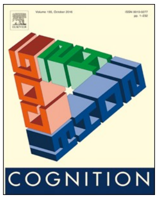
Cognition 194 (2020) 104056

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit



Big data suggest strong constraints of linguistic similarity on adult language learning

Job Schepens[a,*], Roeland van Hout[b], T. Florian Jaeger[c]

[a] *Centre for Cognitive Neuroscience, Free University Berlin, Germany*
[b] *Centre for Language Studies, Radboud University Nijmegen, Netherlands*
[c] *Department of Brain and Cognitive Sciences, Department of Computer Science, University of Rochester, United States*

ARTICLE INFO

Keywords:
Second language learning
Transfer
Phonological similarity
Accents
Speech production
Speech perception

ABSTRACT

When adults learn new languages, their speech often remains noticeably non-native even after years of exposure. These non-native variants ('accents') can have far-reaching socio-economic consequences for learners. Many factors have been found to contribute to a learners' proficiency in the new language. Here we examine a factor that is outside of the control of the learner, linguistic similarities between the learner's native language (L1) and the new language (L*n*). We analyze the (open access) speaking proficiencies of about 50,000 L*n* learners of Dutch with 62 diverse L1s. We find that a learner's L1 accounts for 9–22% of the variance in *Ln* speaking proficiency. This corresponds to 28–69% of the variance explained by a model with controls for other factors known to affect language learning, such as education, age of acquisition and length of exposure. We also find that almost 80% of the effect of L1 can be explained by combining measures of phonological, morphological, and lexical similarity between the L1 and the L*n*. These results highlight the constraints that a learner's native language imposes on language learning, and inform theories of L1-to-L*n* transfer during L*n* learning and use. As predicted by some proposals, we also find that L1-L*n* phonological similarity is better captured when subcategorical properties (phonological features) are considered in the calculation of phonological similarities.

0. Introduction

Adult learners of a second or additional language (L*n*) need to acquire syntactic structures, lexical items, morphological paradigms, and phonological properties in order to successfully communicate in the new language. This learning process is complex, and even after years of exposure to an L*n*, many adult learners fail to achieve native-like proficiency. Deviation from native speech can have far-reaching socio-economic consequences for learners. Non-native accents are often the subject of stereotyping (Brennan & Brennan, 1981; Lippi-Green, 2012; Munro, 2003), resulting in harming effects on e.g. individuals' perceived intelligence and employment rate (Fuertes, Gottdiener, Martin, Gilbert, & Giles, 2012; Gluszek & Dovidio, 2010). The success of L*n* learning is known to be influenced by a variety of cognitive and social factors, including age of acquisition (Birdsong, 2014; Bongaerts, Van Summeren, Planken, & Schils, 1997; Flege, 2018; Lenneberg, 1967; Vanhove, 2013), duration of exposure to L*n* (Babcock, Stowe, Maloof, Brovetto, & Ullman, 2012; Granena & Long, 2013; Pica, 1983; Stevens, 1999, 2006), and individual differences in language learning aptitude (Andringa, Olsthoorn, van Beuningen, Schoonen, & Hulstijn, 2012; DeKeyser, 2012; Schumann, Crowell, Jones, Lee, & Schuchert, 2004).

Many of these factors are hardly under the control of the learner. This applies in particular to the factor we focus on here, the learner's native language background (Jarvis & Pavlenko, 2008; Lado, 1957; Schepens, Van der Slik, & Van Hout, 2013b). We use data from about 50,000 adult learners of Dutch, to investigate how learners' language background affects their L*n* Dutch speaking proficiency. Our **first goal** is to estimate how much of the variance in L*n* speaking proficiency is due to the learner's first language (L1) background. We do so while controlling for a number of other factors that influence L*n* learning including: age of acquisition, duration of exposure, and differences in education and literacy. This allows us to assess how strongly, compared to other factors, L1 affect learners' speaking proficiency—and thus, likely, learners' perceived non-nativeness. Our **second goal** is to estimate how much of the variance across L1s can be explained by similarity in the linguistic properties between the languages.

Previous work has assessed the effects of linguistic similarity through controlled production and perception experiments in the lab (e.g. Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997 for phonological similarity) or studies in more ecologically valid environments,

[*] Corresponding author at: Center for Cognitive Neuroscience Berlin, Freie Universitaet Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany.
E-mail address: jobschepens@gmail.com (J. Schepens).

https://doi.org/10.1016/j.cognition.2019.104056
Received 14 December 2018; Received in revised form 13 August 2019; Accepted 19 August 2019

for example, experiments in the classroom (see e.g. Cook, 2013; Major, 1992). Research within this traditional approach has found that similarities between the L1 and L*n* in terms of, for example, their phonology (e.g. Best, 1993; Escudero, 2005; Flege, 1995; Haugen, 1966; Lado, 1957; Weinreich, 1963), lexicon (e.g. Dijkstra, Miwa, Brummelhuis, Sappelli, & Baayen, 2010; Jarvis, 2000; Otwinowska-Kasztelanic, 2009; Ringbom, 2007; Vanhove & Berthele, 2015b), and morphology or syntax (e.g. Bohnacker, 2006; Ionin & Wexler, 2002; Johnson & Newport, 1989; Lupyan & Dale, 2010; McWhorter, 2007) affect L*n* learning and use (for recent reviews, see Jarvis, 2015; Pajak, Fine, Kleinschmidt, & Jaeger, 2016; Yu & Odlin, 2015). These similarities can have positive (facilitatory) and negative (interfering) effects on L*n* learning and use (Jarvis & Pavlenko, 2008; Jarvis, 2015; Odlin, 1989, 2012; Pajak et al., 2016; Ringbom, 2007).

These studies have typically compared small samples of learners (often 10–40 per group) from two or three L1 backgrounds at a time. The L1s and L*n*s considered across studies have largely been drawn from the same small sample, representing only a few language families (typically a small number of Indo-European languages as well as Japanese, Korean, and Chinese). This leaves open whether the conclusions of those studies generalize across language backgrounds that represent the full range of linguistic diversity. For phonological similarity, for example, Flege (2003) concludes: “It will be necessary to study a wide range of L1–L2 pairs and L2 speech sounds in order to draw general conclusions regarding the nature of constraints, if any, on L2 speech learning” (p. 28).

The present study takes an approach markedly different from traditional approaches. Our approach is summarized in Fig. 1. We utilize big data from a state-administered exam (abbreviated as STEX, Schepens, 2015) with about 50,000 participants to assess the effect of 62 different L1s on L*n* Dutch speaking proficiency. The learners in our sample represent native language backgrounds from over ten language families, including about 15,000 learners from non-Indo-European languages, as many immigrants to the Netherlands come from such language backgrounds. By utilizing data from a wide variety of language backgrounds, we reduce the risk that our findings are due to the specific L1s present in our study only.

Another important property of our approach—and why it is best understood as a complement to, rather than substitute for, traditional approaches—is that we aim to estimate the *cumulative* effect of the similarity of *entire phonological, morphological, and lexical systems*. As shown in Fig. 1, we use a single rating per learner as a measure of that learner’s speaking proficiency. This is in line with our goal to assess the cumulative effect of L1-L*n* similarity on L*n* speaking proficiency: it provides a birds-eye view on precisely the type of perceived (lack of) native-likeness that has been found to have far-reaching socio-economic consequences for immigrants. Our focus on the cumulative effect of linguistic similarity contrasts with traditional approaches. Earlier work has often focused on a small number of specific linguistic features—for example, specific phonological features (Bradlow et al., 1997) or syntactic properties (Bardel & Falk, 2007). Both approaches have unique strengths and limitations, to which we return to in the discussion.

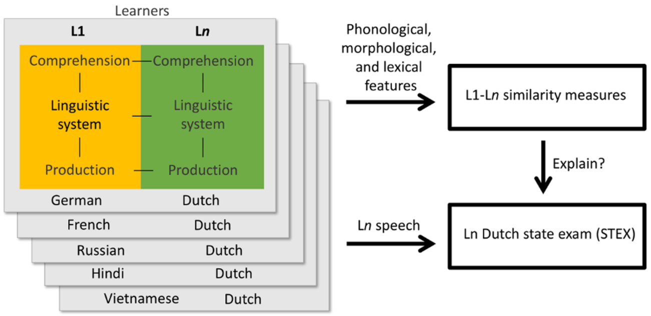


Fig. 1. Big data approach taken in the present study. We use language testing scores from 48,219 adult learners of Dutch with 62 different L1s (bottom right). Learners’ speech was evaluated for speaking proficiency by trained Ln Dutch teaching experts (top right) as part of a state-administered exam (STEX). For each of the 62 L1s, we calculate the L1-to-Ln similarity (bottom left) in terms of phonological (Studies 2 and 3), morphological, and lexical similarity (Study 4), and test whether these similarities explain variability in speaking proficiency. The colored boxes within the same learner (top left) symbolizes potential L1-to-Ln transfer during the comprehension, learning, and production of the Ln, causing the effects we aim to investigate.

We present four studies. Study 1 addresses our first question, by validating that language background indeed accounts for a large amount of variation in L*n* speaking proficiency. Studies 2–4 address our second question, the extent to which the effect of language background is mediated through linguistic similarity between the L1 and L*n*. Studies 2 and 3 develop and compare two new measures for phonological similarity. The first measure (Study 2) assumes that sound categories are atomic units (phonemes). This approach defines similarity as the number of new categories in the sound category inventory of the L*n*. The second measure (Study 3) captures subcategorical similarities between phonemes based on distinctive phonological features—i.e., similarity in terms of phonological features that make up the phonemes. Study 3 also tests whether the effects of distinctive features outweigh, or even subsume, the effects of similarities in the phoneme inventory, as predicted by theories of L1-to-L2 transfer during phonological learning (Best, 1993, 1995; Brown, 1998, 2000; Escudero, 2005; Flege, 1981, 1995). Finally, Study 4 takes a step towards estimating the joint effect of linguistic similarities between an L1 and L*n* on a learner’s speaking proficiency. Specifically, we ask how much of the variance in learners’ speaking proficiency can be explained through the combined effect of phonological, morphological, and lexical similarities between the L1 and L*n*.

1. Study 1: Variability in L*n* learning across language backgrounds

We introduce the STEX database and the statistical approach employed in all studies presented here. We then use these methods to present an estimation of the amount of variation in L*n* speaking proficiency accounted for by learners’ language background.

1.1. Data

We analyzed an open access database of 50,235 language-testing scores from a state exam (STEX) that assesses speaking, writing, reading, and listening proficiency in Dutch as an additional language (Schepens, van Hout, & Jaeger, 2019). STEX contains scores from learners of 70 self-reported language backgrounds. The exam is tailored to higher education; passing it is a requirement for admittance to a Dutch university. The pass level is upper-intermediate, equivalent to the B2 level of the Common European Framework of Reference for Languages. For further information on STEX, see Schepens (2015).

Here we analyzed the speaking scores, because we expected, in particular, the effect of phonological similarity to show clearly, but not exclusively, in speaking proficiency. These compound speaking proficiency scores represent the sum of about 45 ratings based on a total of 14 tasks that took about 30 min to complete. About 18 of the ratings provided 0 or 1 points, and about 27 of the ratings provided up to 3 points, for a maximum total of about 100 points.[1] Two trained, independent examiners evaluated the spoken language on both content and correctness according to a formal protocol. The passing level for the state exam requires a speaking score of about 60. As common in language testing, points were then scaled using Rasch models in order to

[1] There was some variability between learners in the specific number and types of tasks (and thus ratings), based on different versions of the state exam that were administered.

control for variations in exam difficulty. The resulting STEX speaking scores have a range of 415 points from 270 to 685 (mean = 517.6, SD = 37.8).

These speaking scores are a compound measure that captures a variety of aspects of speaking proficiency: about 30% of the score is based on ratings of content, 28% on grammar, 18% vocabulary, 12% pronunciation, 4% tempo and fluency, 3% coherence, 2% word choice, and 2% register. The full set of original ratings that went into the compound scores are available for only a small number of learners in STEX (about 1,359, 2.8%). This prevented us from analyzing only those ratings that would be most likely to be affected by linguistic transfer (e.g., ratings of pronunciation for phonological similarity).

The speaking scores we analyzed thus conflate multiple aspects of learner performance, including pronunciation, morphology, lexicon, and syntax, as well as aspects related to the content of the learners' productions. On the one hand. the use of a coarse-grained compound metric of speaking proficiency may limit our approach: STEX scores incorporate ratings about *content* and are, therefore, likely to be affected by factors not exclusively reflective of a learner's linguistic L*n* proficiency. Such factors might include, for example, learners' knowledge of the topic and comfort with public speaking (regardless of whether they are speaking their L1 or the L*n*). However, this does not necessarily confound our analysis: to the extent that other effects on the compound speaking score are not correlated with L1-L*n* linguistic similarity, the other effects cannot explain our results. We return to this point in the discussion. Importantly, the STEX scores are available for thousands of learners, averaging out individual variation that is independent from language-specific similarity.

We excluded all languages with fewer than 20 learners in STEX, which left 50,235 learners from 70 different L1s. The database we used to obtain linguistic information (described below) lacked information about eight of these languages (Afrikaans, Mandarin, Danish, Estonian, Haitian, Malayalam, Papiamentu, and Slovak). This left 48,219 learners (96% of total) from 62 L1s (89% of total) for analysis. This sample contains languages from 35 different genera and 12 different language families based on the Word Atlas of Language Structures (Dryer & Haspelmath, 2011). Of the 62 languages, 33 are part of the Indo-European family, eight are Niger-Congo, six are Afro-Asiatic, four are Austronesian, three are Altaic, two are Uralic, one each is Dravidian (Tamil), Austro-Asiatic (Vietnamese), Tai-Kadai (Thai) language, and the isolates Japanese and Korean.

STEX provides information on several control variables. These are: gender, educational level, country of birth, length of residence in the Netherlands, age at arrival in the Netherlands, and best additional language learned prior to learning Dutch. We further added a measure of the quality of education for the learner's country of birth (for details, see Schepens et al., 2013b) using World Bank statistics on the ratio of total enrollment into secondary education (UNESCO, 2011). See SOM Table S1 for descriptive statistics.

1.2. The baseline model

All analyses reported below use linear mixed effect regression (Bates, Maechler, Bolker, & Walker, 2014) to analyze the Dutch speaking proficiencies of the 48,219 learners in our sample (for introductions to linear mixed models, see Baayen, Davidson, & Bates, 2008; Jaeger, Graff, Croft, & Pontillo, 2011). The analyses use the *lmer ()* function of the *lme4* library (Bates, Mächler, Bolker, & Walker, 2015), version 1.1.17 in R (Core Team, 2018), version 3.4.4.

Specifically, our studies extend a baseline model of those speaking proficiencies—taken from Schepens, van der Slik, and van Hout (2016). This baseline model includes six control predictors: learner's gender, age at arrival in the Netherlands (as an estimate of age of acquisition), length of residence in the Netherlands (as an estimate of the duration of exposure to Dutch), the amount and quality of education in the learner's home country, and the interaction between the amount and quality of education. Further, the model contains four random intercepts, one each by country of birth, the learner's L1, additional self-reported language background (L2), and L1-by-L2 combinations. This random effect structure avoids inflation of Type I error due to violations of independence, as it captures that learners with shared backgrounds more likely resemble each other compared to learners with different backgrounds.

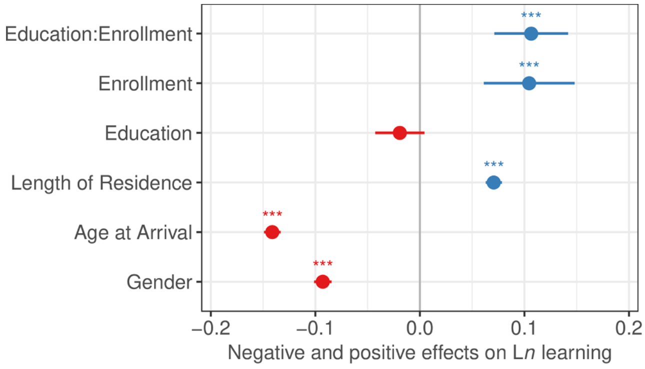


Fig. 2. Standardized coefficients along with 95% confidence interval estimated for the six factors included in the baseline model (plotted with R package sjPlot, Lüdecke, 2019). The standardized coefficients capture how many standard deviations speaking proficiency scores will change, per standard deviation increase in the predictor (e.g., Age at Arrival). For example, an increase of 1 standard deviation of Age at Arrival is predicted to result in a decrease in Ln Dutch speaking proficiency of about 0.15 standard deviations. Standardized coefficients are often used as an estimate of the strength of an effect, indicating here that, for example, Age at Arrival has a larger effect on Ln speaking proficiency than Length of Residency.

The effects of the six control predictors in the baseline model are shown in Fig. 2. All effects, except for the main effect of educational quality, are significant at *ps* $<$ 0.001 based on model comparisons. We left the main effect of educational quality in the model since it is part of the interaction with educational enrollment (for details, see Schepens et al., 2016).

The studies we present below extend this baseline model by adding measures of L1-to-L*n* (Dutch) similarity to the model. This allows us to assess the significance of L1-to-L*n* similarity while controlling for the effects in Fig. 2. If L1-to-L*n* similarity affects L*n* learning and use, we should see significant effects of L1-to-L*n* similarity. In addition, we expect to see that the estimated variance for the random by-L1 intercepts decreases when we add L1-to-L*n* similarity to the model, as the influence of different L1 background variance on L*n* Dutch speaking proficiency should explain L1-to-L*n* similarity.

1.3. The impact of learners' native language (L1) background

Dutch speaking proficiency varies substantially across L1 backgrounds in our sample, as shown in Fig. 3. Grouped by L1 background, Dutch speaking proficiency scores ranged from 482.5 (Somali) to 554.8 (German), with 50% of the scores falling between 497.7 and 526.3. That is, prior to controlling for any confounding factors, the average L*n* Dutch speaking scores of learners with different L1 backgrounds span a range of 72.3 points (17.4% of the total range of scores observed in STEX).

The baseline model lets us assess the importance of L1 learners' background while controlling for learners' gender, education, age at arrival, and length of residency in Holland. Specifically, we can estimate the relative contributions of language background by comparing the total variance explained by the baseline model against the total variance explained by the baseline model without random effects. Together the predictors in the baseline model account for 32% of the total variance in learners' L*n* speaking proficiency. We refer to this as the *explained variance* of the baseline model.

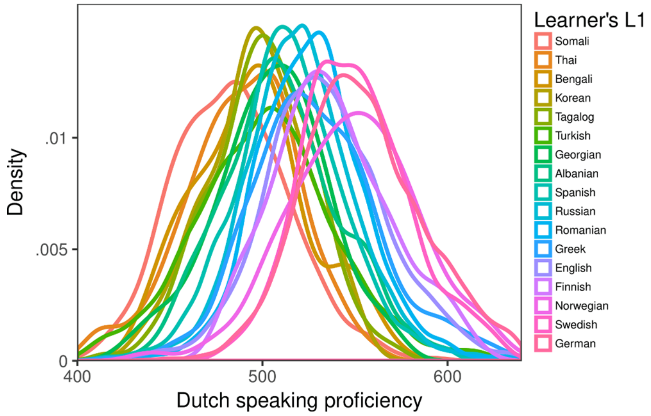


Fig. 3. Distribution of Ln Dutch speaking proficiency scores (x axis) for a number of learner's first language backgrounds (L1) show substantial variation both within and between L1s. L1s in the color legend are sorted from lowest (top) to highest (bottom) mean speaking proficiency.

When the random intercept by L1 is removed from the baseline model, this reduced model accounts for only 23% of the total variance. That is, L1 background accounts for *at least* 9% of the total variance (=32–23%, the difference between the baseline model and the reduced model) in L*n* speaking proficiency. The model without the L1 explains 72% (=23%/32%) of the explained variance so the L1 accounts for at least 28% of the explained variance (=9%/32%). It is important to emphasize that this is a very conservative lower bound of the influence of L1 background, as this way of assessing the influence of L1 background attributes all variance that *could* be explained by L1 background but also could be explained by other variables. Conversely, an *upper* bound of the influence of L1 background is obtained by a model that *only* contains by-L1 random intercepts and none of the other predictors from the baseline model. Such an L1-only model accounts for 22% of the total variance. This is 69% (=22%/32%) of the explained variance. L1 background alone thus accounts for an impressive 9–22% of the total variance in L*n* speaking proficiency among adult learners (28–69% of the explained variance). If L2 and L1-L2 combinations are considered as part of the language background, these estimates increase further to 11–25% of the total variance (34–78% of the explained variance).[2]

We can further quantify the importance of L1 background by comparing the variance of the different random intercepts in the baseline model. The estimated variance associated with the random intercept by L1 is an order of magnitude larger ($\hat{\sigma}^2$= 136.2) than the variance associated with country of birth ($\hat{\sigma}^2$= 44.5). The fact that L1 background captures so much more variance in speaking proficiency than country of birth is particularly noteworthy since the latter is likely to capture (at least) socioeconomic factors in addition to L1 background. The variance associated with learners' L1 background was also an order of magnitude larger than the variance associated with L2 background ($\hat{\sigma}^2$= 14.4) and the variance associated with L1-L2 combinations ($\hat{\sigma}^2$=10.4; for further discussion, see Schepens et al., 2016). This is in line with theories predicting that L1 background is more important than L2 background for L*n* learning (Cenoz, 2001; Escudero, Broersma, & Simon, 2013; Schepens et al., 2016) rather than vice versa (but see Bardel & Falk, 2007; and Pajak et al., 2016 for a discussion).

In summary, L1 background accounts for a *lot* of variance in L*n* Dutch speaking proficiency. Fig. 3 also suggests a reason for the considerable impact of L1 background: the L1 backgrounds with the highest Dutch speaking proficiency are languages that are genealogically closely related to Dutch and share many of its linguistic properties (e.g., German, Swedish, Norwegian). On the other end of the spectrum, we find languages that share comparatively few linguistic properties with Dutch (e.g., Somali, Thai, Bengali, and Korean). This explanation constitutes the focus of Studies 2–4. We begin by comparing measures of phonological similarity between two languages.

2. Study 2: Similarity in the sound inventory

Previous work on single phonological contrasts has found that similarity between the L1 and L*n* seems to facilitate L*n* proficiency (e.g. /r/-/l/; Aoyama, Flege, Guion, Akahane-Yamada, & Yamada, 2004; Bradlow et al., 1997; Flege, 1987). These and similar findings suggest that increased similarity between L1 and L2 phonology can increase the probability of *positive* transfer—not resulting in deviation from native pronunciation (Bohn & Flege, 1992; Haugen, 1966; LaCross, 2015; Lado, 1957; Major, 2008; Pajak & Levy, 2014; Weinreich, 1963). However, in some studies, high similarity between the L1 and L*n* has been found to increase the probability of *negative* transfer (Best, 1995; Flege, 1993, 2003; Flege, MacKay, & Meador, 1999; Kuhl, 1991; Piske, Flege, MacKay, & Meador, 2002). An example of such similarity interference is present in the phonological acquisition of L2 French by L1 English speakers. Experienced L1 English learners of L2 French produce the French /y/ (which does not exist in English) accurately, while their productions of the French sound /u/ (which is similar, but not identical, to English /u/) show influence from English /u/ (Flege, 1987). Similarly, when adult L1 speakers of Italian try to learn to produce the English sound /eɪ/ (as in *play* or *lane*), they seem to assimilate /eɪ/ to the phonologically similar L1 Italian category /e/ (as in *bed,* Piske et al., 2002). In such cases, the existence of a highly similar L1 category interferes with the production of an L*n* category. This could be the result of a failure to learn subtle differences between L*n* sounds and highly similar L1 sounds, possibly because of perceptual assimilation (Best, 1995; Escudero, 2005) or equivalence classification (Flege, 1995, 2003).

In sum, similarities between the L1 and L2 seem to increase the probability of transfer from L1 to L2. Previous work does, however, leave open whether the cumulative effects of such transfer are positive or negative. The goal of Study 2 is to assess whether L*n* speaking proficiency is affected by similarities in the meaning-distinguishing sound inventory (phonemes) between learners' L1 and the L*n* (Dutch). For Study 2, we do not consider the internal structure of sound categories. That is, the measure of the phonological similarity we employ is not sensitive to similarities *between* sound categories (e.g., in terms of their phonological features), but rather measures similarity between languages in terms of how many sound categories they share or do not share.

Specifically, we compare the effect of three different measures of the overall similarity of the L1 and L*n* category inventory (see Fig. 4). The new sounds measure counts the number of new sound categories that are only present in the L*n* and not in the L1 (the complement of L1 sounds in the L*n*). Consider the case of English learners of French. The French high front rounded vowel /y/ (as in *tu* "you") does not exist in the English sound inventory, and would thus constitute 1 new sound. We note that—for the present database—the number of new sounds are perfectly inversely correlated with the number of *shared* sounds between the L1 and L*n*, since the L*n* in our sample is always the same (Dutch). The missing sounds measure counts the number of missing sound categories that are present in the L1 but not in the L*n* (the complement of L*n* sounds in the L1). For example, English / ð/ (as in *that*) does not exist in the French sound inventory. Finally, the different sounds measure of inventory similarity counts the number of different sound categories between the L1 and the L*n*—the combination of the first and second measure.

[2] We note that, unlike for ordinary linear models, the variance components of linear mixed models do not decomposition neatly. We use R^2 (% variance accounted for) as a convenient and intuitive measure here. Leave-one-out model comparison confirms that the random intercept by L1, L2, and L1-by-L2 contributed significantly to the baseline model (all $\chi^2(1) > 21.35$, all $ps < 0.0001$; cf. Schepens et al., 2016 for details).

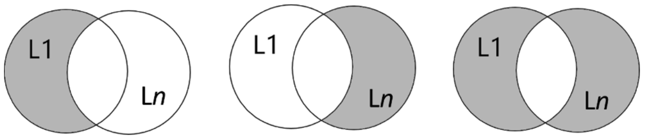


Fig. 4. Overlapping L1 and Ln sound category inventories with missing sounds (left panel), new sounds (middle panel), and different sounds in grey.

It is worth noting that these three measures—new, missing and different sounds—do not directly map onto measures typically used in traditional approaches. Previous research has focused on small numbers of specific L*n* categories that do not exist in the learner's L1. These studies have investigated how the L1 categories that are most similar to that L*n* category affect learners' perception and production of the L*n* category (Best, 1993, 1995; Eckman, Elreyes, & Iverson, 2003; Flege, 1987, 1995; Lado, 1957; Major, 2008; Weinreich, 1963). For the big data approach taken here, the perceptual and articulatory data required for such considerations are not yet available for most of the languages in our data set. We are thus limited to formulating differences between the L1 and L*n* at the phonological level. On the other hand, the present approach captures overall difference in the phonological system, whereas previous work has typically only considered the effects of the most similar categories. By comparing models with the different similarity measures, we can find out whether new or missing sounds affect proficiency more. If both types of differences between the L1 and L*n* weigh equally strongly, the number of different sounds should subsume the effect of the other two measures.

2.1. Determining numbers of new sounds for a language

Measuring similarity between phoneme inventories requires assumptions about the classification of phonological segments across languages. The sound inventories we use come from the PHOIBLE database (Moran, McCloy, & Wright, 2014), which aggregates information from a number of different sources (Crothers, Lorentz, Sherman, & Vihman, 1979; Lev, Stark, & Chang, 2012; Maddieson, 1984; PHOIBLE Online, 2014). For example, PHOIBLE takes the English phoneme inventory from the Stanford Phonology Archive (Crothers, Lorentz, Sherman, & Vihman, 1979), which itself uses information from various sources (Gimson, 1962; Halle, 1973).[3] PHOIBLE is a relatively new tool and has been used in research on cultural evolution (Gray & Watts, 2017; Hammarström, 2016), including studies on the relations between non-linguistic (e.g. population size) and linguistic factors (phoneme inventory) (Moran, McCloy, & Wright, 2012), serial founder effects (Cysouw, Dediu, & Moran, 2012), the role of features in phonological inventories (McCloy, Moran, & Wright, 2013), and language similarity (Skirgård, Roberts, & Yencken, 2017).

Our comparison of phonological systems across languages thus relies on phonological inventories *as provided in available databases*. In particular, we consider sounds with the same IPA symbols as identical. This is a common approach, also known as "phonetic symbol test" or "armchair heuristic" (Bohn, 2002; Flege, 1997). This has two related consequences, both of which are likely to affect the statistical power of our test. First, potential inaccuracy in the phonological inventories will get inherited by our analyses. While some of the phonological inventories in the database we employ are based on perceptual or articulatory experiments, the inventories for many of the languages in our database are the result of phonological classification. Second, even if two sounds are correctly categorized as having identical phonology, their articulatory and perceptual realizations can still differ across languages. This limits the sensitivity of our measures of phonological similarity. Critically, neither of these two limitations constitutes bias with regard the questions we seek to address here, and potential concerns about power are ameliorated by the very large number of learners in our sample.

We used the 2014 release of PHOIBLE to construct phoneme inventories for Dutch (our L*n*) and all L1s in our data. This is the latest release as of 2018, although more languages have been added in the meantime (see github.com/phoible). We first determined the ISO 639.3 codes for all languages. We then retrieved the phoneme inventories for these codes from PHOIBLE. For seven languages for which there was no available PHOBILE entry, we used entries of phonologically closely related neighboring languages (similar to Skirgård et al., 2017). This was the case for the following languages: Bosnian (bos) for Croatian (hrv), Belarus (bel) for Russian (rus), Latvian (lav) for Lithuanian (lit), Serbian (srp) for Croatian (hrv), Tamazight (tzm) for Shilha (shi), Tigrigna (tir) for Tigre (tig), and Urdu (urd) for Hindi (hin). However, we acknowledge that there are important phonological differences between these neighboring languages. For example, Latvian and Lithuanian are very closely related but they are not mutually intelligible because they differ in terms of e.g. stress and phonotactics.

PHOIBLE classifies each sound of a language as a category of the standard international phonetic alphabet (International Phonetic Association, 1999). In PHOIBLE's classification system, Dutch has 39 sound categories (following the most common system used for Dutch phonology, Booij, 1999). The Dutch vowel inventory contains five lax vowels (ɑ, ɔ, ɛ, ɪ, ʏ), seven tense vowels (aː, eː, i, oː, øː, u, y), and three diphthongs (ɔu, ɛi, œy), and finally, the schwa. The Dutch consonant inventory contains six plosives (b, d, k, p, t, ʔ), nine fricatives (f, ɣ, ɦ, s, ʃ , v, χ, z, ʒ), two glides (j, w), two liquids (l, r), and four nasals (m, n, ɲ, ŋ).

The import procedure is described in Moran (2012). Using the PHOIBLE inventories, the L1 backgrounds in our database resulted on average in 20.3 new sounds in L*n* Dutch (SD 2.96, range 13–27), 23.8 missing sounds (SD 14.1, range 6–77), and 44.1 different sounds (SD 14.9, range 24–98). For example, English, Korean, and Arabic differ along these three measures as follows. For L1 English speakers, there are 19 new sounds in Dutch, 21 missing sounds, and 19 shared sounds. For L1 Korean speakers, there are 22 new sounds in Dutch, 24 missing sounds, and 16 shared sounds. For L1 Arabic speakers, there are 24 new sounds in Dutch, 64 missing sounds, and 14 shared sounds.

2.2. Results and discussion

Next, we ask whether the number of new sounds, missing sounds, or different sounds explains more variation in Dutch speaking proficiency. Adding new sounds as a predictor to the baseline model gives a significant improvement (as measured by the change in deviance: $\chi^2(1) = 8.69$, $p < .01$): for each new L*n* sound, speaking proficiency decreases on average by 1.54 points. That is, the number of new sounds explains additional variance in learners' speaking proficiency in L*n* Dutch over and above the controls, including the variance that would be expected by random variation across learners' language backgrounds (random intercept by L1, L2, and L1-L2 combination). Adding the total number of different sounds to the baseline model improved model fit ($\chi^2(1) = 4.76$, $p < .05$), whereas missing sounds did not ($\chi^2(1) = 2.79$, $p = .10$).

Fig. 5 visualizes the effect of new sounds using the 62 adjusted speaking scores (the by-L1 random intercepts obtained from the baseline model described above). We note that these adjusted speaking scores control for the effects of other variables in the baseline model. Overall, these adjusted scores order similarly as the speaking proficiency scores in Fig. 3, but there are some minor differences (e.g., in

[3] Of the 62 L1s considered here, the phonological coding for 41 came from the Stanford Phonology Archive (Crothers et al., 1979), 3 from the UCLA Phonological Segment Inventory Database / UPSID (Maddieson, 1984; Maddieson & Precoda, 1990), and 18 from original sources (Moran, 2012). The procedure by which different coding systems were converted to IPA is described in Moran (2012). Conversion decisions made in PHOIBLE were blind to the goals of the present study, and thus do not introduce obvious bias.

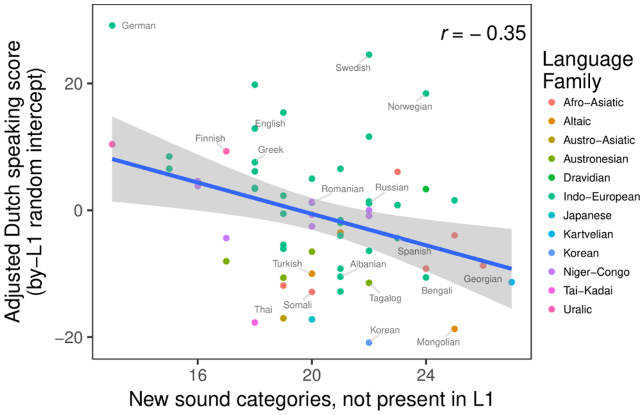


Fig. 5. The relation between adjusted Dutch speaking score (by-L1 random intercepts obtained from the baseline model, i.e. controlling for third factors) and the number of new sounds for every L1. The blue lines represent a linear regression with 95% confidence intervals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 3, Somali learners had the lowest L*n* Dutch speaking proficiency, but in Fig. 5 shows Korean learners to have the lowest adjusted speaking scores after differences in gender, education, age at arrival, and length of residence are taken into account).

The number of NEW SOUNDS correlates significantly ($r = -0.357$, $p < .01$) with the adjusted speaking proficiency scores, while the other two similarity measures did not correlate significantly with adjusted speaking proficiency (MISSING SOUNDS: $r = -0.193$, $p = .13$; DIFFERENT SOUNDS: $r = -0.09, p = .50$). Follow-up analyses found that the trending correlation for missing sounds was driven by four languages: Arabic, Amharic, Hindi, and Urdu. All four languages employ consonantal sound contrasts that rely on phonetic length (geminates). Dutch does not have geminates or other consonantal length distinctions. It is possible that learners from languages with geminates produce consonants that deviate from native Dutch in terms of their average length or variation in their length, resulting in lower Dutch speaking proficiency scores. Excluding these four languages from the analysis removed any evidence for a correlation for missing sounds ($r = -0.09, p = .50$), and did not change the significant correlation for the number of new sounds ($r = -0.35$, $p < .01$). The number of NEW SOUNDS thus explains more variation in L*n* proficiency compared to MISSING SOUNDS or DIFFERENT SOUNDS.

The results of Study 2 provide support from a large-scale sample of 62 different L1s for the hypothesis that L1-to-L*n* phonological dissimilarity between sound inventories impedes L*n* learnability. The number of new sounds that a learner of an L*n* has to acquire has long been hypothesized to affect L*n* learning and proficiency (Eckman et al., 2003; Flege, 1993; Lado, 1957; Major, 2008; Weinreich, 1963). However—to the best of our knowledge—Study 2 is the first quantitative test of this hypothesis against a large sample of L*n* learners from many different L1 backgrounds.

3. Study 3: Similarity of sound inventories in distinctive features

Study 3 focuses on the effects of phonological similarity in terms of distinctive phonological features, going beyond similarities in terms of discrete sound categories. Theories of L*n* learning emphasize the role of perceptual or articulatory similarities between L1 and L*n* categories (Best, 1993, 1995; Escudero, 2005; Flege, 1981, 1995), rather than just the existence of new categories. For example, whether a new phonological category in L*n* is subject to negative transfer is predicted to depend on the existence or absence of *similar* L1 categories. Differences in distinctive features are well-known to correlate with perceptual and articulatory similarity, and—unlike large-scale cross-linguistic perceptual or articulatory data—information about distinct features is available for the 62 L1 in our sample. Distinctive features have previously been successfully employed to approximate perceptual or articulatory effects on the acquisition of individual phonemic contrasts (e.g., to explain difference in r/l, b/v, and f/v learning across Japanese and Mandarin learners of English, Brown, 1998, 2000).

In the feature system used here, phonological features are binary properties of sounds that characterize speech sounds as distinct members of the phonological system (Chomsky & Halle, 1968; Jakobson, 1941, 1968). The feature system we employ is based on articulatory phonology (Hayes, 2009). In this system, described in more detail below, features represent specific movements of oral articulators or the larynx that characterize speech sounds. For example, the feature [sonorant] indicates that production of that sound requires continuous airflow in the vocal tract, [consonantal] requires the (partial) closure of the vocal tract, [continuant] requires incomplete closure of the vocal tract, [syllabic] requires the production of a syllable nucleus, [labial] requires articulation with the lips, [round] requires rounding with the lips and so on and so forth.

We expect that the sum of new features for the set of new sounds captures the cumulative effect of new features in defining phonological similarity. New sounds with many new features compared to their most similar L1 sound lead to more learning difficulty compared to new sounds with just one or two new features. The resulting measure most closely aligns with the assumption of contrastive analysis (Haugen, 1966; Weinreich, 1963); that larger differences in distinct features lead to a lower L*n* learnability. But our measure is also likely to correlate with the prediction of articulatory and perceptual theories of L1-to-L*n* transfer (Best, 1993; Escudero, 2005; Flege, 1981). For the present purpose, this is acceptable, as our goal is not to distinguish between these theories. Rather, we provide the first large-scale test of the hypothesis that similarities in phonological features, rather than just similarities in the phoneme inventory, are relevant to L*n* learning and proficiency. In Study 2, we found an impeding effect of the number of new sounds on L2 learnability. We expect that new features do better and can help to explain why learning new sounds impedes L2 learnability.

In order to measure similarity based on phonological features, we need to specify precisely how to link new sounds to an existing feature geometry. There are many ways to establish such a link. For example, the number of new features of an L*n* sound may be based on all sounds in the L1 sound inventory or a specific subset thereof. Here, we define a similarity measure based on the minimal number of new features of a new sound with respect to any of the existing L1 sounds. We take the sum of the number of new features of all new sounds to compute the overall number of new features between the L1 and L*n* sound inventories.

The distinctive features that we use can either have the value + to indicate presence or − to indicate absence, but they can also have no value, which indicates that a certain movement is not applicable if its "higher-order" movement is absent (e.g. whether or not the lips are [round] when [labial] is absent). We define new features as features that are present in a new sound and either absent or not applicable in the existing L1 neighboring sound.[4]

[4] Because of these differences between presence, absence, and inapplicability, we also considered an alternative measure of new L*n* features based on the number of both new present as well as new absent features in the new L*n* sound, instead of new present features only. This measure also counts differences between absent features and not applicable features instead of differences between present and not applicable features only. Additional post-hoc analyses not reported below found that this measure did not account for variance in L*n* Dutch proficiency.

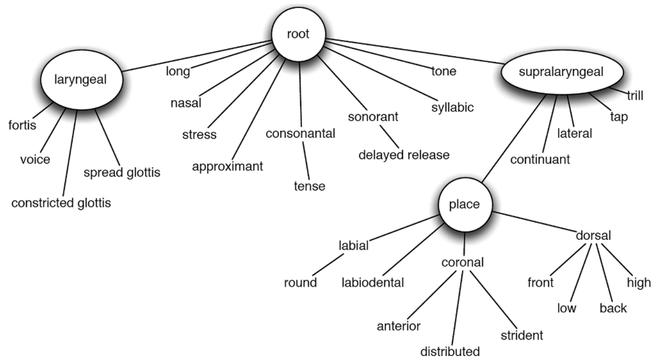


Fig. 6. The feature geometry as based on Hayes (2009) and copied from Moran (2012).

3.1. Calculating feature similarity

Our measures of subcategorical similarity are all based on the information available in the PHOIBLE database (Moran, McCloy, & Wright, 2014). The subcategorical information in PHOIBLE is based on an extension of a widely used phonological feature system (Hayes, 2009). PHOIBLE extends the original set of 30 features (Hayes, 2009) to 37 features (PHOIBLE Online, 2014). The additional features are all absent in Dutch. The geometry is shown in Fig. 6. This system is able to encode the sound inventories of about 71% of the languages of the world (PHOIBLE Online, 2014).

Table 1 uses the example of /ɑ/ and /a:/ to illustrate how we used the feature system to calculate the number of new features between two sounds. The table shows whether a feature is new and whether a feature is shared between sounds. We define shared features as features that are present in both sounds. /ɑ/ and /a:/ differ in two features: /ɑ/ is [back] whereas /a:/ is not and /a:/ is [long] whereas /ɑ/ is not. For an L1 without /a:/, in which /ɑ/ is the closest sound, we would thus count one new feature for the new sound /a:/. Seven features are shared, 17 features are absent in both sounds, and 9 are unspecified (NA) in both sounds. Four other situations can appear in our sound comparisons. It is sometimes the case that a new sound has: (1) a present feature that is not specified in the neighboring sound (+vs NA, e.g. /n/ vs /s/ for the feature [delayed release], which we count as a new feature), (2) an absent feature in the new sound that is not specified in the neighboring sound (− vs NA, e.g. /k/ vs /s/ for the feature [delayed release]), (3) a feature that is not specified in the new sound but present in the neighboring sound (NA vs +), (4) or finally, a feature that is not specified in the new sound and absent in the neighboring sound (NA vs −). Table 2 illustrates how the number of new or shared features for each Dutch category are summed into an L1-specific aggregate similarity score. This is illustrated for two L1 backgrounds, English and Korean. Summing over all Dutch sounds, English learners of Dutch are confronted with 11 new features, compared to Korean learners of Dutch, who are confronted with 22 new features. On average, Dutch has 16.2 new features (SD = 5.5, range = 6–32) and 138.1 shared features (SD = 19.6, range = 84–182) across all L1s.

Table 1
Example of a comparison of the new and shared features for Dutch /a:/ relative to /ɑ/, for a hypothetical L1 without /a:/ in which /ɑ/ is the closest existing sound to /a:/. Only a subset of all 37 features is shown.

Feature	L1 /ɑ/	Dutch / a:/	New feature	Shared feature
[syllabic]	+	+	no	yes
[long]	–	+	yes	no
[sonorant]	+	+	no	yes
[continuant]	+	+	no	yes
[approximant]	+	+	no	yes
[dorsal]	+	+	no	yes
[low]	+	+	no	yes
[back]	+	–	no	no
[periodic glottal source]	+	+	no	yes
[nasal]	–	–	no	no
[delayed release]	NA	NA	no	no

3.2. Results and discussion

Adding the number of new features to the baseline model results in a significantly better model fit ($\chi^2(1) = 15.00$, $p < .001$). Fig. 7 visualizes the effect of new features on adjusted speaking proficiency scores ($r = -0.47$, $p < .001$).[5] Adding the number of shared features also improved the baseline model, but less so than for the new features ($\chi^2(1) = 8.32$, $p < .005$). We then compared the added value of both the shared and the new features similarity measures. Adding the number of shared features to a model that already contained the number of new features did not significantly improve the model ($\chi^2(1) = 3.14$, $p = .08$). In contrast, new features did improve the model that contained the number of shared features ($\chi 2(1) = 9.82$, $p < .005$). Study 3 thus parallels Study 2: differences, rather than commonalities, seem to drive effects of L1-L*n* similarity on speaking proficiency.

Next, we assessed whether the effect of phonological similarity on L*n* proficiency is better captured by differences in phoneme inventories (as assessed in Study (2) or differences in phonological features (Study 3). We compared both the model from Study 2 and the model from Study 3 against a model that included both measures of phonological similarity. Adding feature similarity significantly improved the model from Study 2 ($\chi^2(1) = 7.95$, $p < .005$), whereas adding new sounds does not significantly improve the model from Study 3 ($\chi^2(1) = 1.19$, $p = .273$). Phonological similarities measured in terms of distinctive features thus *subsumes* the effects of new categories on L*n* proficiency *and* explains additional variance. Learners particularly benefit from similarity to new sounds in terms of the minimal number of new features that they need to learn. A higher number of new features has an inhibiting effect on L*n* learning.

This result aligns with theories that emphasize the role of the L1 in second language learning. Experiments on L*n* learning have shown that both L*n* pronunciations and their perception by native listeners are affected through sub-phonemic L1-L*n* similarities (Best, 1995; Chang, 2015; Escudero, 2005; Flege, 1987; Haugen, 1966; Strange et al., 2007; Weinreich, 1963). The role of new features of new sounds may be related to the relative complexity and size of the L1 and L*n* acoustic space. For example, a larger L1 than L*n* vowel inventory may be beneficial because of its extended number of acoustic subspaces (Iverson & Evans, 2009). The results of Study 2 are also compatible with frameworks in which more acoustical or phonological variation in previously acquired languages facilitates generalization to a new language or linguistic variant (Pajak et al., 2016).

4. Study 4: Combining phonological, morphological, and lexical similarity

Now that we have validated the big data approach as well as our measure of phonological similarity, we can address the second question raised in the introduction: how much of the effect on language learning due to L1 background can be attributed to the linguistic similarity between the L1 and L*n*?

Other studies on language learning have found that lexical similarities—such as a high number of cognates that are shared with the

[5] In order to assess the robustness of these results, we repeated the analyses while excluding learners who are familiar with a language besides their L1 and Dutch. Excluding multilingual learners results in a substantially lower number of observations: 8571 learners from 30 L1s (for which we have at least 15 monolingual speakers). The results remained unchanged.

Table 2

Examples of sound comparisons for Dutch (nld) to English (eng, light gray), and Dutch to Korean (kor, dark gray); y: (bold) is discussed in the main text. For each Dutch sound, its most similar phonological neighbor (or one of them in case of equally similar neighbors) in the L1 is shown. Additional columns indicate, from left to right within each L1, whether that nearest sound is shared with Dutch and how many of its features are shared with the Dutch sound (multiple neighbors do not affect the computation).

L*n*-Nld	L1-eng nearest sound	Sound	New features	Shared features	L1-kor nearest sound	Sound	New features	Shared features
œy	uː	new	1	9	y	new	0	10
p	p$^\text{h}$	new	0	2	p	shared	0	2
r	uː	new	2	6	o	new	2	6
s	s	shared	0	6	s	shared	0	6
ʃ	ʃ	shared	0	6	t̠ʃ$^\text{h}$	new	1	5
t	l	new	0	3	t	shared	0	3
u	uː	new	0	11	u	shared	0	11
v	v	shared	0	6	o	new	3	3
w	w	shared	0	10	w	shared	0	10
x	x	shared	0	5	iː	new	2	3
ʏ	uː	new	1	9	uː	new	1	9
yː	**uː**	**new**	**1**	**9**	**uː**	**new**	**1**	**9**
z	z	shared	0	7	s$^\text{ʔ}$	new	1	6
ʒ	ʒ	shared	0	7	t̠ʃ$^\text{h}$	new	2	5
ʔ	ʔ	shared	0	1	ʔ	shared	0	1
	iː	missing	0	10	t̠ʃ$^\text{h}$	missing	0	6
	k$^\text{h}$	missing	0	4	p$^\text{ʔ}$	missing	0	3

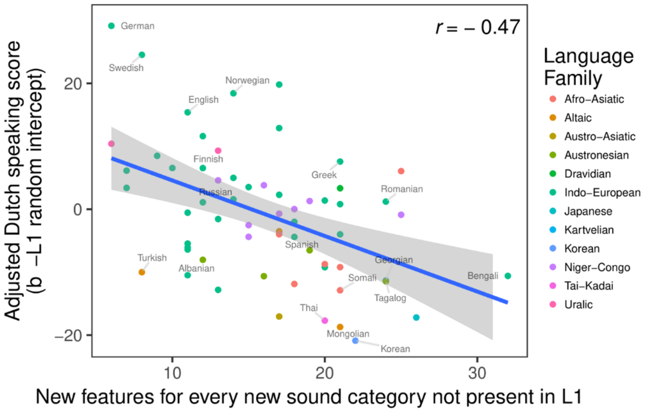


Fig. 7. The relation between adjusted Dutch speaking score (by-L1 random intercepts obtained from the baseline model, i.e. controlling for third factors) and the sum of the minimal number of new features for each new sound. The blue line represents a linear regression fit with 95% confidence interval. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

L1—can facilitate word learning for an L*n* (Dijkstra et al., 2010; Otwinowska-Kasztelanic, 2009; Ringbom, 2007; Vanhove & Berthele, 2015b, 2015a). Similarly, there is evidence that similarities in the morphological or syntactic systems of the L1 and L*n* can affect L*n* learning (Ionin & Wexler, 2002; Johnson & Newport, 1989; Lupyan & Dale, 2010; McWhorter, 2007). Study 4 thus combines the measure of phonological similarity derived in Study 3 with measures of morphological and lexical similarity. As we noted in our description of the STEX database, the speaking scores we analyze are compound measures based on ratings that assess a wide variety of linguistic properties of learners' speech. These scores are likely to be affected by linguistic processes other than articulation or phonological encoding, including difficulty experienced during lexical retrieval, syntactic planning, or discourse planning. As these processes might be subject to similarity effects operating at other levels of linguistic encoding, it should thus be possible to detect effects of L*n* speaking scores on, for example, lexical and morphological similarities between the L1 and L*n*. It is even possible that some of the effects observed in Study 3 might be confounded by L1-L*n* similarities at other levels of linguistic representation: similarities between languages tend to be correlated across different levels of linguistic representation (Croft, 1990; Dunn, Greenhill, Levinson, & Gray, 2011; Nettle, 1999; Trudgill, 2011).

Study 4 thus repeats the analysis from Study 3 while adding control measures for both morphological and lexical L1-L*n* similarities that are scalable to our diverse cross-linguistic data. Both of these measures were found in previous work to predict L*n* learners' speaking proficiency (Schepens, Van der Slik, & Van Hout, 2013a; Schepens et al., 2013b, 2016; van der Slik, van Hout, & Schepens, 2017). These previous works also contain scatterplots between speaking scores and similarity. We close by estimating the joint influence of all three types of similarity on L*n* Dutch speaking proficiency. This allows us to quantify—for the first time and on a large sample of languages—the relative influence of L1-L*n* similarity on the perceived speaking proficiency of L*n* learners of various language backgrounds, while controlling for other variables known to affect L*n* learning. As we described when we introduced the STEX database, language background accounts for a substantial proportion of the variability in L*n* Dutch speaking proficiency across the over 40,000 learners in our sample. Here we ask how much of the effect of L1 language background is due to L1-L*n* similarity.

4.1. Calculating morphological and lexical similarities across languages

For morphology, we measured the increase in morphological complexity in L*n* Dutch compared to the learner's L1 (Schepens et al., 2013a). This measure estimates the amount of additional morphology that a learner of Dutch has to learn, and thus resembles our phonological measure, in that both measures focus on additional features learners of Dutch have to acquire, and ignore features present in the L1 but absent in Dutch. For nine L1s, we were not able to obtain the relevant morphological data (Azerbaijani, Bengali, Éwé, Hebrew, Javanese, Nepali, Pashto, Slovenian, Urdu).

For lexical similarity, we used data from phylogenetic language trees based on expert cognacy judgements (Gray & Atkinson, 2003). This measure is high for languages with many shared cognates. This measure is likely to capture not only lexical, but also phonological and morphological similarities between the two languages as cognacy was determined traditionally using the comparative method. We thus expect this measure to be correlated with both phonological and

Table 3
Correlation matrix of adjusted Ln Dutch speaking scores, new sounds, new features, morphological, and lexical (dis)similarity measures. All correlations are based on estimates for the 53 L1 background for which both morphological and lexical similarity could be estimated.

Measure	Speaking scores	Phonological (dis)similarity	Morphological (dis)similarity
Phonological (dis)similarity	−0.47		
Morphological (dis)similarity	−0.59	0.49	
Lexical (dis)similarity	−0.69	0.48	0.77

morphological similarity. The phylogenetic tree only covers languages in the Indo-European language family, leaving the other 11 language families in our sample without a measure of lexical similarity. For these 11 language families (34 L1s), we substituted the minimal lexical similarity value observed for any of the Indo-European languages. We note that this likely allows our lexical similarity measure to capture effects of linguistic similarity that are not captured by the other two measures, simply because linguistic similarity tends to decrease the less closely languages are genealogically related.

4.2. Correlations between lexical, morphological, and phonological similarity

Table 3 gives the correlations for speaking proficiency with the various similarity measures. These linguistic (dis)similarity measures run from low (very similar) to high (very dissimilar). First, all similarity measures have a negative, significant correlation with speaking proficiency. The larger the similarity between an L1 and the Ln (Dutch), the higher speaking proficiency. The lexical and morphological similarity correlate more strongly with speaking proficiency scores than phonological similarity. Secondly, we can easily see that the three similarity measures are positively correlated with each other, meaning that they partially share their contribution in explaining speaking proficiency. Thirdly, lexical and morphological measures are more strongly correlated with each other ($r = 0.77$) than to new features (0.49 and 0.48), suggesting that phonological similarity can potentially explain additional variance in speaking scores on top of lexical and morphological similarity. See also SOM Fig. S1 for correlations with control variables and Table S2 for descriptive statistics for each L1 group.

4.3. Assessing the joint effects of lexical, morphological, and phonological similarity on Ln speaking proficiency

We repeated the analysis from Study 3 while also including lexical or morphological similarity or both as control predictor. Phonological similarity added significantly to a baseline model with morphological complexity ($\chi^2(1) = 10.79$, $p < .01$), and marginally to a baseline model with lexical similarity or a model with both the lexical and morphological measures ($\chi^2(1) = 3.76$, $p = .052$). Morphological and lexical similarity were significant predictors in all models that they were included in ($\chi^2(1) = 25.50$, $p < .0001$ and $\chi^2(1) = 44.83$, $p < .0001$, respectively). The effects of Study 3 replicate after controlling for linguistic similarity beyond phonology. We again find that native speakers of a language that is phonologically similar to Dutch tend to have higher L*n* speaking proficiency, even after controlling for (among other factors) the length of the learner's exposure to Dutch and the learner's age of acquisition. Fig. 8 illustrates the joint influence of lexical, morphological, and phonological L1-L*n* similarity on L*n* proficiency. In the final part of Study 4, we quantify this joint influence.

4.4. Effects of language background on Ln learning: How much does L1-Ln similarity matter?

The variance of the random by-L1 intercepts from the baseline model serves as an intuitive approximation of the maximal variance any similarity measure (or combination of similarity measures) could theoretically explain. Table 4 shows how much of this variance is explained by the different similarity measures and combinations thereof. Strikingly, all three similarity measures account for unique variance in L*n* speaking proficiency, and the triad of similarity measures performs significantly better than any single measure or pair of measures. Together, the three similarity measures explain almost all of the variance in L*n* proficiency due to L1 background (80%). That is, the linguistic similarity between an L1 and the L*n*—modeled here with just three degrees of freedom—can explain almost all of the variance due to L1 background across tens of thousands of learners of 62 different L1 backgrounds. This points to the enormous influence of L1-to-L*n* similarity during language learning (or specifically, L*n* speech production).

It is also notable that the present measure of lexical similarity explains most of the available variance due to L1 background ($70\% = (91/130) \times 100$). This particular result of Study 4 should, however, be interpreted with caution. Our measure of lexical similarity captures some aspects of phonological and morphological similarity, potentially including aspects of phonological and morphological similarity that are not yet captured by our *measures* of phonological and morphological similarity: lexical similarity was derived here from cognacy judgements and cognate status is arguably more reliably recoverable when phonological and morphological processes have not obscured it. Additionally, it is important to recall that the lexical similarity measure we employ was only available for Indo-European languages; for all non-Indo-European language, we set this measure to the maximum dissimilarity value observed any Indo-European language. This is visible in Fig. 8, where a substantial number of languages (most of them not Indo-European) cluster at the maximal value of lexical dissimilarity. This means that our lexical similarity measure *also* captures any dissimilarity between Indo-European languages and other language families that is not captured by our current measures of phonological and morphological similarity—including, for example, syntactic and other non-lexical similarities (we return to this point in the general discussion).

5. General discussion

The role of phonological transfer in L*n* learning and use is widely acknowledged, but large-scale tests across dozens of language backgrounds have been lacking. We investigated how the Dutch speaking proficiency of 48,219 learners from 62 different L1s is affected by the phonological similarity between their L1 and Dutch. We found that more similarities between the phonological systems of L1 and Dutch facilitate L*n* proficiency and thus likely L*n* learning. Specifically, we found that the internal structure of the new sounds matters: accounting for subcategorical similarities between the L1 and L*n* in terms of phonological features gives a significantly better predictor of L*n* speaking proficiency than a measure that assumes that any new sound category in an L*n* brings about the same degree of difficulty to L*n* learning and use. These results directly inform theories of L*n* learning and production. This study shows that currently available data on sound inventories and their feature representations can be used to capture phonological similarity effects on L*n* learning for many languages at the same time.

Furthermore, the effects of phonological similarity held beyond additional effects of morphological and lexical similarity, despite correlations between the three similarity measures. The phonological,

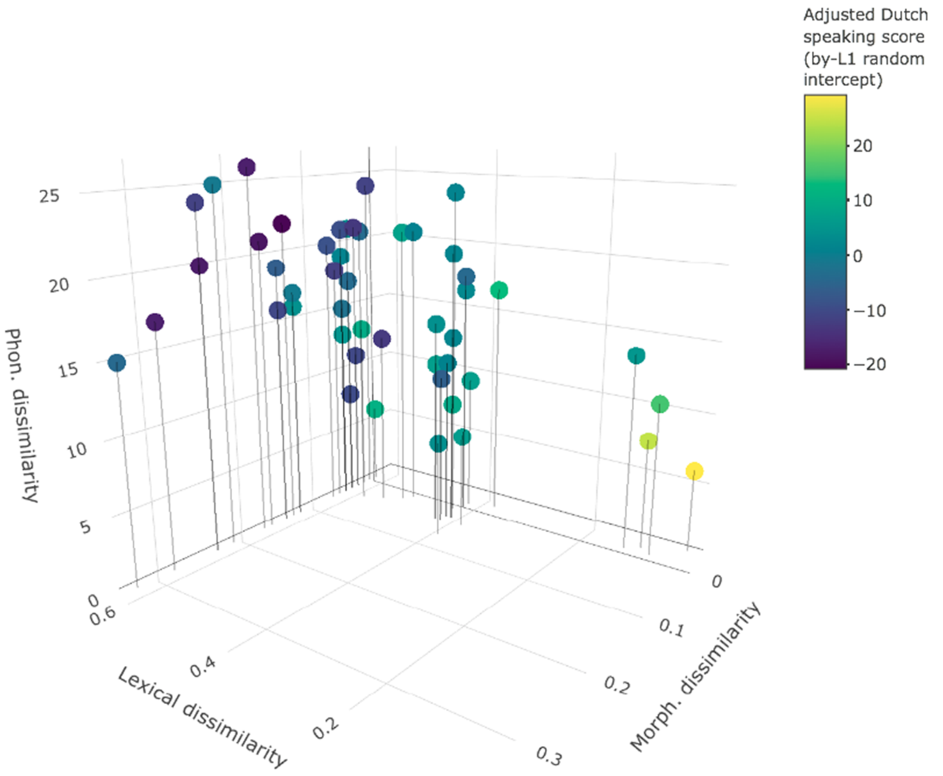


Fig. 8. The relation between adjusted Dutch speaking score (by-L1 random intercepts obtained from the baseline model, i.e. controlling for third factors) and measures of lexical, morphological, and phonological (dis)similarity. Each point shows one L1, colored by adjusted Dutch speaking proficiency. The yellow point on the right represents L1 German learners. The points in the back represent the different non-Indo-European L1s that are least closely related to Dutch. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Explained and remaining L1 variance by models containing no similarity measures (baseline), one similarity measure, or all similarity measures.

Models	Explained L1 variance	Remaining L1 variance
Baseline	0 (0%)	130
Lexical	91 (70%)	40
Phonological	58 (44%)	72
Morphological	41 (31%)	90
Lexical + morphological + phonological	104 (80%)	26

lexical and morphological components of language define their own sources in explaining L*n* learnability, but they overlap at the same time because they are dependent on each other. We first discuss these general effects, and how they might come to affect speaking proficiency. We end by discussing our approach.

5.1. Complementary effects of phonological, lexical, and morphological similarity on Ln learning

We have shown that adult L*n* learning generally depends on at least three linguistic domains. The combination of lexical, morphological, and phonological similarity measures together leads to the strongest reduction in unexplained variance. Although the three similarity measures we considered here are moderately to strongly correlated, lexical, morphological, and phonological similarity each have their independent contribution to the L*n* speaking proficiency. This is perhaps not entirely surprising given that our measure of proficiency was a compound measure, including ratings about various aspects of grammar, vocabulary, and pronunciation. Our approach here leaves open the extent to which transfer of specific linguistic properties from an L1 to an L*n* is driven by (i) L1-L*n* similarity with regard to only that property, (ii) L1-L*n* similarity with regard to related linguistic properties (e.g., overall syntactic similarity affecting the likelihood of transfer for a specific syntactic property), or (iii) overall L1-L*n* similarity across all linguistic domains. There is some evidence, for example, that learners sometimes over-generalize from one L1-L*n* similarity to (falsely) assuming other similarities (for review and references, see Pajak et al., 2016; Rothman, 2015). Because all three domain-specific measures define autonomous indices of L*n* learnability, there is not one domain that forms the overall greatest challenge to all Ln learners. What our approach approximates, however, is the cumulative effect of various linguistic similarities on the perceived proficiency of L*n* learners. And much like the ratings in STEX, this perception is likely to be affected by similarities across linguistic domains.

5.2. Similarity effects on Ln perception, Ln production, and perceived Ln proficiency

There are at least three ways in which linguistic similarities between an L1 and L*n* might come to affect the perceived speaking proficiency of an L*n* learner. The present approach does not distinguish between these potential sources for similarity effects, but rather investigates their joint impact on L*n* speaking proficiency. First, L1-to-L*n* transfer can affect learners' *perception* of L*n*. This is perhaps most evidence for the perception of phonological categories (Best, 1995), affecting the phonological representations learners acquire for the L*n*. Beginning learners especially might experience difficulty perceiving non-native phonological contrasts. A commonly cited example is the difficulty Japanese learners of English have in perceiving the distinction between /r/ and /l/, a contrast present in English but not in Japanese (Aoyama et al., 2004; Bradlow et al., 1997).

Second, L1-to-L*n* transfer might also directly affect the planning (linguistic encoding) and actual articulation of L*n* speech. As a consequence of either L1 transfer effects on perception or L1 transfer effects on production, learners' productions can deviate from native L*n* productions. This, too, can be exemplified for phonological encoding (e.g., Flege, 1987). Japanese learners of English often produce pronunciations of /r/ and /l/ that deviate phonetically from native productions (Aoyama et al., 2004; Bradlow et al., 1997). This contrasts with learners who have an L1 phonology more similar to English. Similar effects have also been observed for morphological, lexical, or syntactic aspects of L*n* learners' productions (e.g., in terms of the types of errors made by learners, Díaz-Negrillo, Ballier, & Thompson, 2013; Granger, Gilquin, & Meunier, 2015; Lüdeling, Hirschmann, & Shadrova, 2017).

Finally, a third similarity effect on the *perceived* speaking proficiency of L*n* learners originates in native L*n* listeners' perception (Porretta, Kyröläinen, & Tucker, 2015). Just as L*n* learners' perception of the L*n* is affected by their L1, native L*n* listeners' perception of L1-accented L*n* speech is affected by their native L*n* knowledge (Best, 1993, 1995; Escudero, 2005; Flege, 1993; Flege, Schirru, & MacKay, 2003), although this influence seems to decrease with increasing exposure to the L1-accented speech (Adank, Evans, Stuart-Smith, & Scott, 2009; Banks, Gowen, Munro, & Adank, 2015; Bradlow & Bent, 2008; Clarke & Garrett, 2004; Porretta, Tucker, & Järvikivi, 2016; Xie et al., 2018). For example, similar-sounding non-native pronunciations of two different L*n* categories will reduce the intelligibility of accented speech.

5.3. Complementing traditional approaches to Ln learning with big data: Future directions

The approach taken in the present study effectively complements other approaches to the study of L*n* acquisition and use, such as production and perception experiments in the lab (e.g. Bradlow et al., 1997; Flege et al., 1999) and studies in the classroom (Derwing, 2008). Different from the majority of previous work, the present approach allowed us to assess the effect of linguistic similarity across a large, heterogeneous, and linguistically diverse group of learners and L1 backgrounds.

The present approach makes a number of simplifying assumptions. In particular, our new phonological similarity measures assume that the role of multiple competitors in learning new sounds can be captured by taking the most similar sound for every L1 sound. Our results validate this as a feasible approach. However, there is evidence that the difficulty of perceiving and producing a new L*n* sound depends on its specific position in **phonetic, articulatory, or perceptual—rather than phonological–space** (e.g. Aoyama et al., 2004; Bradlow et al., 1997). Further, there is evidence that a sound's position in this space is best understood to relative to its surrounding sound*s*, rather than just the most similar sound (Aoyama et al., 2004; Best, 1995; Bradlow et al., 1997; Escudero, 2005). Relatedly, some theories of L*n* learning (Pajak, 2012) emphasize that learners need to acquire phonetic (as well as articulatory and perceptual) *distributions*, and that differences and similarities in these distributions affect L*n* learning. For example, when the feature [long] is present in the L1 on vowels only and the learner needs to acquire a long consonant, generalization from long vowels to long consonants may be easier (Tsukada, Hirata, & Roengpitya, 2014). The big data approach can theoretically be extended to model such more fine-grained articulatory and perceptual effects on L*n*. As of yet, however, large scale articulatory, phonetic, or perceptual databases that contain information about dozens of languages do not yet exist (cf. Dingemanse, Torreira, & Enfield, 2013).

Another obvious limitation of the present study pertains to what types of similarity our measures capture. With regard to phonological similarity, we only considered the inventory and similarity of sound categories. We did not consider (dis)similarities between the L1 and L*n* in terms of **supra-segmental** properties, including lexical stress and intonation. For example, difficulty in producing native-like **prosodic** emphasis and phrasing can interfere strongly with native listeners' ability to segment, and thus understand, non-native speech (e.g. Munro, 2008). Relevant databases might become available in the future (see e.g. Gallagher & Graff, 2012).

Future work should also explicitly address suprasegmental L1-L*n* differences in **phonotactic** rules and syllable structures. For example, native speakers of L1 Dutch use their Dutch version of the Obligatory Contour Principle (OCP) to segment Dutch. Native speakers of Mandarin Chinese, which is not restricted by an OCP-Place constraint, do not make use of OCP-Place constraints when they start to learn Dutch. The benefit from OCP-Place as a cue for L*n* speech segmentation may depend on the prevalence of OCP-Place constraints in the language background of the learner (Boll-Avetisyan, 2012). Similarly, Japanese learners of languages like Dutch or English struggle with the production of syllable final consonants (Japanese syllables either end in a vowel or a nasal), pronouncing words like "MacDonald's" as [makudonanodo] (Major, 2008). Similar limitations apply to our measure of lexical and morphological similarity (for discussion, see Schepens et al., 2013b, 2013a, 2016; van der Slik et al., 2017).

Relatedly, our studies did not consider the **relative frequency and importance of different linguistic properties**. For example, some phonemic contrasts of Dutch—while attested—occur less frequently or are less critical to language understanding (e.g., because the contrast distinguishes between fewer words, a property known as functional load, Jakobson, 1931; Mathesius, 1931; Wedel, Kaplan, & Jackson, 2013). Everything else being equal, such phonemic contrasts might contribute less to the perceived difficulty of understanding an L*n* Dutch speaker. Future studies could integrate lexical information from, for example, CELEX (Baayen, Piepenbrock, & Van Rijn, 1993) to weigh the predicted similarity effects of individual phonemic categories by their functional load (see Burchill & Jaeger, 2017). There are large **frequency** differences in the usage of Dutch sounds (Luyckx, Kloots, Coussé, & Gillis, 2007). Moreover, sounds that are most often different across L1s and Dutch have a relatively low frequency in Dutch (e.g. front rounded vowels rank consistently low on frequency of use; /γ/ ranks 26 in the Dutch sound frequency ranking of Luyckx, Kloots, Coussé, & Gillis, 2007). One may argue that less frequent sounds are less important for a speaker's production of intelligible speech. Future work should try to carve out the role of frequency, both in the L1 and in the L*n*.

Another trade-off relates to the **granularity of the dependent measure** we analyzed: the speaking scores provide a compound measure based on ratings that cover various aspects of speaking proficiency, including (morpho-)syntactic, content-focused and pronunciation-focused ratings. This can be seen both as an advantage and as a disadvantage of the present approach. On the one hand, it serves the present purpose of estimating the cumulative effect of linguistic similarity on the perceived proficiency of non-native talkers. On the other hand, the mixed nature of ratings that combine into the proficiency score likely introduces some statistical noise into our analyses. Future work can explore whether different ratings are affected differently by different similarity measures.

Finally, we have assumed so far that linguistic similarities are not correlated with learner variables beyond the ones that our analysis controlled for (education and gender). Additional learner variables might affect our results only when they are both (a) correlated with linguistic similarities, and (b) are likely to affect the STEX test scores. Obvious candidates are cross-cultural differences when they relate to specific language groups or families, such as differences in language attitude (e.g., the cross-cultural differences in the perceived acceptability of producing accented or otherwise non-native speech). However, cross-cultural differences may not be problematic if they do not affect test scores. There are many studies on the topic of culture and language assessment (e.g. Roever & McNamara, 2006), and cross-cultural studies on the link between language and cultural differences in reasoning and thinking (e.g. Ji, Zhang, & Nisbett, 2004). We inspected the random intercepts of the L1s in our study and we did not observe potential biases, but future big data approaches should aim to include measures of cross-cultural differences.

To conclude, the perceived accentedness of non-native talkers can have social consequences. There is evidence that accentedness can affect the ability to obtain employment, as well as perceptions of intelligence, education status, and the like (Fuertes et al., 2012; Gluszek & Dovidio, 2010). In this context, it is worth repeating the central finding of Study 4: a person's native language has a *large* effect on their perceived speaking proficiency in the L*n*, and this effect is strongly determined by the similarity between the L1 and L*n*. For example, only the best 5% of Arabic learners of Dutch in our sample scored higher than the worst 50% of German learners. Critically, these effects are outside the control of the learner, and they hold over and above effects that might be reasoned to be at least partially under the learner's control (such as the amount of exposure to Dutch and the motivation to learn Dutch).

Declaration of Competing Interest

The authors do not have any competing interests.

Acknowledgments

The authors thank Steven Moran for the use of the PHOIBLE database and the secretary of the State Examination of Dutch as a Second Language for providing the data used in our study. Example elicitation instruments can be downloaded from https://www.hetcvte.nl/item/voorbeeldexamens_en. We thank Theo Bongaerts, Crystal Lee, Bozena Pajak, and, in particular, Xin Xie for valuable feedback. This study was partially supported by a Fulbright Fellowship awarded to JS to visit the Human Language Processing Lab at the University of Rochester, and by funding from NICHD R01 HD075797 to TFJ. The views expressed here are those of the authors, and do not necessarily reflect those of the funding agencies.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2019.104056.

References

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance, 35*(2), 520.

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: An individual differences approach. *Language Learning, 62*, 49–78. https://doi.org/10.1111/j.1467-9922.2012.00706.x.

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., & Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics, 32*(2), 233–250. https://doi.org/10.1016/S0095-4470(03)00036-6.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005.

Baayen, R. H., Piepenbrock, R., & Van Rijn, H. (1993). *The CELEX lexical database [cd-rom].* Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Babcock, L., Stowe, J. C., Maloof, C. J., Brovetto, C., & Ullman, M. T. (2012). The storage and composition of inflected forms in adult-learned second language: A study of the influence of length of residence, age of arrival, sex, and other factors. *Bilingualism: Language and Cognition, 15*(4), 820–840.

Banks, B., Gowen, E., Munro, K. J., & Adank, P. (2015). Cognitive predictors of perceptual adaptation to accented speech. *The Journal of the Acoustical Society of America, 137*(4), 2015–2024.

Bardel, C., & Falk, Y. (2007). The role of the second language in third language acquisition: The case of Germanic syntax. *Second Language Research, 23*(4), 459–484. https://doi.org/10.1177/0267658307080557.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. Retrieved from http://CRAN.R-project.org/package=lme4.

Best, C. T. (1993). Emergence of language-specific constraints in perception of non-native speech: A window on early phonological development. In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. McNeilage, & J. Morton (Eds.). *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 289–304). . https://doi.org/10.1007/978-94-015-8234-6_24.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), Speech perception and linguistic experience: Issues in cross-language research (pp. 171–206). Retrieved from http://ci.nii.ac.jp/naid/10018033931/.

Birdsong, D. (2014). Dominance and age in bilingualism. *Applied Linguistics, 35*(4), 374–392.

Bohn, O.-S. (2002). *On phonetic similarity.* Trier: Wissenschaftlicher Verlag Trier.

Bohn, O.-S., & Flege, J. E. (1992). The production of new and similar vowels by adult German learners of English. *Studies in Second Language Acquisition, 14*(2), 131–158.

Bohnacker, U. (2006). When swedes begin to learn german: From V2 to V2. *Second Language Research, 22*(4), 443–486. https://doi.org/10.1191/0267658306sr275oa.

Boll-Avetisyan, N. (2012). Phonotactics and its acquisition, representation, and use: An experimental-phonological study. LOT, 298. Retrieved from http://dspace.library.uu.nl/handle/1874/241655.

Bongaerts, T., Van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition, 19*(4), 447–465.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English/r/and/l: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America, 101*(4), 2299–2310.

Brennan, E. M., & Brennan, J. S. (1981). Measurements of accent and attitude toward Mexican-American speech. *Journal of Psycholinguistic Research, 10*(5), 487–501.

Brown, C. (1998). The role of the L1 grammar in the L2 acquisition of segmental structure. *Second Language Research, 14*(2), 136–193. https://doi.org/10.1191/026765898669508401.

Brown, C. (2000). The interrelation between speech perception and phonological acquisition from infant to adult. *Second Language Acquisition and Linguistic Theory,* 4–63.

Burchill, Z., & Jaeger, T. F. (2017). Grounding sound change in ideal observer models of perception. In Proceedings of the 7th workshop on cognitive modeling and computational linguistics (CMCL 2017) (pp. 20–28).

Cenoz, J. (2001). The effect of linguistic distance, L2 status and age on cross-linguistic influence in third language acquisition. In J. Cenoz, B. Hufeisen, & U. Jessner (Eds.). *Cross-linguistic influence in third language acquisition: Psycholinguistic perspectives* (pp. 8–20). Clevedon, UK: Multilingual Matters.

Chang, C. B. (2015). Determining Cross-linguistic phonological similarity between segments. In E. Raimy & C. E. Cairns (Eds.), The segment in phonetics and phonology (pp. 199–217). https://doi.org/10.1002/9781118555491.ch9.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* New York: Harper & Row.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America, 116*(6), 3647–3658.

Cook, V. (2013). *Second language learning and language teaching.* Routledge.

Croft, W. (1990). *Typology and universals.* Cambridge: Cambridge University Press.

Crothers, J. H., Lorentz, J. P., Sherman, D. A., & Vihman, M. M. (1979). Handbook of phonological data from a sample of the world's languages: A report of the stanford phonology archive.

Cysouw, M., Dediu, D., & Moran, S. (2012). Comment on "phonemic diversity supports a serial founder effect model of language expansion from Africa". *Science, 335*(6069), https://doi.org/10.1126/science.1208841 657–657.

DeKeyser, R. M. (2012). Interactions between individual differences, treatments, and structures in SLA. *Language Learning, 62*, 189–200. https://doi.org/10.1111/j.1467-9922.2012.00712.x.

Derwing, T. M. (2008). Curriculum issues in teaching pronunciation to second language learners. *Phonology and Second Language Acquisition,* 347–369.

Díaz-Negrillo, A., Ballier, N., & Thompson, P. (2013). *Automatic treatment and analysis of learner corpus data, Vol. 59*. John Benjamins Publishing Company.

Dijkstra, T., Miwa, K., Brummelhuis, B., Sappelli, M., & Baayen, R. H. (2010). How cross-language similarity and task demands affect cognate recognition. *Journal of Memory and Language, 62*(3), 284–301.

Dingemanse, M., Torreira, F., & Enfield, N. J. (2013). Is "Huh?" a universal word? Conversational infrastructure and the convergent evolution of linguistic items. *PloS One, 8*(11), e78273.

Dunn, M., Greenhill, S. J., Levinson, S. C., & Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature, 473*(7345), 79–82. https://doi.org/10.1038/nature09923.

Eckman, F. R., Elreyes, A., & Iverson, G. K. (2003). Some principles of second language phonology. *Second Language Research, 19*(3), 169–208. https://doi.org/10.1191/0267658303sr219oa.

Escudero, P. (2005). Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization (Universiteit Utrecht). Retrieved from https://www.lotpublications.nl/linguistic-perception-and-second-language-acquisition-linguistic-perception-and-second-language-acquisition-explaining-the-attainment-of-optimal-phonological-categorization.

Escudero, P., Broersma, M., & Simon, E. (2013). Learning words in a third language: Effects of vowel inventory and language proficiency. *Language and Cognitive Processes, 28*(6), 746–761. https://doi.org/10.1080/01690965.2012.662279.

Flege, J. E. (1981). The phonological basis of foreign accent: A hypothesis*. *TESOL Quarterly, 15*(4), 443–455. https://doi.org/10.2307/3586485.

Flege, J. E. (1987). The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics, 15*(1), 47–65.

Flege, J. E. (1993). Production and perception of a novel, second-language phonetic contrast. *The Journal of the Acoustical Society of America, 93*(3), 1589–1608. https://doi.org/10.1121/1.406818.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research, 92*, 233–277.

Flege, J. E. (1997). English vowel production by Dutch talkers: More evidence for the "similar" vs "new" distinction : Second-Language Speech Structure and Process. In A. James & J. Leather (Eds.), Second-language speech: Structure and process (pp.

11–52). Retrieved from https://doi.org/10.1515/9783110882933.11.

Flege, J. E. (2003). Assessing constraints on second-language segmental production and perception. *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities,* 319–355.

Flege, J. E. (2018). It's input that matters most, not age. *Bilingualism: Language and Cognition, 21*(5), 919–920. https://doi.org/10.1017/S136672891800010X.

Flege, J. E., MacKay, I. R., & Meador, D. (1999). Native Italian speakers' perception and production of English vowels. *The Journal of the Acoustical Society of America, 106*(5), 2973–2987.

Flege, J. E., Schirru, C., & MacKay, I. R. A. (2003). Interaction between the native and second language phonetic subsystems. *Speech Communication, 40*(4), 467–491. https://doi.org/10.1016/S0167-6393(02)00128-0.

Fuertes, J. N., Gottdiener, W. H., Martin, H., Gilbert, T. C., & Giles, H. (2012). A meta-analysis of the effects of speakers' accents on interpersonal evaluations. *European Journal of Social Psychology, 42*(1), 120–133. https://doi.org/10.1002/ejsp.862.

Gallagher, G., & Graff, P. (2012). The role of similarity in phonology. *Lingua, 122*(2), 107–111. https://doi.org/10.1016/j.lingua.2011.11.002.

Gimson, A. C. (1962). *An introduction to the pronunciation of English.* London: Edward Arnold.

Gluszek, A., & Dovidio, J. F. (2010). The way they speak: A social psychological perspective on the stigma of nonnative accents in communication. *Personality and Social Psychology Review*. https://doi.org/10.1177/1088868309359288.

Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research, 29*(3), 311–343.

Granger, S., Gilquin, G., & Meunier, F. (2015). *The Cambridge handbook of learner corpus research.* Cambridge University Press.

Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature, 426*(6965), 435–439. https://doi.org/10.1038/nature02029.

Gray, R. D., & Watts, J. (2017). Cultural macroevolution matters. *Proceedings of the National Academy of Sciences, 114*(30), 7846–7852. https://doi.org/10.1073/pnas.1620746114.

Halle, M. (1973). Stress rules in English: A new version. *Linguistic Inquiry, 4*(4), 451–464.

Hammarström, H. (2016). Linguistic diversity and language evolution. *Journal of Language Evolution, 1*(1), 19–29. https://doi.org/10.1093/jole/lzw002.

Haugen, E. I. (1966). Linguistics and language planning. *Sociolinguistics,* 50–71.

Hayes, B. (2009). *Introductory phonology.* West Sussex: Blackwell.

International Phonetic Association (1999). *Handbook of the International Phonetic Association (IPA).* Cambridge, UK: Cambridge University Press.

Ionin, T., & Wexler, K. (2002). Why is 'is' easier than '-s'?: Acquisition of tense/agreement morphology by child second language learners of English. *Second Language Research, 18*(2), 95–136. https://doi.org/10.1191/0267658302sr195oa.

Iverson, P., & Evans, B. G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers). *The Journal of the Acoustical Society of America, 126*(2), 866–877. https://doi.org/10.1121/1.3148196.

Jaeger, T. F., Graff, P., Croft, W., & Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology, 15*(2), 281–319.

Jakobson, R. (1931). Prinzipien der historischen Phonologie. *Prague Linguistic Circle Papers, 4*, 246–267.

Jakobson, R. (1941). *Kindersprache, Aphasie, und allgemeine Lautgesetze.* Uppsala: Almqvist and Wiksell.

Jakobson, R. (1968). *Child language: Aphasia and phonological universals (A. Keller, Trans.).* The Hague: Walter de Gruyter.

Jarvis, S. (2000). Methodological rigor in the study of transfer: Identifying L1 influence in them interlanguage lexicon. *Language Learning, 50*(2), 245–309. https://doi.org/10.1111/0023-8333.00118.

Jarvis, S. (2015). Influences of previously learned languages on the learning and use of additional languages. In M. Juan-Garau & J. Salazar-Noguera (Eds.), Content-based language learning in multilingual educational environments (pp. 69–86). https://doi.org/10.1007/978-3-319-11496-5_5.

Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition.* New York: Routledge.

Ji, L.-J., Zhang, Z., & Nisbett, R. E. (2004). Is it culture or is it language? Examination of language effects in cross-cultural research on categorization. *Journal of Personality and Social Psychology, 87*(1), 57.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology, 21*(1), 60–99. https://doi.org/10.1016/0010-0285(89)90003-0.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics, 50*(2), 93–107. https://doi.org/10.3758/BF03212211.

LaCross, A. (2015). Khalkha Mongolian speakers' vowel bias: L1 influences on the acquisition of non-adjacent vocalic dependencies. *Language, Cognition and Neuroscience,* 1–15. https://doi.org/10.1080/23273798.2014.915976.

Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers.* Michigan: University of Michigan Press.

Lenneberg, E. H. (1967). *Biological foundations of language, Vol. 68*. New York: Wiley.

Lev, M., Stark, T., & Chang, W. (2012). South American phonological inventory database. Retrieved from http://linguistics.berkeley.edu/saphon/en/.

Lippi-Green, R. (2012). *English with an accent: Language, ideology and discrimination in the United States.* Routledge.

Lüdecke, D. (2019). sjPlot: Data Visualization for Statistics in Social Science. https://doi.org/10.5281/zenodo.1308157.

Lüdeling, A., Hirschmann, H., & Shadrova, A. (2017). Linguistic models, acquisition theories, and learner corpora: Morphological productivity in SLA research exemplified by complex verbs in German. *Language Learning, 67*(S1), 96–129. https://doi.org/10.1111/lang.12231.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE, 5*(1), e8559. https://doi.org/10.1371/journal.pone.0008559.

Luyckx, K., Kloots, H., Coussé, E., & Gillis, S. (2007). Klankfrequenties in het Nederlands. In Tussen taal, spelling en onderwijs: essays bij het emeritaat van Frans Daems (pp. 141–154). Retrieved from http://hdl.handle.net/1854/LU-430290.

Maddieson, I. (1984). *Patterns of sounds.* Cambridge: Cambridge University Press.

Maddieson, I., & Precoda, K. (1990). Updating UPSID. *UCLA working papers in phonetics: Vol. 74*, (pp. 104–111). Department of Linguistics, UCLA.

Major, R. C. (1992). Transfer and developmental factors in second language acquisition of consonant clusters. *New Sounds, 90*, 128–136.

Major, R. C. (2008). Transfer in second language phonology. In J. G. H. Edwards, & M. L. Zampini (Vol. Eds.), *Phonology and second language acquisition: Vol. 36*, (pp. 63–94). Amsterdam: John Benjamins Publishing.

Mathesius, V. (1931). Zum Problem der Belastungs-und kombinationsfähigkeit der Phoneme. *Prague Linguistic Circle Papers, 4*, 148–152.

McCloy, D. R., Moran, S., & Wright, R. A. (2013). Revisiting 'The role of features in phonological inventories'. Retrieved from http://students.washington.edu/drmccloy/pubs/McCloyEtAl2013_cunyFeatureConf.pdf.

McWhorter, J. H. (2007). *Language interrupted: Signs of non-native acquisition in standard language grammars.* New York: Oxford University Press.

Moran, S. (2012). Phonetics information base and lexicon (University of Washington). Retrieved from https://digital.lib.washington.edu/researchworks/handle/1773/22452.

Moran, S., McCloy, D., & Wright, R. (2012). Revisiting population size vs. phoneme inventory size. *Language, 88*(4), 877–893.

Moran, S., McCloy, D., & Wright, R. (Eds.) (2014). PHOIBLE Online. Retrieved from http://phoible.org/.

Munro, M. J. (2003). A primer on accent discrimination in the Canadian context. *TESL Canada Journal, 20*(2), 38–51.

Munro, M. J. (2008). Foreign accent and speech intelligibility. In J. G. Hansen Edwards, & M. L. Zampini (Eds.). *Phonology and second language acquisition* (pp. 193–219). Amsterdam: John Benjamins.

Nettle, D. (1999). *Linguistic diversity.* Oxford: Oxford University Press.

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning.* Cambridge University Press.

Odlin, T. (2012). Crosslinguistic Influence in second language acquisition. *The Encyclopedia of Applied Linguistics*. https://doi.org/10.1002/9781405198431.wbeal0292.

Otwinowska-Kasztelanic, A. (2009). Raising awareness of cognate vocabulary as a strategy in teaching English to Polish adults. *Innovation in Language Learning and Teaching, 3*(2), 131–147. https://doi.org/10.1080/17501220802283186.

Pajak, B. (2012). Inductive inference in non-native speech processing and learning (University of California at San Diego). Retrieved from http://www.bcs.rochester.edu/people/bpajak/pdfs/Pajak_2012_diss.pdf.

Pajak, B., Fine, A. B., Kleinschmidt, D. F., & Jaeger, T. F. (2016). Learning additional languages as hierarchical probabilistic inference: Insights from first language processing. *Language Learning, 66*(4), 900–944.

Pajak, B., & Levy, R. (2014). The role of abstraction in non-native speech perception. *Journal of Phonetics, 46*, 147–160. https://doi.org/10.1016/j.wocn.2014.07.001.

PHOIBLE Online (2014). Retrieved from http://phoible.org/.

Pica, T. (1983). Adult acquisition of English as a second language under different conditions of exposure. *Language Learning, 33*(4), 465–497. https://doi.org/10.1111/j.1467-1770.1983.tb00945.x.

Piske, T., Flege, J. E., MacKay, I. R., & Meador, D. (2002). The production of English vowels by fluent early and late Italian-English bilinguals. *Phonetica, 59*(1), 49–71.

Porretta, V., Kyröläinen, A.-J., & Tucker, B. V. (2015). Perceived foreign accentedness: Acoustic distances and lexical properties. *Attention, Perception, & Psychophysics, 77*(7), 2438–2451. https://doi.org/10.3758/s13414-015-0916-3.

Porretta, V., Tucker, B. V., & Järvikivi, J. (2016). The influence of gradient foreign accentedness and listener experience on word recognition. *Journal of Phonetics, 58*, 1–21.

R Core Team (2018). R: A language and environment for statistical computing. Retrieved from https://www.R-project.org/.

Ringbom, H. (2007). *Cross-linguistic similarity in foreign language learning.* Clevedon, UK: Multilingual Matters.

Roever, C., & McNamara, T. (2006). Language testing: The social dimension. *International Journal of Applied Linguistics, 16*(2), 242–258.

Rothman, J. (2015). Linguistic and cognitive motivations for the Typological Primacy Model (TPM) of third language (L3) transfer: Timing of acquisition and proficiency considered. *Bilingualism: Language and Cognition, 18*(2), 179–190.

Schepens, J. (2015). Bridging linguistic gaps: The effects of linguistic distance on the adult learnability of Dutch as an additional language. Retrieved from https://www.lotpublications.nl/Documents/383_fulltext.pdf.

Schepens, J., Van der Slik, F., & Van Hout, R. (2013a). Learning complex features: A morphological account of l2 learnability. *Language Dynamics and Change, 3*(2), 218–244. https://doi.org/10.1163/22105832-13030203.

Schepens, J., Van der Slik, F., & Van Hout, R. (2013b). The effect of linguistic distance across Indo-European mother tongues on learning Dutch as a second language. In L. Borin, & A. Saxena (Eds.). *Approaches to measuring linguistic differences* (pp. 199–230). Berlin: De Gruyter Mouton.

Schepens, J., van der Slik, F., & van Hout, R. (2016). L1 and L2 distance effects in learning L3 Dutch. *Language Learning, 66*(1), 224–256. https://doi.org/10.1111/lang.12150.

Schepens, J., van Hout, R., & Jaeger, T. F. (2019). Dataset for: "Big data suggest strong constraints of linguistic similarity on adult language learning". https://doi.org/10.5281/zenodo.2863533.

Schumann, J. H., Crowell, S. E., Jones, N. E., Lee, N., & Schuchert, S. A. (2004). *The neurobiology of learning: Perspectives from second language acquisition.* Routledge.

Skirgård, H., Roberts, S. G., & Yencken, L. (2017). Why are some languages confused for others? Investigating data from the Great Language Game. *PLOS ONE, 12*(4), e0165934. https://doi.org/10.1371/journal.pone.0165934.

Stevens, G. (1999). Age at immigration and second language proficiency among foreign-born adults. *Language in Society, 28*(4), 555–578.

Stevens, G. (2006). The age-length-onset problem in research on second language acquisition among immigrants. *Language Learning, 56*(4), 671–692.

Strange, W., Weber, A., Levy, E. S., Shafiro, V., Hisagi, M., & Nishi, K. (2007). Acoustic variability within and across German, French, and American English vowels: Phonetic context effects. *The Journal of the Acoustical Society of America, 122*(2), 1111–1129.

Trudgill, P. (2011). *Sociolinguistic typology: Social determinants of linguistic complexity.* Oxford: Oxford University Press.

Tsukada, K., Hirata, Y., & Roengpitya, R. (2014). Cross-language perception of Japanese vowel length contrasts: Comparison of listeners from different first language backgrounds. *Journal of Speech Language and Hearing Research, 805*. https://doi.org/10.1044/2014_JSLHR-S-12-0416.

UNESCO (2011). The World Bank, Institute for Statistics, School enrollment, secondary (% gross). Retrieved January 3, 2013, from http://data.worldbank.org/indicator/SE.SEC.ENRR.

van der Slik, F., van Hout, R., & Schepens, J. (2017). The role of morphological complexity in predicting the learnability of an additional language: The case of La (additional language) Dutch. *Second Language Research*. https://doi.org/10.1177/0267658317691322.

Vanhove, J. (2013). The critical period hypothesis in second language acquisition: A statistical critique and a reanalysis. *PLoS ONE, 8*(7), e69172. https://doi.org/10.1371/journal.pone.0069172.

Vanhove, J., & Berthele, R. (2015a). Item-related determinants of cognate guessing in multilinguals. In G. De Angelis, U. Jessner, & M. Kresić (Eds.), Crosslinguistic influence and crosslinguistic interaction in multilingual language learning (pp. 95–118).

Vanhove, J., & Berthele, R. (2015b). The lifespan development of cognate guessing skills in an unknown related language. *International Review of Applied Linguistics in Language Teaching, 53*(1), 1–38.

Wedel, A., Kaplan, A., & Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition, 128*(2), 179–186.

Weinreich, U. (1963). *Languages in contact.* The Hague: Mouton.

Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *The Journal of the Acoustical Society of America, 143*(4), 2013–2031.

Yu, L., & Odlin, T. (2015). *New perspectives on transfer in second language learning.* Bristol: Multilingual Matters.