PRIFYSGOL BANGOR / BANGOR UNIVERSITY

Taking taxonomy seriously in Linguistics: intelligibility as a criterion of demarcation between languages and dialects.

Tamburelli, Marco

Lingua

DOI:
https://doi.org/10.1016/j.lingua.2021.103068

Published: 01/06/2021

Peer reviewed version

Cyswllt i'r cyhoeddiad / Link to publication

Dyfyniad o'r fersiwn a gyhoeddwyd / Citation for published version (APA):
Tamburelli, M. (2021). Taking taxonomy seriously in Linguistics: intelligibility as a criterion of demarcation between languages and dialects. *Lingua*, *256*, Article 103068. https://doi.org/10.1016/j.lingua.2021.103068

Hawliau Cyffredinol / General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

24. Apr. 2024

Taking taxonomy seriously in Linguistics: intelligibility as a criterion of demarcation between languages and dialects.

Abstract

In Linguistics, a principled definition of what constitutes a ‘language’ in opposition to a ‘dialect’ has been notoriously elusive. The intelligibility criterion, possibly the only criterion that could form the basis of such definition, has often been considered inadequate, leading to the widespread conclusion that languages may not be linguistically definable objects at all (e.g. Chambers and Trudgill, 1996).

This paper reconsiders some of the objections typically raised against the intelligibility criterion and argues that one of these objections — namely that intelligibility is a scale to which no meaningfully discernible segmentation may be applied— can be formulated as a testable empirical claim. Three experiments are then presented with the explicit aim to test this claim.

Results indicate that, contrary to what has been frequently claimed, the intelligibility scale does allow for potentially meaningful segmentation, providing empirical evidence in favour of adopting intelligibility as an empirically sound criterion of demarcation for the identification of languages and dialects.

Keywords: intelligibility criterion, linguistic taxonomy, languages, dialects.

1. Introduction

A systematic taxonomy of a discipline's objects of inquiry is at the basis of scientific enterprises (see for example Feigelson, 2012, on astronomy; Gupta, 2007, on genetics; Hospenthal & Rinaldi, 2007, on diagnostic medicine; Wheeler, 2004, on biology). Similarly, many areas within linguistics depend on a definition of the concept "language" as the basis for their field of inquiry, language enumeration being perhaps the most obvious example. We can only count the languages of the world and – by extension – the languages of Asia, Africa, or the number of endangered languages in Europe if we have some criteria for identifying the entity "language", particularly in opposition to and distinguished from that of (its) "dialects" (e.g. Moseley, 2008). While the dependence of inquiry upon taxonomic classification may not be so straightforward in all linguistic sub-disciplines, examples of such dependence abound. Studies on bi- and multi-lingualism, for instance, often necessitate a definition of "language", as the question of who speaks two or more languages can only be answered (and, arguably, fully addressed) if we can identify what qualify as "two or more languages", a concept that ultimately relies on defining the entity "language" (for an overview of how defining "language" affects multilingualism research, see Kemp, 2009). Similarly, identification and understanding of language contact phenomena is predicated on knowledge of what constitutes two or more languages being in contact as opposed to "just […] dialect mixture" (Appel & Muysken, 2005:3. See also Thomason, 2001). The study of linguistic rights is perhaps even more desperately dependent on identifying what qualifies as a "language". As Dunbar put it:

> "While language is referred to in many international instruments, none address the fundamental question of what constitutes a language, of what forms of expression are entitled to protection" (2001: 96. Emphasis mine. See also Kibbee, 1998; Tulloch, 2006).

Dunbar's point echoes a view that is widespread in the sciences, namely that taxonomic and classificatory understanding is fundamental particularly – though not exclusively – to the development of conservation efforts (e.g. Lyal et al., 2008; Mace, 2004; Peterson, 2006; Wheeler, 2004; among many others). Despite this, a definition of "language" - particularly in opposition to that of "dialect" - has been elusive, and an increasing number of language researchers have accepted that "[l]inguists have failed to determine criteria by which languages can be distinguished from dialects" (Fasold, 2005:1. See also De Swaan, 1991; Romaine, 2000, *inter alia*). It is probably this perceived failure that has led to a tendency for linguists to avoid the question altogether, with a general "linguists' refusal to address the language-dialect business head on" (Nunberg, 1997:675. For examples of this avoidance

strategy, see Benincà & Price, 2000; Comrie, 2009; Posner, 1996). Despite its elusiveness, however, an objective definition has often been seen as a desideratum at least since Kloss (1967), who suggested that it was possible to define "language" as a dialect cluster that forms a "linguistic unit" (1967:29) which he calls a language by *Abstand,* definable independently of socio-political bias, and thus separately from what he called "sociological" entities, namely languages by *Ausbau*. More recently, Salminen (2007) pointed out the possibility of a definition entirely based on the structural properties of a language as opposed to the mere ideological construction and socio-political achievements of its speakers (for an overview of the pitfalls of a purely sociological / socio-political definition, see Author, 2014). As Salminen (2007) put it:

> "While there certainly are borderline cases, not least in Europe, it is usually quite easy to say which linguistic isoglosses amount to language boundaries and which do not, and the truly problematic cases are better regarded as challenges rather than obstacles" (2007: 211).

A~~The same~~ similar stance is taken by the Ethnologue (Lewis, Simons, & Fennig, 2014) and by the Encyclopedia of the World's Endangered Languages (Moseley, 2008), which put structural- linguistic considerations at the centre of their classifications. A perhaps more developed version of this position, factoring in the communicative properties of language, is found in Dixon (1997):

> "[o]nce political considerations are firmly discarded, it is generally not a difficult matter to decide whether one is dealing with one language or with more than one in a given situation." (1997: 7).

On this basis, Dixon calls upon the concept of intelligibility as a criterion of demarcation for the term "language" in a "linguistic sense" (1997: 7), stating that "two forms of speech which are mutually intelligible are regarded as dialects of one language" (1997: 7).

These authors are not alone in regarding intelligibility as the criterion of demarcation between "languages" and "dialects". There is at least one discipline within linguistics which rests rather heavily on the concept of (loss of) intelligibility. In historical linguistics, languages are often said to be formed through the process of "dialect split", which is defined as the process through which "[d]ialects, as they diverge more and more in the course of time, cease to be mutually intelligible and rank as separate languages" (Greenberg, 1971: 176. See also Hawkins, 2009; Jochnowitz, 2013; Kalyan & Francois, 2019). Similarly, the concept of intelligibility is relied upon in defining pidginisation (e.g. Trudgill, 1996), as well as successful attainment in second language learning where intelligibility levels, both measured and perceived, have been repeatedly shown to be of fundamental importance, to the extent

that "that intelligibility is a crucial concept in communication […] is not disputed" (Rajadurai, 2007: 89. See also Iwashita et al, 2008; Sewell, 2010).

However, the idea of intelligibility as a criterion of demarcation is not without problems. Firstly, despite the optimistic views quoted above from Dixon's (1997) and Salminen's (2007) work, linguistics still lacks an empirically grounded proposal for the implementation of the intelligibility criterion. Secondly, the idea that the intelligibility criterion can be implemented at all has been questioned, and negative conclusions have often been drawn. The next section will demonstrate that these conclusions may have been too hasty and possibly due to a conceptual misunderstanding. The remainder of the paper is then dedicated to a set of empirical studies which show evidence that that the intelligibility criterion can indeed function as an objective criterion of demarcation for an empirically grounded taxonomy of languages and dialects.

2. The intelligibility criterion[1]: a workable solution?

When considering intelligibility as a criterion of demarcation, scholars have often raised two main objections[2]. The first, which we may call the "political objection", is exemplified in the following quote by Chambers and Trudgill (1998: 3-4. See also Comrie, 2009; Janson, 2011; Lepschy, 2002; among many others. A similar stance is subsequently taken in Dunbar, 2001):

> "if we consider, first, the Scandinavian languages, we observe that Norwegian, Swedish and Danish are usually considered to be different languages. Unfortunately for our [intelligibility] definition, though, they are mutually intelligible."

This purported objection is so taken for granted that it is invariably repeated and conceded in linguistics textbooks[3] (e.g. Fromkin, Rodman, & Hyams, 2013) as well as in any of the relatively few reviews that discuss the dialect/language distinction (e.g. Pereltsvaig, 2017; Siegel, 2010; Stavans & Hoffmann, 2015; Wei, 2000; Woll, Sutton-Spence, & Elton, 2001). However, as pointed out in Author (2014), the objection is misguided, as it requires that we collapse the two concepts of *Abstand* language and *Ausbau* language into a single, generic and unidimensional concept. As soon as we follow Kloss' (1967) insight in considering

[1] For reasons of space, the concept of intelligibility will be considered on its own, and without addressing issues of "mutuality". However, see Hammarström (2008) and Schuppert (2011) for a rebuttal of the objections typically raised against the mutual component of the intelligibility criterion.
[2] There are in fact three typical objections, the third one being that of "variety chains". I will not discuss this here, however, as it has been exhaustively addressed by others (Hammarström 2008; Author, 2014).
[3] Unless the question is carefully avoided altogether, e.g. Yule 2014.

Abstand languages and *Ausbau* languages as separate entities identifiable by separate sets of criteria and for different purposes, then the apparent contradiction melts away. This is because the purported objection tacitly demands that there be an absolute correspondence between two distinct sets of entities, namely structural linguistic systems (*Abstand* languages), and socio-political constructions (*Ausbau* languages). Such demand for correspondence is fallacious. It is analogous to demanding that we reject the political scientists' definition of "republic" on the basis that it forces us to classify the Democratic People's Republic of Korea as a non-republic, a result that is in clear conflict with the country's official name as well as the belief of a number of its inhabitants. A cursive look at the political science literature is enough to show that such demand would be absurd. Political scientists have no qualms about stating that the Democratic People's Republic of Korea is a "dictatorship" (e.g. Jeong & Kim, 2016: 21), and "neither democratic, for the people, nor a republic" (Tan, 2016: 162), regardless of the country's official name or the government's insistence to the contrary. This is of course positive, as it is neither necessary nor indeed desirable to require that taxonomic categorisations resulting from objective, replicable measurements correspond to official government positions or to socially shared beliefs (see Ammon, 1989 for a similar point with regard to linguistics in particular). The "political objection" is therefore invalid as it rests on the conflation of two ontologically distinct concepts. Accordingly, if it turns out that the intelligibility criterion can be implemented, there will be no contradiction in stating that varieties X and Y are dialects of one *Abstand* language, even though they may have reached high levels of social construction such that they are commonly *perceived* to be or *officially acclaimed* as different languages, and may thus be classed as two *Ausbau* languages in sociologically oriented analyses.

The second common objection is based on the concept of "degree". As Hudson (1996) put it:

> "[…] intelligibility is a matter of *degree,* ranging from total intelligibility down to total unintelligibility. How high up this scale do two varieties need to be in order to count as members of the same language? This is clearly a question which is best avoided, rather than answered, since any answer must be arbitrary."
>
> (1996: 35, emphasis original).

The position exemplified in the quote above is widespread even today (e.g. Kauffeld, 2016; Kurpaska, 2019; Pereltsvaig, 2017), and it is essentially based on the idea that a linear scale does not involve any objectively identifiable threshold(s) and can therefore only be divided arbitrarily, supposedly leading to the conclusion that any attempts at divisions are therefore

futile. Leaving aside philosophical questions as to whether dividing scales is an *a priori* futile enterprise[4], the position exemplified in Hudson's quote is far from being the foregone conclusion it is often claimed to be, chiefly because it does not take into account important developments in intelligibility studies since the work of Smith (1982) and Munro and Derwing (1995a) (for more recent developments, see Hilton, Gooskens, & Schüppert, 2013; Kachru, 2008. See Sewell, 2010 for an overview). Specifically, for the "degree" problem to be considered fatal, one needs to rely on a unidimensional view of intelligibility both as a generic term for "understanding" and as a simple linear scale that runs from 0 (totally unintelligible) to 100 (totally intelligible) with no empirically identifiable thresholds. However, if we follow intelligibility researchers (e.g. Bamgbose, 1998; Smith, 1982; Smith and Nelson, 1985; Jenkins, 2000; among others) in breaking the process down into *comprehensibility* (recognising an utterance) and *intelligibility* (successfully retrieving the propositional content encoded in the utterance), it can no longer be maintained *a priori* that all and any ranges across the intelligibility scale will be equal, and thus that any partitioning of the scale will inevitably be arbitrary. It is at least possible in principle that, below a certain intelligibility level, hearers fail to decode messages with any reliability, perhaps even with no more reliability than if intelligibility were at 0%. If such cases exist, then we would be faced with instances in which it would make no taxonomical sense to classify the speaker's variety as belonging to "the same language" as the hearer's, since speaker and hearer fail to achieve communication through linguistic means. In other words, the linguistic code that the speaker utilises when building his/her utterances is unknown to the hearer to such an extent that the hearer is either (i) unable to decode those utterances or (ii) ends up with an output that does not match the intended message (see also Malmberg, 2012, on this point). Both scenarios (i) and (ii) lead to failure in retrieving the intended message from the phonetic stimuli produced by the speaker. In these cases, the speaker and hearer must necessarily be considered users of separate linguistic systems (i.e. separate *Abstand* languages)[5].

Further, it is also possible in principle that, below a certain intelligibility level, hearers *feel* that the variety being spoken to them is too different from their own variety for successful communication to be considered possible or achievable. This would be where the hearer perceives the process of decoding the speaker's variety as excessively arduous and the speaker's variety as potentially beyond comprehension. While this measurement relies on more "subjective" metrics (e.g. Saunders & Cienkowski, 2002), it would also give us some

[4] But see examples of how it has helped researchers in education (Le, Loll, & Pinkwart, 2013), agriculture (Peterson, Wysocki, & Harsh, 2001), psychiatry (Linscott & Van Os, 2010) to cite but a few.

[5] Here I am referring to the varieties being measured. It is of course possible that speaker and hearer share some other language in which they can communicate successfully, as in the case of multilingualism.

indication of a level beyond which it would be at least dubious to classify the speaker's variety as belonging to "the same language" as the hearer's.

Therefore, taking a multidimensional view of intelligibility allows us to ask the following questions:

1. do speakers feel unable to retrieve the propositional content of utterances if the intelligibility level falls below a certain point on the intelligibility scale?
2. is there a point along the intelligibility scale (0%-100%) beyond which speech becomes so poorly intelligible that it can no longer be said to "form part of a message"? (Sewell, 2010: 258).

A positive answer to the first question would give us evidence of a non-arbitrary threshold of minimal comprehensibility along the intelligibility scale. While intelligibility itself would remain measurable on a linear scale, the interaction between intelligibility and the comprehensibility levels that derive from it would indicate a possible intelligibility threshold, casting doubt on the idea that the concept of intelligibility is linear in nature, i.e. the idea that any point on the intelligibility scale is equal to any other point, a property which is necessarily true, for example, of mathematical scales[6]. Similarly, a positive answer to the second question would give us evidence that, while intelligibility levels are measurable on a linear scale, intelligibility is not itself a linear concept, as not all points along the intelligibility scale would qualify as equal. In either case, a positive answer would provide evidence against the widely held assumption that intelligibility is simply a "matter of degree" with "no clear-cut" segmentation (Comrie, 2009: 3). Conversely, if the answer to the second question turns out to be negative, i.e. we find that intelligibility simply decreases in a steadily incremental manner, we would have empirical evidence that intelligibility is potentially a linear measure without any discernible segmentation and is therefore likely unusable as a criterion of demarcation between languages and dialects, as previously assumed. Likewise, if it turns out that comprehensibility decreases linearly at a comparable rate as intelligibility levels decrease, then the conclusions drawn from the "degree" argument would have empirical confirmation that intelligibility is indeed a linear concept with no identifiable comprehensibility thresholds.

The present series of experiments investigated these issues by adapting and extending three paradigms, originally devised by Kalikow et al. (1977), Munro and Derwing (1995a, 1995b) and Anderson-Hsieh and Koehler (1998) for intelligibility scores, comprehensibility scores and listening comprehension respectively.

[6] Equating the intelligibility scale to mathematical scales is presumably at the origin of the degree objection, though no explicit reference has ever been made to this as far as I am aware.

3. Experiment I

The first experiment addresses research question 1, namely whether speakers feel unable to retrieve the propositional content of utterances once intelligibility levels fall below a certain point on the intelligibility scale. The view that intelligibility is "just a scale" without any empirically identifiable thresholds predicts that comprehensibility will simply decrease at the same rate as intelligibility, with no degree of intelligibility having any more or less of an impact on comprehensibility than any other degree, as dictated by the concept of scale. Experiment I was designed to test this prediction by testing for a potential interaction effect between comprehensibility and intelligibility, thus also investigating the possibility of a non-arbitrary threshold of minimal comprehensibility along the intelligibility scale. In order to achieve this, two sets of scores were obtained in this experiment: a functional intelligibility score and a comprehensibility score (sometimes also called intelligibility "judgement" or "opinion" score, e.g. Tang & van Heuven, 2009), which were then tested for interaction via an analysis of variance.

3.1 Method

Intelligibility scores were obtained as percentages of correct responses to sentence stimuli. The sentences were selected from the 'high predictability' list originally developed for English by Kalikow et al. (1977) and more recently shown to be an accurate measure of intelligibility across related varieties (Author, 2014; Tang and van Heuven, 2009; Wang, 2007). Comprehensibility scores were obtained following conventional methodology in comprehensibility studies (Derwing & Munro, 2009; Isaacs & Thomson, 2013; Sheppard et al., 2017; as originally developed by Munro and Derwing 1995a, 1995b), whereby stimulus sentences are scored on a 9-point Likert scale. Each participant was asked to assign a value by circling the number they felt was most reflective of the effort involved in retrieving the message, with 1 indicating "very easy to understand" and 9 indicating "impossible to understand".

3.2 Participants

Forty-two British undergraduates (11 M – 31 F) between the ages of 18 and 23 took part in the experiment in partial fulfilment of a course requirement. All participants were studying at a UK university, and they were screened for linguistic background to ensure that only monolingual English speakers with little or no knowledge of a second language were included.

272
273 **3.3 Materials**
274 *3.3.1 Stimuli*
275 Both intelligibility scores and comprehensibility scores were obtained as responses to
276 auditory stimuli. To produce the auditory stimuli, 34 sentences were randomly selected from
277 the high-predictability sentence lists of the intelligibility test developed by Kalikow et al.
278 (1977). These include declarative SVO sentences, imperatives, and passives, all with a
279 prepositional adjunct, as in the following examples:
280
281 1a. Declarative He caught the fish in his net
282 1b. Imperative Keep your broken arm in a sling
283 1c. Passive Her hair was tied with a blue bow
284
285 In this test, listeners are requested to write down the final word (i.e. the target) of each
286 sentence they hear. These sentences are classed as high predictability because they
287 provide contextual information leading up to the target word, thus linking target recognition to
288 the overall understanding of the sentence (the underlined word is the target):
289
290 2. He caught the fish in his net
291
292 All sentences had a total length of between six and eight words and between 17 and 22
293 phones.
294 The initial 34 sentences were recorded in a soundproof booth by a female speaker of
295 Standard British English with a mild Northern English accent. The speaker was a trained
296 linguist and was instructed to keep pace and intonation constant throughout the recordings.
297 The 34 sentences were then manipulated electronically to produce four sets of stimuli (A, B,
298 C, and D), each containing the same 34 sentences but with varying levels of phonetic
299 distance and thus decreased intelligibility (for the link between phonetic distance and
300 intelligibility see for example Gooskens, 2007). For set A, each of the 34 sentences had two
301 phones replaced. For instance, for the sentence in the example in (2) above, the first phone
302 in “fish” (i.e. the fricative [f]) was replaced with [v], while the third phone in “net” (i.e. the
303 plosive [t]) was replaced with [θ]. The segmental positions to be manipulated were selected
304 randomly, while the replacement sound was chosen based on plausible but unattested
305 historical changes, namely changes that could have happened in the development of some
306 English dialect (as indicated by attested Indo-European processes reported for example in
307 Ringe, 2017; Ringe & Taylor, 2014) but that are actually unattested in any currently living
308 dialect of English. For instance, in the example above, the change in manner of articulation

from the plosive [t] to the fricative [θ] reflects a plausible though unattested case of word-final lenition. All changes involved one feature dimension, namely either place, manner, or voicing for consonants and either height, backness or roundness for vowels. In keeping with research arguing that "phonetic sensitivity" (Nerbonne & Heeringa, 2010:553) needs to be incorporated into considerations of phonetic distance by keeping consonants and vowels distinct in order to achieve a "linguistically responsible" process of phone substitution (Nerbonne, Colen, Gooskens, Kleiweg, and Leinonen, 2011: 73), sound replacements always involved substituting vowels for vowels and consonants for consonants. Note that the same premise also follows from the concept of "plausible but unattested historical changes" described above, as historical changes tend to affect consonants and vowels differently.

For the remaining sets of stimuli, each had one more phone manipulated per sentence in the manner describe above, so that each sentence in set B had a total of three phones replaced (the same two phones as in set A plus an additional one), while sentences in set C had a total of four phones replaced and those in set D a total of five. This corresponded approximately to 10% of the total phones being manipulated for the sentences in set A, 15% for set B, 20% for set C and 25% for set D.

Four sets of stimuli were therefore produced, each of which exemplified a possible but non-existent dialect of English with varying degrees of phonetic distance. This ensured that all participants were being tested on a linguistic variety to which they had no previous exposure, following a similar logic to non-word tasks, which involve possible but non-existent words in order to avoid the confound of previous exposure (e.g. Gathercole et al., 1994). By analogy with the term "non-word" (which refers to a possible but non-existent word) we might call this possible but non-existent dialect a "non-dialect". For the comprehensibility scores, using "non-dialect" stimuli may also minimise potential attitudinal effects whereby participants might otherwise provide overly low or overly high ratings to the auditory stimuli due to the social preconceptions associated with a specific, familiar dialect (e.g. Smith & Bailey, 1980).

3.3.2 Design

Four separate lists were generated using a Latin square design. Each list contained 32 test sentences preceded by two practice sentences (T = 34), and with a 5.0 second pause in between each sentence. The 32 test sentences comprised of eight test sentences for each condition (A, B, C, D), with each condition varying in phonetic distance as described above (i.e. A=10%, B=15%, C=20%, D=25%). Each participant was randomly assigned to one of the four lists.

3.3.4 Procedure

Participants heard the stimuli through high-fidelity Sennheiser-HD201 headphones. They were instructed to provide two responses immediately after hearing each sentence. First, to write down what they thought the final word of the sentence was (i.e. the target), and then assign a perceived comprehensibility rating to the sentence. They were specifically asked to assign this rating with the whole sentence in mind, not just the final word. Participants took part in the experiment in individual sessions, in a quiet room.

3.4 Results

Inter-rater agreement for comprehensibility judgements was measured using Intra-Class Correlation Coefficient. The average measure ICC was .827 with a 95% confidence interval from .642 to .908 ($F(23,69) = 5.208$, $p < .001$), showing consistency across participants' rating of the stimuli. Mean comprehensibility ratings were therefore computed for each condition.

To test the hypothesis that comprehensibility ratings decrease at the same rate as intelligibility scores, a two-way repeated measure multivariate analysis of variance (MANOVA) was conducted. This enabled the evaluation of changes across measurement types (i.e. comprehensibility ratings vs intelligibility scores) over increases in phonetic distance (i.e. across 10%, 15%, 20% and 25% phonetic distance). The results show a statistically significant interaction between phonetic distance and test type (with Huynh-Feldt correction, $F(2.6, 106.72) = 15.51$, $p < .001$, $\eta_p^2 = .274$), revealing that the rate of decrease differs across measurement types, with comprehensibility ratings decreasing less rapidly than intelligibility scores.

Post-hoc paired-samples t-tests were performed on the log-transformed data to compare intelligibility scores and comprehensibility ratings at each level of phonetic distance. Results revealed a statistically significant difference only at 25% phonetic distance ($t(41) = -6.796$, $p < .001$), while no significant difference emerged at 10% and 15% phonetic distance ($p > .409$). Significance was approached at 20% phonetic distance ($p = 0.056$). Furthermore, the effect size for 25% phonetic distance ($d = 1.048$) exceeds Cohen's (1988) convention for a large effect.

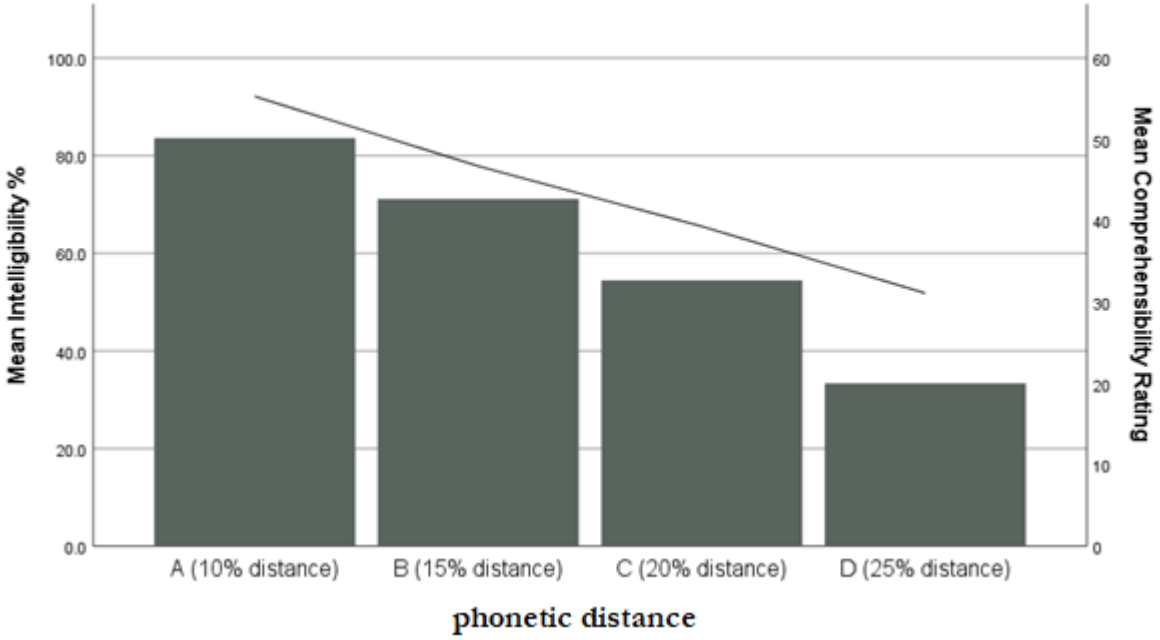

Figure 1: comparison between intelligibility scores and mean comprehensibility ratings across the four conditions.

In keeping with the literature on the relationship between intelligibility and phonetic distance (Gooskens, 2007; Gooskens, Heeringa, & Beijering, 2008; Speelman, Impe, & Geeraerts, 2014; Yang, 2012; among others), results also showed a main effect for phonetic distance (with Huynh-Feldt correction, $F(2.37 , 97.14) = 253.57$, $p < .001$, $\eta_p^2 = .861$), confirming that increasing phonetic distance predictably decreased intelligibility scores.

3.5 Discussion

Experiment I aimed to address the following research question: "do speakers feel unable to retrieve the propositional content of utterances if the intelligibility level falls below a certain point on the intelligibility scale?"

In order to address this question, the experiment tested whether comprehensibility ratings and intelligibility scores decrease at the same rate, as predicted by the view that intelligibility is "just a scale" with no identifiable thresholds. Results showed that the two variables decrease at statistically significantly different rates, providing empirical evidence against the widely held view that intelligibility is a linear measure without any discernible segmentation. However, the manner in which the two measures differ is somewhat surprising. As suggested by the research question, in case of a different rate of decrease between the two measures, a potential outcome could have been the decrease of

comprehensibility over and above what can be accounted for by decreased intelligibility, thereby suggesting that listeners feel unable to retrieve the propositional content of utterances once the intelligibility level falls below a certain point on the intelligibility scale. However, results showed that the difference between the two measures is due to comprehensibility decreasing less rapidly, not more rapidly, than intelligibility. This suggests that far from feeling unable to retrieve the propositional content of utterances, listeners actually become unable to reliably estimate how little they do understand, thus rating sentences as relatively easily comprehensible while at the same time failing to successfully retrieve their propositional content. While these results reveal a state of affairs that diverges from what research question 1 suggested, they nevertheless provide counterevidence to the assumption that all points on the intelligibility scale are equal, while also providing some evidence of a comprehensibility threshold along the intelligibility scale. Specifically, the results show that listeners' reliability decreases more rapidly from the 20% decay mark, where intelligibility falls below 70%. Interestingly, this matches several suggestions from various disciplines where the figures of 70% and 75% have often been proposed as potential thresholds of minimally acceptable intelligibility (e.g. Aniansson & Peterson, 1983; Casad, 1974; Moore, 1989; Wang et al., 2012).

Moreover, the magnitude of this overestimation increases as intelligibility decreases, suggesting that the less intelligible an utterance becomes, the more listeners become unable to reliably judge its degree of comprehensibility. We can therefore conclude that while we have not found a threshold of minimal comprehensibility, we have nevertheless identified a potential threshold along the intelligibility scale, albeit in the form of reliable comprehensibility ratings.

Note that, while intelligibility was manipulated by increasing phonetic distance, this is not to claim that intelligibility may only be affected by phonetics. It is indeed the case that intelligibility is affected by lexical, syntactic and/or morphological differences (e.g. Gooskens, Heeringa, & Beijering, 2008). However, recall that the aim of this study was to test the prediction that comprehensibility will decrease at the same rate as intelligibility. To this end, the *reasons* why intelligibility may have decreased is tangential to the aims of the study. The finding from Experiment I, namely that intelligibility and comprehensibility do not decrease at the same rate, constitutes evidence against the claim that intelligibility is "just a scale" with no empirically identifiable thresholds. Such evidence stands regardless of how intelligibility happened to decrease, as the core point here is that its rate of decrease differed from that of comprehensibility.

3.6 Testing the concept of "intelligibility threshold"

The next set of experiments addresses research question 2, namely whether there is an identifiable point along the intelligibility scale beyond which speech becomes so poorly intelligible that it can no longer be said to "form part of a message" (Sewell, 2010: 258), suggesting that the hearer's decoding system and the speaker's encoding system are too dissimilar to be considered part of the same *Abstand* languages. Specifically, Experiment II investigates the possibility that – below a certain intelligibility level – hearers may fail to be able to decode messages beyond chance levels, which is the same level at which we would expect speakers of two unintelligible languages to perform.

Experiment II approaches this issue from the perspective of single sentences, investigating at what point the hearer can no longer reliably decode the message encoded in a sentence (in the absence of non-linguistic cues), while Experiment III approaches the issue from the perspective of a longer communicative piece, where issues of short-term-memory and broader contextual information are also at play.

4. Experiment II

4.1 Method

A forced-choice procedure was used for this experiment, where participants were asked to judge whether a spoken sentence matched an accompanying picture. The sentences were selected from the same list as in Experiment I above (i.e. from Kalikow et al. 1977), following the same procedure detailed in Experiment I.

Each participant could only judge each sentence-picture pair as either a match or a mismatch, and the participant's score consisted of the total number of correctly identified matches (see below for details). In keeping with the force-choice paradigm, participants were not allowed the option of skipping an item.

4.2 Participants

Sixty-one adult speakers of British English (25 M – 36 F) between the ages of 18 and 40 were included in the experiment. A further three participants were tested but excluded from the analysis due to being fluent bilinguals. Participants were recruited through social media, and they were screened for linguistic background to ensure that only monolingual English speakers with little or no knowledge of a second language were included.

4.3 Materials

Materials included auditory and visual stimuli. The auditory stimuli were produced following the same procedure detailed in Experiment I. A total of 32 sentences were recorded by the same female speaker of Standard British English who produced the auditory stimuli for Experiment I. The 32 sentences were then manipulated electronically to produce four sets of stimuli (A, B, C, and D) with varying levels of phonetic distance and thus decreased intelligibility, in the same manner detailed in Experiment I.

The visual stimuli consisted of 32 pictures. Half of these pictures (N=16) matched one of the 16 stimulus sentences and formed the test items (i.e. matching sentence-picture pairs). These test items all consisted of matching sentence-picture pairs due to the fact that the aim of the experiment was to investigate how reliably participants could retrieve the message from the stimulus sentences, a result that only successful identification of a matched sentence-picture pair could indicate. The participant's score consisted of the total number of correctly identified matches. Each correctly identified match was assigned a score of 1, allowing for a maximum score of 4 per condition (16 test sentences / 4 conditions). Incorrect responses were scored as 0.

The remaining 16 pictures did not match any of the stimulus sentences. These formed the foil items (i.e. mismatching sentence-picture pairs) which introduced mismatches into the task to avoid response set effects.

The pictures were all in colour and were made using open source clipart and a picture editing software.

Figure 2: example of visual stimulus (match stimulus for the sentence "He caught the fish in his net").

Figure 3: example of visual stimulus (mismatch stimulus for the sentence “They marched to the beat of the drum”).

4.3.1 Design

Four separate lists were generated using a Latin square design. Each list contained 16 test items (matching picture-sentence pairs) and 16 foil items (mismatching picture-sentence pairs) preceded by two practice items (T = 34). The 16 test items comprised of four test sentences for each condition (A, B, C, D), with each condition varying in phonetic distance as described above (i.e. A=10%, B=15%, C=20%, D=25%). Each participant was randomly assigned to one of the four lists.

4.3.2 Procedure

Participants accessed the experiment through online software (www.gorilla.sc. See Anwyl-Irvine et al, 2019) via a personal computer or laptop. Upon accepting to take part in the experiment, each participant was instructed to connect a set of headphones before agreeing to move on to the next screen. The next screen described the task to participants, as follows:

> Picture Matching Task
> In this task you will hear some sentences accompanied by a picture. Each sentence contains sounds that have been manipulated in order to reduce intelligibility. For each sentence-picture pair, your task is to click the smiley face ☺ if you think that the sentence matches the picture, and the frowney face ☹ if you think that the sentence does not match the picture.
> You might find that some sentences are unintelligible, in which case you can take a guess. You will only be allowed to listen to each sentence once.
> The task will take approximately five minutes to complete.

Participants would then need to click a button labelled “I’m ready” to begin the task

4.4 Results

A Friedman test revealed a statistically significant difference between the four conditions ($\chi^2(3) = 84.132$, $p < 0.001$.). Follow-up Wilcoxon’s signed ranks tests with Bonferroni correction (p =< .016) determined that there was a statistically significant difference between performance in conditions B and C ($Z = -1.807$, $p = .004$) and in conditions C and D (p < .001), but no significant difference between performance in conditions A and B (p = .059).

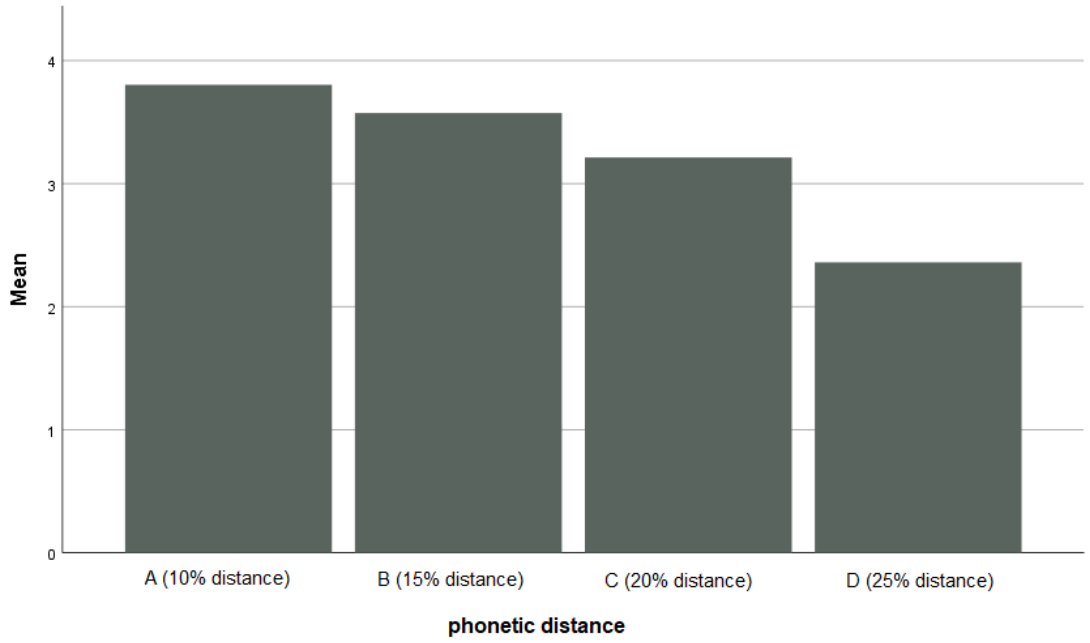

Figure 4: comparison of forced-choice scores across the four conditions.

A binomial test indicated that the proportion of correctly identified matches in Condition D (.49) did not differ significantly from chance (.50), p > .999 (2-sided).

4.5 Discussion

Experiment II investigated the possibility that - below a certain intelligibility level - hearers may consistently fail to decode messages beyond chance levels, thus casting doubt on the widespread assumption that intelligibility is simply a linear scale with no useful or even interesting thresholds across it. The experiment addressed the question of whether there is a

point along the intelligibility scale beyond which speech becomes so poorly intelligible that it can no longer be said to form part of a message, and specifically whether participants would be able to correctly identify sentence-picture matches where the sentences had been manipulated to gradually increase phonetic distance and thus decrease intelligibility[7]. Results show that once phonetic distance reaches 25%, participants can no longer retrieve the intended message beyond chance level. In other words, reducing phonetic distance by 25% does not simply have a negative effect on listeners' ability to retrieve information. The 25% threshold leads to a performance that is no different from a situation where phonetic distance between speakers' variety and listeners' variety is 100%, as listeners would also be able to perform at chance level in a task where their variety is maximally different from the speakers'. These results cast serious doubt on the widespread *a priori* assumptions that (i) intelligibility cannot involve any objectively identifiable thresholds (e.g. Kauffeld, 2016; Pereltsvaig, 2017), (ii) that – due to the fact that intelligibility is a linear scale - nothing may be gained by investigating its partitioning (e.g. Hudson 1996), and (iii) that any partitioning of such scale can only be done arbitrarily (e.g. Wei, 2000).

Firstly, we now have empirical evidence that, although *measurable* on a linear scale, intelligibility does not behave in the manner expected of a liner notion where each point on its scale is equivalent to any other point. Experiment II has shown that reducing intelligibility by the equivalent of 15% phonetic distance does not significantly impair listeners' ability to decode a message more than when phonetic distance is at 10%. Once the distance goes beyond 15%, however, intelligibility becomes drastically impaired, and when distance reaches 25% listeners' rates of linguistic decoding drop to chance level. We can therefore conclude that a phonetic distance of between 16% and 25% (or, inversely, phonetic equivalence between 84% and 75%) is the most likely candidate for a threshold of minimal intelligibility. Following the results of Experiment I, this stands between 34% and 71% intelligibility on the sentence-level intelligibility test (see Experiment I for details).

Secondly, Experiment II has shown that, far from being a futile endeavour, investigating the properties of different ranges across the intelligibility scale revealed a range within which proposition retrieval consistently fails, at last at sentence level. While this range remains arguably wide and further research is needed in order to establish a more fine-grained level between the 34% and 71% currently identified, we have nevertheless made progress in addressing the question of "how much distance is enough" before we must necessarily consider two varieties as separate *Abstand* languages. Furthermore, we now have evidence that when the intelligibility level between two varieties is =<34% it becomes linguistically unsound to suggest that the speaker's and the hearer's varieties belong to the

"same language", as the degree of *Abstand* between the two varieties is such that sentences uttered in the speaker's variety cannot be successfully decoded by relying on the hearer's variety any more than if the hearer relied on a maximally distant *Abstand* language where phonetic overlap is at 0%.

Thirdly, result from Experiment II provide further evidence to refute the assumption that any partitioning of the intelligibility scale must necessarily be arbitrary. A number of studies have already shown that certain intelligibility levels are more desirable than others in ways that are not only empirically identifiable but that also have predictable consequences in applied domains (see for example Garinther, Whitaker, & Peters, 1995 on intelligibility in military performance; Gordon-Brannan & Hodson, 2000 on intelligibility as a diagnostics in speech and language pathology; Yang & Hodgson, 2006 on intelligibility thresholds in sound-system engineering). What our results have contributed is evidence that intelligibility levels at 34% are insufficient for the successful retrieval of a sentential proposition, and that the minimum level of intelligibility required lies between the rates of 34% and 71%. Insofar as one believes that for the statement "John and Mary speak the same language" to be true it is necessary that John be consistently able to retrieve the propositional content of the sentences spoken by Mary (and vice versa), then we can also conclude that the same range applies to the identification of *Abstand* languages in linguistic continua.

Besides improving our understanding of our discipline's object of inquiry, this finding may also be of value to the applied linguist. Indeed, as Leonardi (2016) pointed out, there are several social and educational pitfalls directly linked to the pervasive insistence on favouring *Ausbau* considerations when classifying varieties that are separated by considerable *Abstand,* and ignoring *Abstand* considerations leads to pernicious assumptions about speakers' "mother tongue" as well as to unnecessarily protracted stages of semilingualism. Following Experiment II, we are now a step closer to defining this hitherto elusive concept of "considerable *Abstand*".

However, while Experiment II gives us an indication of the intelligibility range within which proposition retrieval consistently fails at the sentential level, it also raises the question of how operating at this range impairs one's ability to function socially in a community where a related but different variety is the established *Ausbau* language. This particular set-up is virtually impossible to test in established linguistic communities because speakers of languages that are related to but different from the established *Ausbau* language tend to have had considerable amounts of exposure to the *Ausbau* language in question, virtually by definition. This is where the "non-dialect" paradigm presented in this paper can provide a useful testing ground, as described in the next experiment.

5. Experiment III

Experiment III also addressed the question of whether there is a point along the intelligibility scale beyond which speech becomes so poorly intelligible that it can no longer be said to form part of a message. However, while Experiment II did so from the perspective of single sentences, (i.e. at what point is one no longer reliably receiving the message encoded in a sentential proposition, in the absence of non-linguistic cues?), Experiment III approaches the question from the perspective of a longer communicative piece. More specifically, the question that Experiment III aims to address is the following: how far apart on the intelligibility scale do two varieties need to be in order for speakers of one to be unable to function as communicatively competent in the other? And, by extension, in order for speakers of one variety to be unable to function as communicatively competent members of a speech community where the other variety is the established *Ausbau* language?

5.1 Method

While communicative competence in everyday exchanges involves a number of contextual as well as linguistic cues (Duran, & Kelly, 1985; Knutson & Posirisuk, 2006), a speaker's / listener's communicative competence has been shown to be reliably measured via language tests. For example, the Test of English as a Foreign Language (TOEFL) has been shown to be a highly reliable indicator of learners' actual communicative competence in ordinary conversation (Bridgeman et al., 2012) having been developed with the specific aim of communicative competence in mind (Carrell, 2007; Taylor & Angelis, 2008).

A modified version of a TOEFL listening comprehension task was therefore used for this experiment. The task was modified in accordance with the "non-dialect" paradigm used in experiments I and II, as detailed in the Materials section below.

TOEFL listening tasks have been selected as a particularly fitting method to address the research question above due to the fact that TOEFL scores have been shown to correspond closely to English language skills required in order to successfully function in higher education (e.g. Powers, 1985; Rosenfeld, Leung, & Oltman, 2001; Sawaki & Nissan, 2009) as well as in professional roles (Farnsworth, 2013; Wagner, 2016). TOEFL scores have also been shown to accurately measure cross-dialectal intelligibility (Kang, Moran & Thomson, 2018) and have been employed widely in the measurement of linguistic variation, particularly phonetic variation (e.g. Kang, Moran & Thomson, 2018; Major et al., 2002; Ockey, Papageorgiou, & French, 2016).

5.2 Participants

A total of 122 British undergraduates (35 M – 87 F) between the ages of 18 and 23 took part in the experiment in partial fulfilment of a course requirement. All participants were studying at a UK university, and they were screened for linguistic background to ensure that only monolingual English speakers with little or no knowledge of a second language were included.

5.3 Materials

5.3.1 Stimuli

Following the procedure employed in the listening section of the TOEFL, the stimuli comprised of a monologic lecture and a set of nine questions designed to test participants' understanding of the lecture content. In the TOEFL, questions are designed to test for both basic comprehension and pragmatic understanding, including the use of contextual information to draw inference from some of the speaker's statements.

The lecture transcript consisted of a discussion of bee behaviour, specifically on the characteristics and hypothesised purposes of the "waggle dance". The total length of the transcript was 719 words. This lecture transcript was first transcribed in IPA, totalling 2418 phones, and then manipulated to produce three lecture stimuli (A, B, C), each with different levels of phonetic distance from the original transcript. For each auditory stimulus, a percentage of the total phones were replaced: 7.5% for stimulus A (N= 189 out of 2418), 12% for stimulus B (N= 283 out of 2418) and 15% for stimulus C (N= 375 out of 2418). This ensured that the stimuli for condition C were comparable to the stimuli in one of the conditions in Experiment I, namely condition B, which also involved substituting 15% of the original phones. However, unlike Experiment I, Experiment III did not include conditions beyond 15% phonetic distance in order to avoid possible floor effects due to the additional complexities of the task at hand and the more significant challenges that longer, more complex clauses pose to working memory recall (e.g. Blauberg & Braine, 1974; Montgomery, 2000).

Following the steps outlined in Experiment I, the segmental position of each phone to be replaced was selected randomly across the transcript, with the exclusion only of a proper name which appeared five times in the text. The replacement sounds were chosen on the basis of the non-dialect procedure defined above, namely by applying plausible but unattested historical changes to each of the randomly selected phones, and by substituting vowels for vowels and consonants for consonants, involving one feature dimension per change (either place, manner, or voicing for consonants and either height, backness or roundness for vowels).

Each modified transcript was subsequently recorded in a soundproof booth by a female speaker of Standard British English with a mild Northern English accent. The

speaker, who had linguistic training and could read the IPA, was coached by two trained assistant researchers who ensured that she pronounced each component phrase and sentence with natural intonation and at a speed comparable to that of recordings used on the TOEFL test (as set out at https://www.ets.org/toefl), changing only the pronunciation of the relevant phones. Each recording totalled between 4min 56sec and 5min 15sec in length.

5.3.2 Procedure

Participants were tested in a classroom environment and allocated to one of three separate groups (one group per condition). The auditory stimuli were played through the classroom speaker system. Following standard practice in TOEFL testing, participants were allowed to take notes during the test, and were asked to wait quietly until everyone had finished before leaving the room.

Each participant was randomly assigned to one of the three groups. Scores were calculated on the nine question items as follows: questions one to seven were assigned 1 point for correct answers and 0 points for incorrect answers, while questions eight and nine were assigned 2 points for correct answers and 0 points for incorrect answers, for a possible total of eleven points per participant. Questions eight and nine were assigned more points in keeping with common practice in the TOEFL Listening test and in TOEFL preparation tests (e.g. https://www.test-guide.com), as correctly answering these questions requires that the listener go beyond basic understanding of the text, collating more than one item of information from the lecture content in order to apply some amount of pragmatic inference.

5.4 Results

A one-way ANOVA revealed a statistically significant difference between groups ($F(2,119) = 9.391$, $p < .001$). A Tukey post hoc test revealed that test scores were statistically significantly lower at 15% phonetic distance (M = 5.36, SD = 1.92) compared to 7.5% phonetic distance (M = 7.05, SD = 1.58, $p < .001$) and to 12% phonetic distance (M = 6.50, SD = 1.79, $p < .012$), but there was no statistically significant difference between groups at 7.5% and 12% phonetic distance ($p = .351$).

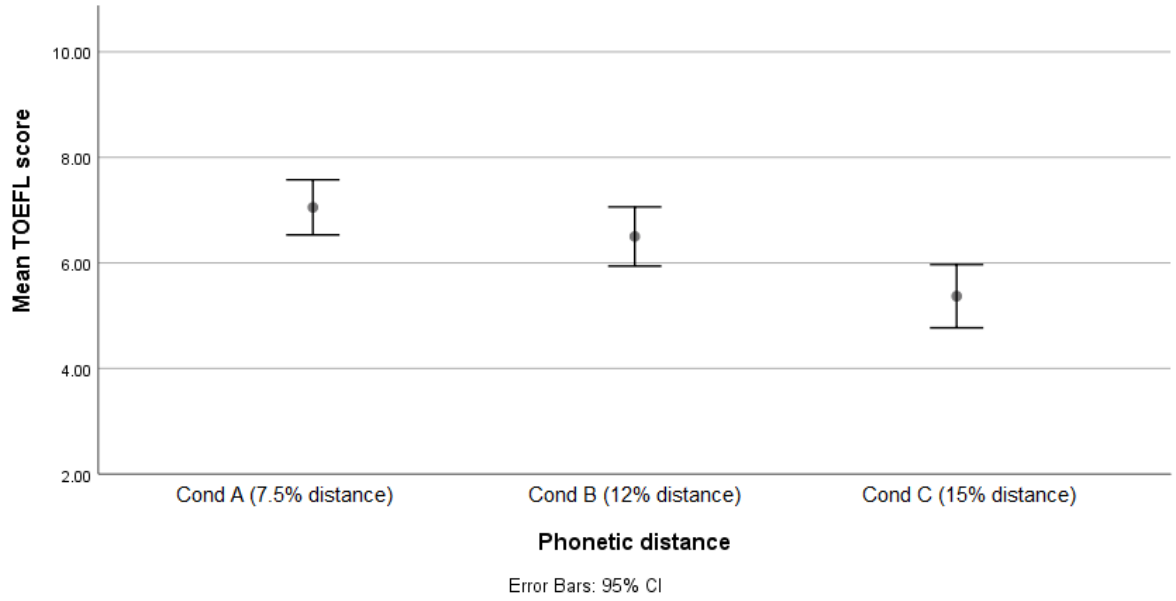

Fig. 5: participant scores on listening task by condition.

To enable comparison of the participants' results with average TOEFL scores, we converted mean scores into percentages and subsequently calculated scaled scores corresponding to the TOEFL listening section (following Hicks, 1989). Scaled scores constitute the scores on TOEFL score reports and are the scores on which TOEFL requirements are based (ETS, 1998). Corresponding percentile rank is also presented for comparison (ETS, 2017).

	Cond A	**Cond B**	**Cond C**
% correct answers	78.4%	72.2%	59.7%
TOEFL raw score equivalent[8]	27	25	20
Corresponding TOEFL scaled score	20	18	13
Corresponding percentile rank	45	36	18

[8] This is calculated by applying the percentages above to the maximum raw score obtainable on the TOEFL Listening section (i.e. 34), rounded to the nearest integer.

Table 1: participants' scores in percentages, conversions to scaled scores and corresponding percentile ranks.

As indicated in Table 1, despite the relatively low amount of phonetic distance (as compared to Experiments I and II), participants performed rather poorly on the listening tasks, with condition C (15% phonetic distance) leading to a result that would place participants within the 18th percentile.

5.5 Discussion

Similarly to Experiment II, Experiment III addressed the question of whether there is a point along the intelligibility scale beyond which speech becomes so poorly intelligible that it can no longer be said to form part of a message. However, while Experiment II focused on the sentence level, Experiment III was concerned with a longer communicative piece. In doing so, Experiment III asked a specific, arguably more fine-grained version of the question "how much *Abstand* is too much", namely how much *Abstand* is enough to prevent a listener/speaker from functioning as a communicatively competent member of a speech community. Results showed that, at 15% phonetic distance, participants scored the equivalent of a 13 score on the TOEFL Listening test. This is considerably lower than what is considered a "clearly adequate" (Moglen, 2015: 11) level of language skills necessary for university students (set at between 21 and 25), and lower than what is considered "less than adequate" (i.e. between 16 and 20, Moglen, 2015: 11), as well as being considerably lower than the minimum requirement for admission to undergraduate programmes, e.g. in Canada (Simner & Mitchell, 2007). It is also considerably lower than what is considered "just enough […] to perform the job of an entry-level nurse" (O'neill, Tannenbaum, & Tiffen, 2005: 137) or what is considered acceptable by Irish professional bodies, which require a minimum score of 22 (Merrifield, 2012). This strongly suggests that a 15% phonetic distance is more than enough to render a speaker/listener unable to function as a linguistically competent member of a speech community at an educated and/or professional level. In other words, a 15% phonetic distance may force members of a community into lower social and socioeconomic positions than what they would have otherwise been able to access had the phonetic distance not been as high, an effect potentially comparable to the negative impact that illiteracy has on job opportunities and socioeconomic status (e.g. Messias, 2003). This suggests that the language used in condition C of Experiment III cannot rationally be described as "the same language" as the participants' mother tongue.

In fact, even at 7.5% phonetic distance, participants could only achieve an equivalent score of 20, which – at best - is at the margins of acceptability for most universities and for

765 the professional bodies cited above. Among other things, this confirmed that the additional
766 complexities of the task at hand and the more significant challenges that longer, more
767 complex clauses pose to working memory recall (e.g. Blauberg & Braine, 1974; Gooskens,
768 2013; Montgomery, 2000) lead to performance being highly negatively affected even at
769 lower levels of phonetic distance. At sentence level (i.e. Experiment I) intelligibility becomes
770 seriously impaired from 20% distance onwards, while longer, more demanding structures
771 can lead to poor intelligibility at 7.5% phonetic distance.
772 In addition, these results provide further empirical evidence that, although
773 *measurable* on a linear scale, intelligibility does not necessarily have the characteristics of a
774 liner notion, as not all points on its scale are equal. Specifically, reducing intelligibility by the
775 equivalent of 15% phonetic distance impairs listeners' ability to decode longer, articulated
776 messages to such an extent that they would be unable to function as communicatively
777 competent members of a speech community whose language is 15% phonetically distant
778 from their own. Following the results of Experiment I, this is equivalent to 71% intelligibility
779 on the sentence-level intelligibility test (see Experiment I for details). Once again, this
780 matches suggestions from other disciplines where figures between 70% and 75%
781 intelligibility are often proposed as potential thresholds of minimal acceptability (e.g. Wang et
782 al., 2012).
783 Furthermore, and in keeping with the results of Experiment II, Experiment III has also
784 shown that investigating the properties of different ranges across the intelligibility scale is far
785 from a futile endeavour (contra e.g. Hudson, 1996). Specifically, the results of Experiment III
786 suggest that maintaining that two varieties at 15% phonetic distance are "the same
787 language" may lead to issues of social injustice in the form of impaired social mobility,
788 strongly suggesting that it is unwise to continue to perpetuate the habit of favouring
789 sociolinguistic notions when defining or identifying "languages" (i.e. the "*Ausbau*-centrism" of
790 Author, 2014). In fact, the results of Experiment III strongly suggest that favouring *Ausbau*
791 considerations over *Abstand* relations can unwittingly lead to "linguistic injustice" (see e.g.
792 Craft et al., 2020 on this notion), with speakers being systematically reported as or expected
793 to be "native" in some *Ausbau* language, when in fact their actual mother tongue is too
794 phonetically distant from the *Ausbau* language in question to be rationally considered "the
795 same language". The result is that these speakers are not communicatively functional in the
796 language that – due to our bias for sociolinguistic considerations over *Abstand*
797 characteristics – is their supposed mother tongue. For similar reasons, issues of injustice
798 also arise in relation to people who – besides speaking some highly *Ausbau*-ized language –
799 also speak some other variety classed as a "dialect" of that language on purely *Ausbau*
800 grounds, and are therefore routinely identified as being "monolingual" despite the fact that

they know and regularly use two *Abstand* languages, and have likely had to learn as an L2 the language they are supposedly monolingual in (see Leonardi, 2016, for an example).

To reiterate, insofar as one believes that for the statement "John and Mary speak the same language" to be true it is necessary that John be consistently able to retrieve the propositional content of the sentences spoken by Mary (and vice versa), the results of Experiment III show that a 15% phonetic distance is a good indicator that we are dealing with two *Abstand* languages, as 15% distance causes John to be unable to function in Mary's linguistic community (and vice versa). Failing to consider this indicator may lead to unwelcome consequences for speakers of related varieties that are taken to belong to the "same language" purely on *Ausbau* considerations.

6. Overall Discussion and Tentative Conclusions

Identifying the object of inquiry is an important step in any scientific discipline. In the case of Linguistics, a definition of this "object" has been rather elusive (e.g. Fasold, 2005), a fact that has led to the worryingly widespread assumption that languages cannot be defined linguistically (e.g. Chambers and Trudgill, 1996) with some authors even welcoming the discipline's failure to provide a definition as a positive result (e.g. Otheguy, García, & Reid, 2015). Nevertheless, many linguistic subfields continue to depend on or even tacitly assume some form of definition of "language" as a structural, linguistic object in opposition to that of (its) "dialects"; language enumeration, multilingualism research, historical linguistics, to name but a few. This continues to beg the question of what criterion of demarcation could provide a potential solution to the taxonomical problem or "language" and "dialects".

In this paper I suggested that the intelligibility criterion is most probably our best candidate. I argued that one of the typical objections raised against intelligibility (i.e. the "political" objection) is based on a fallacy and should therefore be abandoned. I then pointed out that a second objection typically raised against the workability of an intelligibility criterion (i.e. the "degree" objection) amounts to an empirical claim and that – as such – it is therefore testable. Specifically, the degree objection states that because intelligibility can be measured on a scale from 0% to 100%, it automatically follows that no objective threshold can be identified, presumably because all points on the intelligibility scale must inherently be equal (a property which is necessarily true of mathematical scales). However, this logical leap is hardly warranted, given the successes in other disciplines where not only has it been shown that linear scales can be partitioned in objective and meaningful ways, but also that such partitioning can lead to a better understanding of a range of phenomena, e.g. in education (Le, Loll, & Pinkwart, 2013), agriculture (Peterson, Wysocki, & Harsh, 2001), and psychiatry (Linscott & Van Os, 2010). Even linguistics has had some successes in partitioning scales,

838 leading to a better understanding of phonological perception, specifically perception of voice
839 onset time and how phonological representations partition an acoustic continuum into
840 discrete categories (Casserly & Pisoni, 2010). In addition, and most importantly for our
841 purposes, such claim has typically been maintained as an *a priori* truism without any
842 empirical testing to support it. The core aim of this paper was to analyse this claim in more
843 details and then proceed to test it through a series of empirical studies.
844 The studies presented here addressed two separate yet interconnected questions,
845 namely (1) whether speakers feel unable to retrieve the propositional content of utterances if
846 the intelligibility level falls below a certain point on the intelligibility scale and (2) whether
847 there is an identifiable point along the intelligibility scale beyond which speech becomes so
848 poorly intelligible that listeners can no longer rely on the linguistic knowledge of their own
849 variety as a valid basis for the retrieval of the encoded message.
850 In relation to question (1), a view that takes intelligibility as being "just a scale"
851 without any empirically identifiable thresholds predicts no interaction effect between
852 intelligibility and comprehensibility, expecting that comprehensibility simply decreases as
853 intelligibility decreases, since all points on the intelligibility scales are assumed to be equal.
854 Contrary to this assumption, Experiment I showed that comprehensibility ratings and
855 intelligibility scores do not decrease at the same rate, providing evidence that intelligibility
856 does allow for potentially meaningful segmentation. However, no evidence was found in
857 support of the idea implicit in question (1), namely that such segmentation would be provided
858 by listeners reporting an inability to retrieve the propositional content of utterances with low
859 intelligibility. Instead, the results showed that – below 70% intelligibly – listeners' estimation
860 of how much propositional content they were able to retrieve becomes unreliable,
861 consistently rating sentences as comprehensible while actually failing to retrieve their
862 propositional content. This interaction provided quantitative, experimental evidence of a
863 threshold at approximately 70% intelligibility, in line with several theoretical suggestions and
864 some anecdotal evidence from the literature (e.g. Aniansson & Peterson, 1983; Casad,
865 1974; Moore, 1989; Wang et al., 2012). Interestingly, the degree of overestimation in
866 comprehensibility ratings increases as intelligibility decreases, suggesting that listeners'
867 inability to reliably judge comprehensibility of an utterance is inversely proportional to the
868 utterance intelligibility. Consequently, we may conclude that while intelligibility itself is by
869 definition measurable on a linear scale, we can nevertheless achieve a meaningful
870 partitioning of the scale based on listeners' comprehensibility.
871 In response to question (2), results from both experiments II and III revealed a
872 positive answer, providing evidence that while intelligibility is measurable on a linear scale,
873 the concept of intelligibility is not itself linear, as it is not the case that all points along the
874 intelligibility scale are equal to all other points. Specifically, Experiment II showed that there

is an intelligibility range (i.e. between 34% and 71%) within which listeners consistently fail to retrieve sentential propositions beyond chance levels. This provides an initial answer to the question of how much distance is necessary before two varieties must be considered separate *Abstand* languages. While further research is necessary to narrow this range further and address the "how much is enough?" question more precisely, the results of Experiment II did provide a lower threshold of 34% intelligibility, thus suggesting an initial answer to a different yet related question, namely: beyond what point is it no longer tenable to talk of "same language"? Based on the results of Experiment II, future research is likely to find that this point is above 34% intelligibility.

Looking at the results from Experiment III, it is likely that the threshold of minimal intelligibility is closer to the upper point of 71% than to 34%. This is because while Experiment II was concerned with absolute failure to retrieve a propositional content beyond chance, the aim of Experiment III was to investigate the point beyond which a speaker cannot function as a successful member of a speech community whose variety is related to by phonetically distance from his or her own. This threshold would necessarily be higher than the one in Experiment II, since a speaker can fail to be functional in a speech community even though s/he is occasionally able to retrieve some propositional content, albeit not consistently and not always reliably. In this case, we saw that reducing intelligibility to 71% (i.e. the equivalent of 15% phonetic distance) renders listeners unable to reach the minimum TOEFL scores necessary to function at a social and professional level commensurate with their other, non-linguistic skills. Evidence of this comes from the fact that although participants were all undergraduates at a British university, when tested in a non-dialect that was only 71% intelligible with standard English, they were unable to meet the minimum language threshold for admission to undergraduate programmes. Note that participants received instructions both written and spoken, in standard English, something which would not have been the case had the non-dialect been the *Ausbau*-language of the society in which they were expected to be functioning. This suggests that 71% is likely to be a conservative threshold.

This is probably the finding with largest scope for applied linguistics, since it relates to the concept of speakers' functionality rather than to absolute failure to retrieve propositional content (the latter being a more extreme measure). Comparing results from Experiment III to those form Experiment I and to the literature on acceptable TOEFL scores (e.g. Moglen, 2015, see above for details) we can conclude that, when it comes to communicative functionality, the intelligibility threshold is firmly between the much narrower window of 70%-75%.

Given the potential as well as documented challenges facing people who are constrained to function in a speech community within which their native variety is only

partially reliable for successful communication and linguistic development (e.g. Bulatović, Schüppert, & Gooskens, 2019; Ibrahim & Aharon-Peretz, 2005; Leonardi, 2016; Saiegh–Haddad, 2003, *inter alia*) it seems pernicious to continue to maintain that two groups whose varieties stand at 15% phonetic distance speak "the same language", or to continue to define "languages" primarily on the basis of ideological construction and socio-political achievements, insisting on *Ausbau* considerations at the expense of *Abstand* measurements, as so many linguists have done (e.g. Comrie, 2009; Chambers and Trudgill, 1998; Janson, 2011; among many others).

In conclusion, the studies presented here have provided evidence against the widely held and hitherto untested assumption that intelligibility is simply a "matter of degree" with "no clear-cut" segmentation (e.g. Comrie, 2009), and revealed that intelligibility can be an empirically sound criterion of demarcation for the identification of languages and dialects. In view of these results, perhaps the time has come to reconsider the possibility that language might be a linguistic object after all.

References

Ammon, Ulrich. (1989): *Status and function of languages and language varieties*. Berlin, Walter de Gruyter.

Anderson-Hsieh, Janet, and Kenneth Koehler. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language learning*, *38*(4), 561-613.

Aniansson, Gunnar, and Yvonne Peterson. (1983). Speech intelligibility of normal listeners and persons with impaired hearing in traffic noise. *Journal of Sound and Vibration*, *90*(3), 341-360.

Anwyl-Irvine, Alexander, Jessica Massonnié, Adam Flitton, Natasha Kirkham, and Jo K. Evershed. (2019). Gorilla in our Midst: An online behavioral experiment builder. *Behavior research methods, 1-20.*

Appel, René, and Pieter Muysken. *Language contact and bilingualism*. Amsterdam University Press, 2005.

Bamgbose, Ayo. (1998). Torn between the norms: Innovations in world Englishes. *World Englishes*, *17*(1), 1-14.

Benincà, Paola, and Glanville Price. (2000). "Italy (Romance vernaculars): Introduction." In *Encyclopedia of the Languages of Europe*, edited by G. Price, 251–254. Oxford: Blackwell.

946 Blauberg, Maija S., and Martin D. Braine. (1974). Short-term memory limitations on decoding
947 selfembedded sentences. *Journal of Experimental Psychology, 102*, 745–748.

948 Bridgeman, Brent., Donald Powers, Elizabeth Stone, and Pamela Mollaun. (2012). TOEFL
949 iBT speaking test scores as indicators of oral communicative language
950 proficiency. *Language Testing*, *29*(1), 91-108.

951 Bulatović, Stefan., Ania Schüppert, and Charlotte Gooskens. (2019). Receptive
952 multilingualism versus ELF: How well do Slovenes understand Croatian compared to
953 Croatian speakers' English?. *Journal of English as a Lingua Franca*, *8*(1), 37-65.

954 Carrell, Patricia L. (2007). Notetaking strategies and their relationship to performance on
955 listening comprehension and communicative assessment tasks. *ETS Research*
956 *Report Series*, *2007*(1).

957 Casad, Eugene H. (1974). *Dialect intelligibility testing* (Microfiche ed.). Dallas: SIL.

958 Casserly, Elizabeth D., and David B. Pisoni. (2010). Speech perception and
959 production. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 629-647.

960 Chambers, Jack K., and Peter Trudgill. (1998). *Dialectology*. Cambridge University Press.

961 Cohen, Jacob. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.).
962 Hillsdale, NJ: Lawrence Earlbaum Associates.

963 Comrie, Bernard. (Ed.). (2009). *The world's major languages*. Routledge.

964 Craft, Justin T., Kelly E. Wright, Rachel Elizabeth Weissler, and Robin M. Queen. (2020).
965 Language and Discrimination: Generating Meaning, Perceiving Identities, and
966 Discriminating Outcomes. *Annual Review of Linguistics*, *6*.

967 De Swaan, Abram. (1991). Notes on the emerging global language system: regional,
968 national and supranational. *Media, Culture & Society*, *13*(3), 309-323.

969 Derwing, Tracey M. and Murray J. Munro. (2009) Comprehensibility as a factor in listener
970 interaction preferences: Implications for the workplace. *Canadian Modern Language*
971 *Review 66*, 181–202.

972 Dixon, Robert M. W. (1997). *The rise and fall of languages*. Cambridge, Cambridge
973 University Press.

974 Dunbar, Robert. (2001). Minority language rights in international law. *International and*
975 *comparative law quarterly*, *50*(01), 90-120.

Duran, Robert. L., and Lynne Kelly. (1985). An investigation into the cognitive domain of communication competence. *Communication Research Reports, 2*, 112-119.

ETS. (1998). Computer-based TOEFL score user guide.

ETS. (2017). TOEFL iBT test and score data.

Farnsworth, Timothy L. (2013). An investigation into the validity of the TOEFL iBT speaking test for international teaching assistant certification. *Language Assessment Quarterly*, *10*(3), 274-291.

Fasold, Ralph W. (2005). Making languages. In *Proceedings of the 4 th International Symposium on Bilingualism*.

Feigelson, Eric. (2012). Classification in Astronomy. *Advances in Machine Learning and Data Mining for Astronomy*, 1. In Way, M. J., Scargle, J. D., Ali, K. M., & Srivastava, A. N. (Eds.) *Advances in machine learning and data mining for astronomy*. CRC Press.

Fromkin, Victoria., Robert Rodman, and Nina Hyams. (2013). *An introduction to language*. Cengage Learning.

Garinther, George R., Leslie A. Whitaker, and Leslie J. Peters. (1995). The effects of speech intelligibility on military performance. In *Proceedings of the Symposium on Speech Communication Metrics and Human Performance*, pp. 72-82.

Gathercole, Susan. E., Catherine S. Willis, Alan D. Baddeley, and Hazel Emslie. (1994). The children's test of nonword repetition: A test of phonological working memory. *Memory*, *2*(2), 103-127.

Gooskens, Charlotte. 2007. 'The contribution of linguistic factors to the intelligibility of closely related languages.' *Journal of Multilingual and multicultural development*, 28(6), 445-467.

Gooskens, Charlotte. (2013). Experimental methods for measuring intelligibility of closely related language varieties. *The Oxford handbook of sociolinguistics*, 195-213.

Gooskens, Charlotte, Wilbert Heeringa, and Karin Beijering. (2008). Phonetic and lexical predictors of intelligibility. *International Journal of Humanities and Arts Computing*, 2(1-2), 63-81.

Gordon-Brannan, Mary., and Barbara W. Hodson. (2000). Intelligibility/severity measurements of prekindergarten children's speech. *American Journal of Speech-Language Pathology*, *9*(2), 141-150.

Greenberg, Joseph H. (1971). *Language, culture, and communication*. Stanford University Press.

Gupta, P. K. (2007). *Genetics: Classical to modern*. Rastogi Publications.

Hammarström, Harald. (2008). Counting Languages in Dialect Continua Using the Criterion of Mutual Intelligibility*. *Journal of Quantitative Linguistics*, *15*(1), 34-45.

Hawkins, John A. (2009). Germanic languages. In Comrie, B. (ed.). *The world's major languages* (pp. 66-73) Routledge.

Hicks, Marilyn M. (1989). The TOEFL computerized placement test: Adaptive conventional measurement. *ETS Research Report Series*, *1989*(1), 1-29.

Hilton, Nanna. H., Charlotte Gooskens, and Anja Schüppert. (2013). The influence of non-native morphosyntax on the intelligibility of a closely related language. *Lingua*, *137*, 1-18.

Hospenthal, Duane R., and Michael G. Rinaldi. (Eds.). (2007). *Diagnosis and treatment of human mycoses*. Springer Science & Business Media.

Hudson, Richard A. (1996). *Sociolinguistics*. Cambridge University Press.

Ibrahim, Raphiq, and Judith Aharon-Peretz. (2005). Is literary Arabic a second language for native Arab speakers?: Evidence from semantic priming study. *Journal of Psycholinguistic Research*, *34*(1), 51-70.

Isaacs, Talia, and Ron I. Thomson. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135-159.

Iwashita, Noriko, Annie Brown, Tim McNamara, and Sally O'Hagan. (2008). Assessed levels of second language speaking proficiency: How distinct?. *Applied linguistics*, *29*(1), 24-49.

Janson, Tore. (2011). *The History of Languages: An Introduction*. Oxford: Oxford University Press.

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford University Press.

Jeong, Hoi Ok, and Yoon Sil Kim. (2016). North Korean women defectors in South Korea and their political participation. *International Journal of Intercultural Relations*, *55*, 20-31.

Jochnowitz, George. (2013) *Dialect boundaries and the question of Franco-Provençal*. Vol. 147. Walter de Gruyter.

Kachru, Braj B. (2008). The first step: the Smith paradigm for intelligibility in world Englishes. *World Englishes*, *27*(3-4), 293-296.

Kalikow, Daniel N., Kenneth N. Stevens, and Lois L. Elliott. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, *61*(5), 1337-1351.

Kalyan, S. and Francois, A. (2019). When the waves meet the trees. *Journal of Historical Linguistics, 9*(1), 168-177.

Kang, Okim, Ron I. Thompson, and Meghan Moran. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, 68, 115–146.

Kang, Okim, Ron I. Thompson, and Meghan Moran. (2018). Which features of accent affect understanding? Exploring the intelligibility threshold of diverse accent varieties. *Applied Linguistics*, online first, 1-29.

Kauffeld, Cynthia. (2016). Andalusian Spanish: a diachronic survey of its origins and footprint in the Americas. In Núñez Méndez, E (ed.) *Diachronic Applications in Hispanic Linguistics* (pp. 167-178). Cambridge Scholars Publishing.

Kemp, Charlotte. (2009). Defining multilingualism. *The Exploration of Multilingualism: Development of research on L3, multilingualism and multiple language acquisition*, *6*, 11.

Kibbee, Douglas A. (Ed.). (1998). *Language Legislation and Linguistic Rights: Selected Proceedings of the Language Legislation and Linguistic Rights Conference, the University of Illinois at Urbana-Champaign, March, 1996* (Vol. 2). John Benjamins Publishing.

Kloss, Heinz. (1967). 'Abstand languages' and'ausbau languages'. *Anthropological linguistics*, 29-41.

Knutson, Thomas J., and Sutirat Posirisuk. (2006). Thai relational development and rhetorical sensitivity as potential contributors to intercultural communication effectiveness: JAI YEN YEN. *Journal of Intercultural Communication Research, 35*, 205-217.

Kurpaska, Maria. (2019). Dialects or Sinitic languages? In Huang, C. R., Jing-Schmidt, Z., & Meisterernst, B. (Eds.) *The Routledge Handbook of Chinese Applied Linguistics*. London: Routledge.

Le, Nguyen-Thinh., Frank Loll, and Niels Pinkwart. (2013). Operationalizing the continuum between well-defined and ill-defined problems for educational technology. *IEEE Transactions on Learning Technologies*, *6*(3), 258-270.

Leonardi, Mara M.V. (2016). Bilingualism or Trilingualism? Social versus linguistic views: Evidence from the Germanic-speaking language group in South Tyrol (Italy). PhD Dissertation.

Lepschy, Giulio C. 2002. *Mother Tongues & Other Reflections on the Italian Language*. Toronto: University of Toronto Press.

Lewis, Paul M. Gary F. Simons, and Charles D. Fennig. (2014). Ethnologue: Languages of the World, 17th edn, Dallas: SIL International.

Linscott, Richard J., and Jim van Os. (2010). Systematic reviews of categorical *versus* continuum models in psychosis: evidence for discontinuous subpopulations underlying a psychometric continuum. Implications for DSM-V, DSM-VI, and DSM-VII. *Annual Review of Clinical Psychology* 6.

Lyal, Chris, Paul Kirk, David Smith, and Richard Smith. (2008). The value of taxonomy to biodiversity and agriculture. *Biodiversity*, *9*(1-2), 8-13.

Mace, Georgina. M. (2004). The role of taxonomy in species conservation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359*(1444), 711-719.

Malmberg, Bertil. (2012). *Structural linguistics and human communication: an introduction into the mechanism of language and the methodology of linguistics* (Vol. 2). Springer Science & Business Media.

Merrifield, Glenys. (2012). An impact study into the use of IELTS by professional associations and registration entities: Canada, the UK and Ireland. *IELTS Research Reports Volume 11, 2012, 2nd edition*, 1.

Messias, Erick. (2003). Income inequality, illiteracy rate, and life expectancy in Brazil. *American Journal of Public Health*, *93*(8), 1294-1296.

Moglen, Daniel. (2015). The Re-Placement Test: Using TOEFL for Purposes of Placement. *CATESOL Journal*, *27*(1), 1-26.

Montgomery, James W. (2000). Verbal working memory and sentence comprehension in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *43*(2), 293-308.

Moore, Thomas J. 1989. Speech Technology in the Cockpit. In *Proceedings of Aviation Psychology*, edited by R.S. Jensen, 50-65. Aldershot: Gower Technical.

Moseley, Christopher. (2008). *Encyclopedia of the world's endangered languages*. Routledge.

Munro, Murray J., and Tracey M. Derwing. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning*, *45*(1), 73-97.

Munro, Murray J., and Tracey M. Derwing. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and speech*, *38*(3), 289-306.

Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P., and Leinonen, T. (2011). Gabmap-a web application for dialectology. *Dialectologia*, special issue 2: 65-89.

Nerbonne, J. and Heeringa, W. (2010). Measuring dialect differences. In: Auer, P. and Schmidt, J. E. (eds.), Language and Space. An international Handbook of Linguistic Variation. Volume 1: Theories and Methods. Berlin and New York: De Gruyter Mouton, pp. 550-567.

Nunberg, Geoffrey. (1997). Double Standards. *Natural Language and Linguistic Theory*. 15: 667-675.

Major, Roy C., Susan F. Fitzmaurice, Ferenc Bunta, and Chandrika Balastubramanian. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly, 36*, 173–190.

Ockey, Gary. J., Spiros Papageorgiou, and Robert French. (2016). Effects of strength of accent on an L2 interactive lecture listening comprehension test. *International Journal of Listening, 30*(1–2), 84–98.

O'neill, Thomas R., Richard J. Tannenbaum, and Jennifer Tiffen. (2005). Recommending a minimum English proficiency standard for entry-level nursing. *Journal of Nursing Measurement*, *13*(2), 129-146.

Otheguy, Ricardo, Ofelia García, and Wallis Reid. (2015). Clarifying translanguaging and deconstructing named languages: A perspective from linguistics. *Applied Linguistics Review, 6*(3), 281–307.

Pereltsvaig, Asya. (2017). *Languages of the world: An introduction*. Cambridge: Cambridge University Press.

Peterson, Townsend, A. (2006). Taxonomy is important in conservation: a preliminary reassessment of Philippine species-level bird taxonomy. *Bird Conservation International*, *16*(02), 155-173.

Peterson, Christopher H., Allen Wysocki, and Stephen B. Harsh. (2001). Strategic choice along the vertical coordination continuum. *The International Food and Agribusiness Management Review*, *4*(2), 149-166.

Posner, Rebecca. (1996). *The Romance languages.* Cambridge University Press.

Powers, Donald. E. (1985). A survey of academic demands related to listening skills. *ETS Research Report Series*, *1985*(2), i-63.

Rajadurai, Joanne. (2007). Intelligibility studies: A consideration of empirical and ideological issues. *World Englishes*, *26*(1), 87-98.

Ringe, Don. (2017). *From Proto-Indo-European to Proto-Germanic* (Vol. 1). Oxford University Press.

Ringe, Don, and Ann Taylor. (2014). *The Development of Old English* (Vol. 2). OUP Oxford.

Romaine, Susanne. (2000). *Language in society: An introduction to sociolinguistics*. Oxford University Press.

Rosenfeld, Michael, Susan Leung, and Philip K. Oltman. (2001). The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels (Research Memorandum No. RM-01-03). Princeton, NJ: Educational Testing Service.

Saiegh–Haddad, Elinor. (2003). Linguistic distance and initial reading acquisition: The case of Arabic diglossia. *Applied Psycholinguistics*, *24*(3), 431-451.

Salminen, Tapari. (2007). Endangered Languages in Europe. In Brenzinger, M. (Ed.). (2007). *Language diversity endangered* (Vol. 181). Walter de Gruyter.

Saunders, Gabrielle. H., and Kathleen M. Cienkowski. (2002). A test to measure subjective and objective speech intelligibility. *Journal of the American Academy of Audiology*, *13*(1), 38-49

Sawaki, Yasuyo, and Susan Nissan. (2009). Criterion-related validity of the TOEFL iBT listening section (Research Report No. RR-09-02). Princeton, NJ: Educational Testing Service.

Sewell, Andrew. (2010). Research methods and intelligibility studies. *World Englishes*, *29*(2), 257-269.

Sheppard, Beth. E., Nancy C. Elliott, and Melissa M. Baese-Berk. (2017). Comprehensibility and intelligibility of international student speech: Comparing perceptions of university EAP instructors and content faculty. *Journal of English for Academic Purposes*, *26*, 42-51.

Schüppert, Anja. (2011). *Origin of Asymmetry. Mutual intelligibility of spoken Danish and Swedish.* University of Groningen: Grodil 94.

Siegel, Jeff. (2010). *Second dialect acquisition*. Cambridge: Cambridge University Press.

Simner, Marvin L., and John B. Mitchell. (2007). Validation of the TOEFL as a Canadian university admissions requirement. *Canadian Journal of School Psychology*, *22*(2), 182-190.

Smith, M. K., & Bailey, G. H. (1980). Attitude and activity: Contextual constraints on subjective judgments. In *Language* (pp. 209-215).

Smith, Larry E., and Cecil L. Nelson. (1985). International intelligibility of English: Directions and resources. *World Englishes*, *4*(3), 333-342.

Speelman, Dirk., Leen Impe, and Dirk Geeraerts. (2014). Phonetic distance and intelligibility in Dutch. *Pluricentricity: Language Variation and Sociocognitive Dimensions*, *24*, 227-242.

Stavans, A., & Hoffmann, C. (2015). *Multilingualism*. Cambridge University Press.

Tan, Morse. (2016). North Korea now: turning point for a regime of rightlessness?. In *Routledge Handbook of Korean Culture and Society* (pp. 176-190). Routledge.

Tang, Chaoju, and Vincent J. van Heuven. (2009). Mutual intelligibility of Chinese dialects experimentally tested. *Lingua*, *119*(5), 709-732.

Taylor, Carol A., and Paul J. Angelis. (2008). The evolution of the TOEFL. In Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.) *Building a validity argument for the Test of English as a Foreign Language*. Routledge, 27-54.

Thomason, S. G. (2001). *Language contact*. Edinburgh, Edinburgh University Press.

Trudgill, Peter. (1996). Dual-source pidgins and reverse creoloids: northern perspectives on language contact. In Jahr, E. H., & Broch, I. (Eds.) *Language contact in the Arctic: Northern pidgins and contact languages* (Vol. 88). Walter de Gruyter.

Tulloch, S. (2006). Preserving dialects of an endangered language. *Current Issues in Language Planning*, *7*(2-3), 269-286.

Wagner, Elvis. (2016). A study of the use of the TOEFL iBT® test speaking and listening scores for international teaching assistant screening. *ETS Research Report Series*, *2016*(1), 1-48.

Wang, Hong-yan. (2007). *English as a lingua franca: Mutual intelligibility of Chinese, Dutch and American speakers of English*. Netherlands Graduate School of Linguistics (LOT), Utrecht.

Wang, W. H., Hwang, T. Z., Chang, C. H., & Lin, Y. C. (2012). Reconstruction of pharyngeal defects with a submental island flap after hypopharyngeal carcinoma ablation. *ORL*, *74*(6), 304-309.

Wei, Li. (2000). Dimensions of Bilingualism. In Wei, L. (ed.) *The Bilingualism Reader*. London: Routledge.

Wheeler, Quentin D. (2004). Taxonomic triage and the poverty of phylogeny. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *359*(1444), 571-583.

Woll, Bence, Rachel Sutton-Spence, and Frances Elton. (2001). Multilingualism: The global approach to sign languages. *The sociolinguistics of sign languages*, 8-32.

Yang, C. (2012). Classifying Lalo languages: Subgrouping, phonetic distance, and intelligibility. *Linguistics of the Tibeto-Burman Area, 35*(2), 113.

Yang, Wonyoung., and Murray Hodgson. (2006). Auralization study of optimum reverberation times for speech intelligibility for normal and hearing-impaired listeners in classrooms with diffuse sound fields. *The Journal of the Acoustical Society of America*, 120(2), 801-807.

Yule, George. (2014). *The Study of Language: An Introduction*, 5th edition. Cambridge, Cambridge University Press.

1223

1224

1225

1226