of the phonological system more often goes hand in hand with greater complexity in others.” (Maddieson, 2011:30.

Our second question was: *Are the relationships among linguistic levels the same for Germanic, Romance and Slavic?*

It is clear from Table 7 (left part), and the discussion following it, that the relationships among the linguistic distance measures for lexis, phonetics and syntax do not run parallel in the three language families we examined. In the Germanic group the correlation was strongest for Lexis-Phonetics, weaker for Lexis-Syntax, and weakest for Phonetics-Syntax (LP-LS-PS = 1–2–3). In the Romance group, the order of the three domain pairs found was 2–3-1, while the order in the Slavic group was 1–3-2. When we correlate the strength of these correlations for each pair of language families, we find a significant and negative correlation for the pair Romance-Slavic (r = –0.999, p <.05). We do not find significant correlations for the pairs Germanic-Romance (r = 0.233, n.s.) and Germanic-Slavic (r = –0.177, n.s.). Although it is suggestive that the domain pair Lexis-Phonetics has the strongest correlation in two out of three languages families, the overall concordance is corrupted by the inconsistencies in the ordering of the remaining pairs. As a consequence of this, Kendal’s overall coefficient of concordance is poor and insignificant, W = 0.111 (p =.717, asymptotic).

The general conclusion here must be that the relationships among linguistic distances in each of the linguistic domains do not follow a uniform (let alone universal) pattern: they differ from one language family to the next. The validity of this conclusion is limited, however, by the small number of language families. We cannot exclude the possibility that the agreement would increase if a larger set of language families were sampled.

The last question we asked is: *To what extent are the traditional family relationships that were put forward by comparative linguists for the languages in our three groups, correlated with the distances we measured in the linguistic domains?*

As a preliminary answer to this question we may observe, in the rightmost column of Table 7, that distances between language pairs in all three linguistic domains correlate positively with the genealogical distance in the cladistic tree of their language family. The correlations are significant in six out of nine cases, and can still be qualified as “strong”, i.e., $r \geq 0.4$ if they fail to reach significance (due to the small number of language pairs in the Germanic and Romance samples).

Predicting the distance between language pairs in the family tree is already successful, even when the prediction is based on a single structural distance parameter. The best prediction in all three families is afforded by the lexical distance, i.e., the percentage of cognates in the vocabularies of the languages. The success of the prediction is expressed by the coefficient of determination R^2, which can be interpreted as a percentage.[7] Consequently, lexical distance explains 72% of the tree distance in the Germanic group, 79% in the Romance group, and 76% in the Slavic group. However, since the other two linguistic levels also correlate substantially with the cladistic tree distance, we may attempt to make a better prediction of the tree distance based on multiple structural distance parameters. This is what we did in Table 8. It then turns out that the R^2 rises to 87% for the Germanic group, 82% for the Romance family and 77% for Slavic. The values show that the phylogenetic trees proposed by linguists can be predicted to a large extent from structural distance measures, of which the lexical distance makes the largest contribution to the prediction. Second is the phonetic distance, and the smallest contribution is made by the syntactic level.

The overall conclusion, then, would be that our three-level quantitative analysis of a relatively small set of materials (which were used to establish degrees of cross-lingual intelligibility in an experimental setting, see Gooskens and Van Heuven, 2017, 2020; Gooskens et al., 2018) allows us to predict the traditional phylogenetic trees proposed by linguists to a high degree, with an accuracy between 77 and 87%.

The order of importance made by the three linguistic levels, i.e., lexis first, phonetics second and syntax last, would seem to reflect the way historical linguists have argued about family relationships among languages. The comparative method rests on finding cognate words in the vocabulary of languages. When arguing phylogeny, the proportion of cognates in the lexicons of two languages has traditionally been a strong indicator of the degree of their genealogical relatedness. Two varieties are said to be dialects of the same language if their lexicostatistics (percentage of cognates) exceeds a threshold of 70. If the lexicostatistics remain below 70%, the varieties should be considered different languages: “No instance is known of mutually intelligible dialects scoring below 70%.” (Dyen et al., 1992:9). The problem here is how to define cognates and how to minimize the risk of mistaking accidental similarity for cognacy. Two forms are cognate if they have both descended *in unbroken lines* from the same ancestor. Descent does not include borrowing from one speech variety to another. The two forms must be related by the *regular phonetic changes* which connect the

[7] When R^2 = 1 (or 100%) the prediction is perfect (the prediction error is 0); the predictor accounts for (“explains”) 100% of the variance in the criterion variable. When R^2 = 0.5, the prediction error is reduced to 50% relative to predicting that the score on the criterion variable equals the mean of that variable (which would minimize the prediction error). In the case of an R^2 of zero, 0% of the variance in the criterion scores is accounted for.

two target language varieties (Dyen et al., 1992:95). Determining what the regular sound changes are in the historical development presupposes a set of cognates that can be compared over time, which injects circularity in the definition of cognacy. Once the subset of cognates shared between varieties is identified, we can determine the degree of phonetic similarity between the cognates. The similarity can be expressed in a relatively simple manner in terms of (relative) number of phonemes (or combinations thereof, i.e., bigrams or trigrams, e.g., Lambert et al., 2001) or of distinctive features shared between the forms. More sophisticated ways are computing the edit distance between the strings (i.e., Levenshtein distance, as we did in the present article) or determining the complexity of the rule set that would be needed to convert all the cognate forms in variety A to their counterpart forms in variety B, e.g., Cheng's (1997) "Mutual Intelligibility", renamed Phonological Correspondence Index by Tang and Van Heuven (2007). In view of this interleaving of determining cognacy in the lexicon and establishing the set of regular sound changes, it seems difficult to separate the two information sources used in the construction of phylogenetic language trees. We can only observe that the lexicostatistics correlate somewhat better with the phylogenetic trees for the three language families we studied than the phonetic similarity in the cognate sets.

At the same time, our results suggest that syntactic similarity and difference plays a minor role in accounting for the phylogenetic tree structures we sampled from the linguistic literature. This would seem to reflect the minor or even absent role of syntactic criteria used in the construction of the traditional phylogenetic trees. This conclusion would seem to be at odds with findings reported by Ceolin et al. (2020), who estimated syntactic distance between all possible unique pairs of 69 languages samples from thirteen different language families. Syntactic distance was quantified in terms of presence vs. absence of 94 binary syntactic properties (universally definable parameters) used to capture the diversity of nominal structures. From these data, we calculated the Jaccard distances among the languages, separately for each of our three languages groups.[8, 9] In the Germanic group we substituted Norwegian for Swedish (which was not included in the 69-language sample), and in the Slavic groups we had to leave out Czech and Slovak (neither of which was included) in Ceolin et al. We correlated the Jaccard distances with our trigram distances and found large effect size correlation coefficients: r = 0.730 (p <.05) for Germanic, r = 0.604 (n.s.) for Romance and r = 0.554 (n.s.) for Slavic. There is also substantial correspondence between Ceolin et al.'s syntactic affinity trees (based on UPGMA clustering) and those established by ourselves. Substituting Norwegian for Swedish and disregarding Afrikaans, Icelandic and Faroese (which were not part of our sample) yields isomorphous tree structures in both approaches for Germanic. The subtree established by Ceolin et al. for the Romance languages (omitting the eleven Italian dialects from their total set of 16 languages) is isomorphous with our own syntactic affinity tree, with the exception of Italian, which was even more remote from the French-Spanish-Portuguese cluster than Romanian was. Our own syntactic affinity tree does, in fact, match the phylogenetic tree for Romance better than Ceolin et al.'s. In contrast to this, the Slavic subtree we extracted from Ceolin et al. matches our own tree structure rather poorly. Ceolin et al.'s sample contains five Slavic languages, four of which match our own. In both sets, Croatian and Slovenian cluster at the lowest level. The historically motivated primary split into West and South Slavic languages, which was reflected in our own syntactic affinity trees, was not reproduced in Ceolin et al. These correspondences suggest that, indeed, our PoS-tag trigram measure of syntactic distance captures many aspects of a more abstract syntactic structure based on Longobardi's (2005) minimalist parametric model. Both approaches, then, bear out that syntactic variation does reflect phylogenetic relationships among languages, even though the strength of the relationships may be less than that of lexis and phonetics.

Finally, although we can account for a high percentage of the variance in the phylogenetic trees by our relatively simple measures of lexical, phonetic and syntactic similarity, there still remains a part that cannot be explained. Rather than assuming that the part of the variance not accounted for is pure noise in the data, we would like to explore the possibility that including morphological similarity criteria might improve the predictive success. In the present paper we have compared the phonetic shapes of cognates without attempting to decompose the words into their constituent stems and affixes. We would expect that comparing only stems with stems (and only affixes with affixes) will correspond better with the decisions embodied on the traditional phylogenetic trees (Heeringa et al., 2014).

6. FUNDS

The data collection underlying this study was supported by the Netherlands Organisation for Scientific Research (NWO), under grant number 360–70-430 (principal investigators: Charlotte Gooskens and Vincent van Heuven).

[8] The data set is online available at: https://www.frontiersin.org/articles/https://doi.org/10.3389/fpsyg.2020.488871/full#supplementary-material.

[9] We used the web app LED-A at https://www.led-a.org/, in particular the option 'binary item comparison'.

Data availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGEMENTS

We thank Jelena Golubović, Femke Swarte and Stefanie Voigt for collecting and preparing the data for linguistic distance measurement.

APPENDIX A:. THE FOUR TEXTS USED FOR INTELLIGIBILITY TESTING

A1. 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.

A2. Catching a cold

Hello, my name is Christina 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!

A3. 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.

A4. 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.

APPENDIX B:. DISTANCES AMONG GERMANIC LANGUAGES

Table B.1
Lexical distance (asymmetrical).

Source language	Target language				
	Da	Du	En	Ge	Sw
Danish		43.94	49.17	38.95	05.82
Dutch	45.60		39.02	21.52	47.83
English	48.54	38.07		44.47	49.48
German	38.97	19.80	46.41		40.35
Swedish	04.63	46.75	51.38	39.88	

Table B.2
Phonological distance (upper); syntactic distance (lower triangle).

Language A	Language B				
	Da	Du	En	Ge	Sw
Danish		39.65	41.54	40.19	34.45
Dutch	36.80		33.85	29.56	37.20
English	19.10	34.41		37.64	42.41
German	40.50	23.77	43.64		37.36
Swedish	14.65	36.87	22.16	40.65	

Table B.3
Cladistic tree distance.

Language A	Language B			
	Du	En	Ge	Sw
Danish	6	6	6	2
Dutch		4	2	6
English			4	6
German				6

APPENDIX C:. DISTANCES AMONG ROMANCE LANGUAGES

Table C.1
Lexical distance (asymmetrical).

Source language	Target language				
	Fr	It	Pt	Ro	Sp
French		21.63	17.76	49.01	19.96
Italian	24.39		11.53	54.00	14.44
Portuguese	21.02	21.25		51.95	06.02
Romanian	57.81	52.14	53.90		54.03
Spanish	24.30	10.40	01.75	45.91	

Table C.2
Phonological distance (upper); syntactic distance (lower triangle).

Language A	Language B				
	Fr	It	Pt	Ro	Sp
French		40.94	43.27	44.21	38.58
Italian	21.17		41.23	39.19	33.28
Portuguese	23.26	13.81		42.83	33.88
Romanian	31.65	26.50	20.79		36.99
Spanish	16.64	11.13	14.22	26.87	

Table C.3
Cladistic tree distance.

Language A	Language B			
	It	Pt	Ro	Sp
French	5	4	7	4
Italian		5	6	5
Portuguese			7	2
Romanian				7

APPENDIX D:. DISTANCES AMONG SLAVIC LANGUAGES

Table D.1
Lexical distance (asymmetrical).

Source language	Target language					
	Bu	Cr	Cz	Po	Sk	Sl
Bulgarian		21.32	39.53	44.64	39.03	30.42
Croatian	23.88		33.10	44.27	31.15	22.91
Czech	38.78	39.18		25.44	05.99	45.17
Polish	37.62	43.91	17.67		23.43	46.85
Slovak	39.46	37.96	10.20	23.95		40.68
Slovene	36.30	25.80	36.84	47.87	36.44	

Table D.2
Phonological distance (upper); syntactic distance (lower triangle).

Language A	Language B					
	Bu	Cr	Cz	Po	Sk	Sl
Bulgarian		31.90	40.22	38.84	42.29	33.45
Croatian	41.12		28.14	35.02	30.15	29.75
Czech	40.10	29.11		30.85	11.48	37.11
Polish	36.71	23.99	21.70		31.52	39.29
Slovak	41.99	25.04	10.92	23.14		36.57
Slovene	38.66	29.49	27.84	24.41	31.03	

Table D.3
Cladistic tree distance.

Language A	Language B				
	Cr	Cz	Po	Sk	Sl
Bulgarian	4	6	6	6	4
Croatian		6	6	6	2
Czech			4	2	6
Polish				4	6
Slovak					6

References

Bergman, G., 1979. Likt och olikt I de skandinaviska språken. Föreningarna Nordens Förbund, Stockholm.

Bonnet, E., Van de Peer, Y., 2002. Zt: a software tool for simple and partial Mantel tests. Journal of Statistical Software, 7(10), 1–12. https://doi.org/10.18637/jss.v007.i10.

Bossong, G., 2016. Classifications. In: M. Maiden, A. Ledgeway. The Oxford Guide to the Romance languages. Oxford University Press, Oxford, pp. 63–72.

Braunmüller, K., 1999. Die skandinavischen Sprachen im Überblick. Francke Verlag, Tübingen/Basel.

Brink, L., Lund, J., 1975. Dansk rigsmål: lydudviklingen siden 1840 med særligt henblik på sociolekterne i København. Gyldendal, Copenhagen.

Ceolin, A., Guardiano, C., Irimia, M.A., Longobardi, G., 2020. Formal syntax and deep history. Frontiers in Psychology 11, 488871. https://doi.org/10.3389/fpsyg.2020.488871.

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 B 376, 20200197. https://doi.org/10.1098/rstb.2020.0197.

Cheng, C.-C., 1997. Measuring relationship among dialects: DOC and related resources Retrieved from Computational Linguistics & Chinese Language Processing 2 (1), 41–72. Retrieved from https://aclanthology.org/O97-3002.pdf.

Cohen, J., 1988. Statistical power analysis for the behavioral sciences. Routledge Academic, New York, NY.

Council of Europe, 2001. Common European framework of reference for languages. Learning, teaching, assessment. Cambridge University Press, Cambridge. Retrieved from https://rm.coe.int/16802fc1bf.

Cressey, W.W., 1978. Spanish phonology and morphology: a generative view. Georgetown University Press, Washington, DC.

Crystal, D., 2008. A dictionary of linguistics and phonetics. Blackwell, Malden. https://doi.org/10.1002/9781444302776.

Dietrich, W., Geckeler, H., 2000. Einfürung in die spanische Sprachwissenschaft. Erich Schmidt Verlag, Berlin.

Dryer, M.S., Haspelmath, M. (Eds.), 2009. World atlas of language structures. electronic edition. Max Planck Institute of Evolutionary Anthropology, Leipzig., Retrieved from http://wals.info/.

Dunn, R. H., Vitolins, V., 2019. eSpeak NG speech synthesizer. In GitHub repository (Version 1.50). Retrieved from https://github.com/espeak-ng/espeak-ng.

Dyen, I., Kruskal, J.B., Black, P., 1992. An Indoeuropean classification: A lexicostatistical experiment (Transactions of the American Philosophical Society, 82–5). American Philosophical Society, Philadelphia., Retrieved from https://www.jstor.org/stable/1006517.

Emonds, J., Faarlund, J.T., 2014. The language of the Vikings. Palacky University, Olomouc.

Ganesh, S., Cave, V., 2018. P-values, p-values everywhere! New Zealand Veterinary Journal 66 (2), 55–56. https://doi.org/10.1080/00480169.2018.1415604.

Golubović, J., 2016. Mutual intelligibility in the Slavic language area (Groningen Dissertations in Linguistics, 152). CLCG, Groningen, Retrieved from https://pure.rug.nl/ws/portalfiles/portal/31880596/Complete_thesis.pdf.

Gooskens, C., 2007. The contribution of linguistic factors to the intelligibility of closely related languages. Journal of Multilingual and Multicultural Development 28 (6), 445–467. https://doi.org/10.2167/jmmd511.0.

Gooskens, C., Heeringa, W., 2004. The position of Frisian in the Germanic language area. In: Gilbers, D., Schreuder, M., Knevel, N. (Eds.), On the boundaries of phonology and phonetics. University of Groningen, Groningen, pp. 61–87., Retrieved from https://www.academia.edu/5221526/On_the_Boundaries_of_Phonology_and_Phonetics.

Gooskens, C., Heeringa, W., 2006. The relative contribution of pronunciation, lexical and prosodic differences to the perceived distances between Norwegian dialects.. Literary and Linguistic Computing, special issue on Progress in dialectometry: Toward explanation 21 (4), 477–492, Retrieved from https://www.academia.edu/4862206/Phonetic_and_Lexical_Predictdors_of_Intelligibility.

Gooskens, C., Van Heuven, V.J., 2017. Measuring cross-linguistic intelligibility in the Germanic, Romance and Slavic language groups. Speech Communication 89, 25–36. https://doi.org/10.1016/j.specom.2017.02.008.

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. https://doi.org/10.1075/lab.17084.go.

Gooskens, C., Van Heuven, V.J., Golubović, J., Schüppert, A., Swarte, F., Voigt, S., 2018. Mutual intelligibility between closely related language in Europe. International Journal of Multilingualism 15 (2), 169–193. https://doi.org/10.1080/14790718.2017.1350185.

Goslee, S.C., Urban, D.L., 2007. The ecodist package for dissimilarity-based analysis of ecological data. Journal of Statistical Software 22 (7), 1–19. https://doi.org/10.18637/jss.v022.i07.

Greenberg, J.H., 1987. Language in the Americas. Stanford University Press, Stanford, CA.

Grieve, J., 2013. A statistical comparison of regional phonetic and lexical variation in American English. Literary and Linguistic Computing 28 (1), 82–107. https://doi.org/10.1093/llc/fqs051.

Grønnum, N., 1998. Fonetik og fonologi. Almen og dansk, Akademisk Forlag, Copenhagen.

Hall, R.A., 1974. External history of the Romance languages. Elsevier, New York, NY.

Harbert, W., 2007. The Germanic languages. Cambridge University Press, Cambridge.

Heeringa, W., Hinskens, F., 2014. Convergence between dialect varieties and dialect groups in the Dutch language area. In: Szmrecsanyi, B., Wälchli, B. (Eds.), Aggregating dialectology, typology, and register analysis; linguistic variation in text and speech (Linguae et Litterae, 28). De Gruyter, Berlin and Boston, pp. 26–52, 452–453.

Heeringa, W., Swarte, F., Schüppert, A., Gooskens, C., 2014. Modeling intelligibility of written Germanic languages: Do we need to distinguish between orthographic stem and affix variation? Journal of Germanic Linguistics 26 (4), 361–394. https://doi.org/10.1017/S1470542714000166.

Heeringa, W., Swarte, F., Schüppert, A., Gooskens, C., 2018. Measuring syntactical variation in Germanic texts. Digital Scholarship in the Humanities 33 (2), 277–296. https://doi.org/10.1093/llc/fqx029.

Hendriksen, C., Van der Auwera, J., 1994. The Germanic languages. In: König, E., Van der Auwera, J. (Eds.), The Germanic languages. Routledge, New York, NY, pp. 1–18.

Jain, A.K., Dubes, R.C., 1988. Algorithms for clustering data. Prentice-Hall, Upper Saddle River, NJ.

Jakobson, R., 1955. Slavic Languages, a condensed survey. Columbia University Press, New York, NY.. https://doi.org/10.7312/jako92822.

Kortmann, B., 2016. The Viking Hypothesis from a dialectologist's perspective. Language Dynamics and Change 6 (1), 27–30. https://doi.org/10.1163/22105832-00601006.

Kufner, H.L., 1972. The grouping and separation of the Germanic languages. In: Van Coetsem, F., Kufner, H.L. (Eds.), Toward a grammar of Proto-Germanic. Max Niemeyer, Tubingen, pp. 72–97.

Lambert, B.L., Chang, K.-Y., Lin, S.-J., 2001. Effect of orthographic and phonological similarity on false recognition of drug names. Social Science & Medicine 52 (12), 1843–1857. https://doi.org/10.1016/s0277-9536(00)00301-4.

Legendre, P., Lapointe, F., Casgrain, P., 1994. Modeling brain evolution from behavior: A permutational regression approach. Evolution 48, 1487–1499. https://doi.org/10.1111/j.1558-5646.1994.tb02191.x.

Lewis, M.P., Simons, G.F., Fennig, C.D. (Eds.), 2015. Ethnologue: Languages of the world. Eighteenth edition. SIL International, Dallas, TX.

Longobardi, G., 2005. A minimalist program for parametric linguistics. In: Broekhuis, H., Corver, N., Huijbregts, R., Kleinhenz, U., Koster, J. (Eds.), Organizing grammar. DeGruyter Mouton, Berlin, pp. 407–414.

Maddieson, I., 2011. Phonological complexity in linguistic patterning. Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, pp. 28–34. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/PlenaryLecture/Maddieson/Maddieson.pdf.

Maddieson, I., 2013. Tone. In: Dryer, M. S., Haspelmath, M. (Eds.), The world atlas of language structures online. Max Planck Institute for Evolutionary Anthropology, Leipzig. Retrieved from http://wals.info/chapter/13.

Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. Cancer Research 27 (2), 209–220.

Mateus, M., d'Andrade, E., 2000. The phonology of Portuguese. Oxford University Press, Oxford.

Nerbonne, J., Wiersma, W., 2006. A measure of aggregate syntactic distance. In: Nerbonne, J., Hinrichs, E. (Eds.), Proceedings of the Workshop on Linguistic Distances. Association for Computational Linguistics, Sydney, pp. 82–90. Retrieved from https://aclanthology.org/W06-1111.pdf.

Nunnally, J., 1978. Psychometric theory. McGraw-Hill, New York, NY.

Rohlfs, G., 1971. Romanische Sprachgeographie. Geschichte und Grundlagen, Aspekte und Probleme mit dem Versuch eines Sprachatlas der romanischen Sprachen. Beck, München.

Ruhlen, M., 1994. On the origin of languages: studies in linguistic taxonomy. Stanford University Press, Stanford, CA.

Ruhlen, M., 1987. A guide to the world's languages, Vol. 1: Classification. Stanford University Press, Stanford, CA.

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

Spruit, M.R., Heeringa, W.J., Nerbonne, J., 2009. Associations among linguistic levels. Lingua 119 (11), 1624–2142. https://doi.org/10.1016/j.lingua.2009.02.001.

Sussex, R., Cubberley, P., 2006. The Slavic languages. Cambridge University Press, Cambridge.

Swarte, F., 2016. Predicting the mutual intelligibility of Germanic languages from linguistic and extra-linguistic factors. (Groningen Dissertations in Linguistics, 150). CLCG, Groningen. Retrieved from https://pure.rug.nl/ws/portalfiles/portal/29253828/Complete_thesis.pdf.

Tang, C., Van Heuven, V.J., 2007. Mutual intelligibility and similarity of Chinese dialects. Predicting judgments from objective measures. In: Los, B., Van Koppen, M. (Eds.), Linguistics in the Netherlands 2007. John Benjamins, Amsterdam, pp. 223–234. https://doi.org/10.1075/avt.24.21tan.

Torp, A., 2002. The Nordic languages in a Germanic perspective. In: Bandle, O., Braunmüller, K., Jahr, E.H., Karker, A., Naumann, H.-P., Teleman, U. (Eds.), The Nordic languages. An international handbook of the history of the North Germanic languages, Volume I. Walter de Gruyter, Berlin/New York, pp. 13–24.

Trask, R., 1996. Historical linguistics. Arnold, London.

Trips, C., 2022. Morphological change. In: Lieber, R. (Ed.), Oxford research encyclopedia of linguistics. Oxford University Press, Oxford. https://doi.org/10.1093/acrefore/9780199384655.013.260.

Trudgill, P., 2016. Norsified English or anglicized Norse? Language Dynamics and Change 6 (1), 46–48. https://doi.org/10.1163/22105832-00601011.

Trudgill, P., 2020. Sociolinguistic typology and the speed of linguistic change. Journal of Historical Sociolinguistics 6 (2), 20190015. https://doi.org/10.1515/jhsl-2019-0015.

Van Gelderen, E., 2016. Split infinitives in Early Middle English. Language Dynamics and Change 6 (1), 18–20. https://doi.org/10.1163/22105832-00601003.

Van Heuven, V.J., Gooskens, C., Van Bezooijen, R., 2015. Introducing MICRELA: Predicting mutual intelligibility between closely related languages in Europe. In: Navracsics, J., Bátyi, S. (Eds.), First and second language: Interdisciplinary approaches (Studies in Psycholinguistics 6). Tinta Könyvkiadó, Budapest, pp. 127–145. Retrieved from https://scholarlypublications.universiteitleiden.nl/handle/1887/38285.

Van Kemenade, A., 2016. English: The extent of Viking impact remains open. Language Dynamics and Change 6 (1), 24–26. https://doi.org/10.1163/22105832-00601005.

Vikør, L.S., 2002. The Nordic language area and the languages in the north of Europe. In: Bandle, O., Braunmüller, K., Jahr, E.H., Karker, A., Naumann, H.-P., Teleman, U. (Eds.), The Nordic languages. An international handbook of the history of the North Germanic languages, Volume I. Walter de Gruyter, Berlin/New York, pp. 1–12.

Von Wartburg, W., 1936. Ausgliederung der romanischen Sprachräume. Zeitschrift für Romanische Philologie 56, 1–48.

Ward Jr., J.H., 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58, 236–244.

Wieling, M., Prokić, J., Nerbonne, J., 2009. Evaluating the pairwise alignment of pronunciations. In: Borin, L., Lendvai, P. (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. Retrieved from https://dl.acm.org/doi/abs/10.5555/1642049.1642053.

Wieling, M. B., 2012. A quantitative approach to social and geographical dialect variation. (Groningen Dissertations in Linguistics, 103). CLCG, Groningen. Retrieved from http://martijnwieling.nl/Dissertation-Martijn-Wieling.pdf.