SEGÉDKÖNYVEK A NYELVÉSZET TANULMÁNYOZÁSÁHOZ 168.

A NYELV – RENDSZER, HASZNÁLAT, ALKALMAZÁS

Pszicholingvisztikai tanulmányok V.

LANGUAGE – SYSTEM, USAGE, APPLICATION

Studies in Psycholinguistics 5.

Szerkesztette/Editors

BÁTYI SZILVIA – VÍGH-SZABÓ MELINDA

TINTA KÖNYVKIADÓ
BUDAPEST, 2015

SEGÉDKÖNYVEK A NYELVÉSZET TANULMÁNYOZÁSÁHOZ 168.

Sorozatszerkesztő
KISS GÁBOR

Szerkesztette
BÁTYI SZILVIA
VÍGH-SZABÓ MELINDA

Lektorálta
GÓSY MÁRIA
LENGYEL ZSOLT
DAVID SINGLETON

ISSN: 1419-6603
ISBN: 978-963-409-002-1

A kiadvány a Tradeorg nyomdában készült.

A kötet kiadását támogatta a TÁMOP-4.1.2.D-12/1/KONV-2012-0017 számú projekt.

A borítón Erdélyi Ernő *Balatoni stég* című olajfestménye látható.

A kiadásért felelős
a TINTA Könyvkiadó igazgatója

A RELATIVE MEASURE OF THE INTERLANGUAGE SPEECH INTELLIGIBILITY BENEFIT: A META-ANALYTIC EXERCISE

VINCENT J. VAN HEUVEN

1. Introduction

It is commonly recognized that native speakers and native listeners outperform foreign speakers and listeners of the language. For instance, native (L1) listeners generally find fellow native talkers more intelligible than non-native (L2) talkers, particularly in noisy conditions (Munro, 1998; Munro & Derwing, 1995). L1 speakers cause fewer word perception errors than foreign speakers of the language do. In a classical study, word recognition by native listeners for Serbian-, Japanese- and Punjabi-accented English was some 36% poorer than for native English speech in a range of signal-to-noise ratios and filtering conditions (Lane 1967). More recently, it was shown that the word error rate of English spoken with a Mandarin accent was 11% against a mere 4% for native American control speakers, when in both cases the listeners were Americans (Munro & Derwing, 1995). Using a different methodology, native-speaker superiority was measured in terms of the Speech Reception Threshold (SRT). SRT was found to be at a 4-dB poorer signal-to noise ratio when the Dutch listeners responded to Dutch speakers, than when the speakers were British learners of Dutch (Van Wijngaarden, 2001).

By the same token, L1 listeners have better scores, faster recognition times, and withstand more adverse listening conditions than L2 listeners do – at least when the test materials are recorded from fellow L1 speakers. Native listeners are better at recognizing degraded speech (telephone speech, synthetic speech, speech in noise) than non-native speakers. For instance, Dutch listeners could recognize Dutch words from shorter onset portions than English learners of Dutch, even if the latter had resided in the Netherlands for twenty years or more (Nooteboom & Truin 1980).

In the studies summarized above information is exchanged between a native speaker and a native listener as the control condition and a native/nonnative pair of interactants

for the experimental condition. In this comparison the native/non-native pair is consistently outperformed by the native/native control pairs. Note that the comparison does not involve pairs of interactants who are both non-native speakers of the language used. Somewhat surprisingly, it has been observed that non-native speakers may be more intelligible than native speakers when the listener is also non-native. Indeed, second-language learners often report that the speech of a fellow non-native talker is easier to understand than the speech of a native talker. Bent and Bradlow (2003) advanced two hypotheses with respect to this phenomenon. The first hypothesis holds that a foreign talker of a language is more intelligible to any foreign listener of that language than a native speaker is. This is what Bent and Bradlow call the non-matched (or 'mixed') interlanguage speech intelligibility benefit. Early evidence in support of this hypothesis has been provided by Nash (1969). The second, more restricted, hypothesis predicts that a foreign talker will be more intelligible to a foreign listener (than a native talker would be) only if the foreign talker and listener share the same mother tongue. This is what Bent and Bradlow call the matched (or 'shared') interlanguage benefit.

The theoretical underpinning of the unrestricted hypothesis seems somewhat tenuous. It has been observed that non-native talkers speak rather slowly and hesitantly, which would benefit anyone who would have problems with decoding the message. The slow speed of delivery and the insertion of pauses when the speaker is looking for words would allow the non-native listener time to integrate what has been heard and to predict upcoming words. The beneficial effect of insertion of pauses (with compensation for slower rate of delivery) has been demonstrated for low-quality Dutch speech synthesis and for natural Dutch speech in noise (Scharpff & Van Heuven, 1988, Van Heuven & Scharpff, 1991, Scharpff, 1994), as well as for Danish perceived by Swedish listeners (Gooskens & Van Bezooijen, 2014). Moreover, the foreign talker will use a fairly restricted vocabulary comprised of high-frequency words only so that the listeners will not often be confronted with unfamiliar words. The benefit will probably disappear, I would argue, if the test materials were produced by a native speaker of the target language and manipulated such that the words and sentence structures (after minimal correction) and the gross temporal organisation (speed of delivery as well

location and length of pauses) would be the same as that used by the non-native talker. I am not aware of any such study, however, so that my objection remains speculative.

Evidence supporting this more restricted hypothesis has been provided by many studies, e.g. Smith & Rafiqzad (1979), Van Wijngaarden (2001), Van Wijngaarden et al. (2002), Imai (2003), Wang & Van Heuven (2003, 2004, 2006), and Wang (2007). It has been shown on many occasions that native speakers have a vast knowledge of the statistical regularities at all linguistic levels (sounds, syllables, morphemes, words and sentences) and skillfully use any redundancy that may exist in the native language system. These skills are much less developed in non-native listeners. The sound categories of the target language are less well defined in the perceptual representation of non-natives, and transitional probabilities that allow the native listener to predict upcoming sounds (or restore sounds that were missed) are not known to (let alone used by) the non-native listener. This does not only apply to non-native listeners who have learned the foreign language as adults but it has been shown that even the sound categories in a second language that was acquired before the age of four (i.e. by so-called early bilinguals) are less well defined than for monolingual listeners (Sebastian-Galles & Soto Faraco 1999).

Bent and Bradlow (2003) tested both hypotheses in one integrated experiment and found evidence in support of both. They point out that specific combinations of foreign speaker and listener language backgrounds yield better intelligibility scores than combinations involving a native speaker or listener, both when language backgrounds of the foreign speakers and listeners are mixed and when they are shared. However, the authors do not quantify the effect in a way that allows the reader to determine the magnitude of the interlanguage benefit, nor to check whether the benefit is larger for the shared interlanguage than for the mixed interlanguage situation. The purpose of the present article is to provide a simple computational method to express the magnitude of the (shared or mixed) interlanguage benefit and to re-analyse the results of a number of earlier studies on these phenomena. This meta-linguistic exercise will show, first of all, that the proposed relative measure of the interlanguage benefit yields the predicted effects (much more clearly so than when some absolute measure of the benefit is applied), and that the benefit is indeed larger when speakers and listeners have a shared native language between them than when the interactants have different native

languages. Moreover, the meta-analysis will ascertain whether the mutual intelligibility is poorest when one of the interactants (whether in the role of speaker or than of listener) is native and the other is non-native. I will call this the case of the native-speaker handicap.

In the next section, I will first explain the computational procedure that should be applied to compute the proposed relative measure of interlanguage benefit. Here I will use an example taken from Wang (2007). In the later sections I will re-analyse earlier results by Smith and Rafiqzad (1979), by Bent and Bradlow (2003) and the set of six tests used by Wang (2007).

2. Computing the relative interlanguage speech intelligibility benefit (R-ISIB)

In this section I will demonstrate how a relative measure of the Interlanguage Speech Intelligibility Benefit can be computed. The data are taken from a large study on the mutual intelligibility of Dutch, Mandarin and American speakers of English described in more detail by Wang (2007), Van Heuven & Wang (2007) and Wang & Van Heuven (2014); see also experimental detail in section 5. Twenty speakers (10 males, 10 females) from each of these three different native-language backgrounds produced materials in English, i.e. (i) vowels in a /hVd/ context, (ii) consonants and (iii) consonant clusters in intervocalic contexts, (iv) semantically unpredictable sentences (SUS), and (v) semantically meaningful sentences with final target words in unpredictable ('non-pregnant') and (vi) predictable ('pregnant') contexts. The materials of one representative male and one female speaker for each of the three language backgrounds were then offered for identification (of vowels, consonants and clusters) or recognition (of words in sentences) to 36 listeners in each of three countries, so that all nine possible combinations of speaker and listener backgrounds occurred equally often in the experiment.

The results of the first part of the materials, i.e. the vowel perception test, are given in Table 1. The observed scores (column marked 'Obs.') are the mean percent correct vowel identification scores for each of the nine combinations of speaker and listener language backgrounds. In absolute terms, the best intelligibility scores are obtained

when both speakers and listeners are native (75% correct vowel identification). It is not the case, however, that native-non-native speaker-listener combinations yield consistently poorer intelligibility scores than pairs exclusively involving non-native interactants – in contradistinction to what the interlanguage intelligibility benefit hypothesis predicts. In fact, the poorest results are obtained when both speakers and listeners are Chinese (30%), and the best result is found for the combination of Dutch listeners to American speakers (61%). Nor is it the case that non-native speaker-listener combinations that share the same language between them (30% for Chinese-Chinese and 59% for Dutch-Dutch) yield consistently better scores than mixed non-native combinations (34% and 40% for Chinese-Dutch and Dutch-Chinese, respectively). Clearly, then, testing the interlanguage speech intelligibility benefit (ISIB) hypothesis in absolute terms fails miserably.

Now let us look at these results in rather more relative terms. I argue that the 30% correct vowel identification obtained by the Chinese-Chinese speaker-listener combination, although the lowest score of all in absolute terms, is in fact much better than should be expected in comparison with the other scores. Van Heuven and Wang (2007) proposed a fairly simple computational method based on linear modeling to quantify the magnitude of the relative ISIB (or R-ISIB), which is basically the interaction component that remains after the main effects of speaker language and listener language have been factored out. An illustration of the method is given in Table 1.

	Language background of						Exp.	Obs.	Δ
	Listener			Speaker					
1.	Chinese	33	–16	Chinese	39	–10	22	30	**+8**
2.	Chinese	33	–16	Dutch	52	+3	35	34	–1
3.	Chinese	33	–16	Am. English	56	+7	40	34	–6
4.	Dutch	53	+4	Chinese	39	–10	42	40	–2
5.	Dutch	53	+4	Dutch	52	+3	55	59	**+4**
6.	Dutch	53	+4	Am. English	56	+7	60	59	–1
7.	Am. English	61	+12	Chinese	39	–10	50	45	–5
8.	Am. English	61	+12	Dutch	52	+3	63	61	–2
9.	Am. English	61	+12	Am. English	56	+7	68	75	**+7**
	Grand mean		0			0	49	49	0

Table 1. *Expected vowel identification scores (% correct) on the basis of grand mean (= 49%) and main effects for Listener and Speaker L1. Observed scores (Obs.) and residuals (Δ) are indicated. Bolded delta's represent the interlanguage (or native language) benefit. All percentages have been rounded off to the nearest integer*

The computational procedure involves the following steps.

1. Compute the grand mean score across all speaker-listener combinations. This is 49% correct in the present example.
2. Next, compute the mean score for each of the speaker groups (by averaging over the listener groups). For instance, the mean score for Chinese speakers is 39, which is the mean of Chinese speakers combined with Chinese, Dutch and American listeners, with scores of 30%, 40% and 50%, respectively.[1]
3. Likewise, compute the mean scores for each of the listener groups, averaged over speakers. This yields mean scores of 33%, 53% and 61% for Chinese, Dutch and American listeners, respectively.
4. Then compute the deviation of the speaker means from the grand mean by subtraction. For instance, the mean of the Chinese speaker group (39%) is 10 points below the grand mean of 49%, hence a deviation of –10.

[1] On face value, these three numbers should average out at 40% instead of 39%. The discrepancy is due to rounding errors.

5. Similarly, compute the deviation of each listener group mean from the grand mean. The mean of the Chinese listener group (33%) is 16 points below the grand mean, hence a deviation of –16.
6. Then compute the expected score for each speaker-listener combination, by adding the speaker group deviation and the listener group deviation to the grand mean. In the case of the Chinese-Chinese speaker-listener combination this would be 49% (grand mean) – 16 (listener group deviation) – 10 (speaker group deviation) = 22%.[2]
7. Finally, compute the prediction error ('residual') for each speaker-listener combination, which is the difference between the expected and the observed score. For the Chinese-Chinese combination we expect 22% but find 30%, so that the residual equals +8 points. This is the value for R-ISIB.[3] Note that the mean R-ISIB for each row and each column in the matrix, as well as for the matrix in its entirety, should always add up to zero, since positive and negative prediction errors should cancel each other out.

When the listeners are Chinese, Dutch and American, the expected mean scores are –16, +4 and +12 relative to the grand mean; for the three speaker language backgrounds the expected mean should be additionally corrected with –10, +3 and +7, respectively. Note here that the size of the increments/decrements is larger for listener language background than for speaker language background, i.e. the listener effect is larger than the speaker effect.

Generally, the observed scores are correctly predicted or even overestimated by the linear addition of the two main effects. Only in three combinations of factor levels is the observed score substantially better than the prediction. These are precisely the conditions in which the listeners are confronted with vowel tokens spoken by their fellow countrymen ('shared interlanguage', shaded rows in Table 1). The native or interlanguage benefit is 4 to 8 percentage points better than the expected score. It appears that there is no need to differentiate between communication between a native speaker and a native listener (with a R-ISIB of +7 points, which could be called a 'native-language benefit') and communication between a non-native speaker and a non-

[2] On the basis of the values presented in table 1, an expected value of 21% would be expected. The slight discrepancy is due to greater rounding accuracy in the computations underlying the table.
[3] The numbers presented in this table deviate slightly from what was published in Wang (2007) and Van Heuven & Wang (2007). The present numbers are correct.

native listener who share the same native language (+4 and +8 points for Dutch and Chinese matched interlanguage groups, respectively): in both situations the residual is of comparable, positive magnitude.

In the case of a speaker-listener combination with a mixed interlanguage the R-ISIB is very close to zero: –1 for Dutch-Chinese and –2 for Chinese-Dutch). This would indicate that, indeed, the shared interlanguage yields a substantially greater benefit than the mixed interlanguage. There are too few observations to run any meaningful statistics on the difference; this we will do in a later section of this article where we will test this effect on data aggregated over a number of studies.

R-ISIB is most negative when the speaker-listener combination involves one native and one non-native party. Here the R-ISIB ranges between –1 and –6 points. Again, we will defer statistical testing of the significance of this native-language handicap until we have sufficient aggregate data.

3. (R-)ISIB in Smith and Rafiqzad (1979)

The earliest study to compare the intelligibility of native and non-native Englishes in a sufficiently complete matrix of speaker and listener groups with a variety of language backgrounds was probably done by Smith and Rafiqzad (1979). Speakers were educated teachers of English in their own country, between 20 and 40 years of age, who had not lived in an English-speaker country for more than four consecutive months, had not been trained in schools directed by native speakers of English, and who had never lived in English-speaking groups or families. Listeners were educated students or professionals, sampled from a variety of disciplines. 'Typical' materials (selected to the discretion of the speakers) were read to classroom audiences by L2 speakers of English in seven Asian countries, viz. Hong Kong, India, Japan, Korea, Malaysia, Nepal, and the Philippines. Similar materials were collected from native speakers of American English. Unfortunately, the design was incomplete in that no materials of any speaker group were presented to American native listeners. There were also non-native listener groups that were never used as speakers – these I pruned from the matrix below.[4] The

[4] Smith & Rafiqzad (1979) have been criticized for other reasons as well. It has been pointed out that the materials produced by the speaker groups differ substantially in terms of conceptual comprehensibility –

materials were presented to the seven relevant listener groups in a Cloze test, in which listeners saw a printed version of the audible text, with every sixth word replaced by a blank to be filled in.

The results of this experiment are summarized Table 2, which lists the percentage of key words correctly filled in for each combination of eight speaker groups and seven listener groups.

Speakers	**Listeners**							
	HK	In	Ja	Ko	Ma	Ne	Ph	**Mean**
Hong Kong	**80**	58	47	12	60	9	42	**44**
India	89	**92**	71	36	94	55	97	**76**
Japan	95	92	**94**	45	88	43	85	**77**
Korea	86	90	82	**55**	75	36	67	**70**
Malaysia	95	90	73	37	**83**	42	84	**72**
Nepal	84	92	64	45	62	**75**	87	**73**
Philippines	83	89	64	16	81	25	**79**	**62**
USA	78	82	60	29	67	23	74	**59**
Mean	**86**	**86**	**69**	**34**	**76**	**39**	**77**	**67**

Table 2. *Percentage of key words correctly filled in English materials spoken by speakers of eight different native language groups (rows) and listened to by subjects from the same native language groups. Note that no American native listeners participated. Data from Smith and Rafiqzad (1979)*

The results support the hypothesis that native speakers are not necessarily better understood than non-native speakers when the listeners are themselves non-native. In fact, the native speakers consistently rank between the sixth and eighth (i.e. last position) for each of the seven listener groups. The mean score of the American native speakers is the second lowest mean (59% correct), the second poorest mean after the Hong Kong speakers (44% correct). Neither is it true, in absolute terms, that speaker-listener combinations that share the same native language between them yield consistently better scores than any other combination: this is the case only for three out of seven groups. Nevertheless, there is a general tendency for non-native speaker-

so that no straightforward comparisons between speaker and listener groups can be made. This criticism, of course, is no longer valid once we apply the concept of R-ISIB. When the speaker of a language is more difficult, for whatever reason, this will affect the main effect of speaker but not the speaker by listener interaction, i.e. not the R-ISIB.

listener combinations to yield better scores than combinations of a native speaker and a non-native listener. In absolute terms, then, the following results obtain.

(i) Non-native English is generally better understood by non-native listeners than native (American) English. This is the case in 40 out of (8 × 7 =) 56 combinations of speaker and listener language backgrounds, leaving 16 counterexamples in which native English is superior to non-native English.

(ii) It is not the case that the matched interlanguage yields consistently better scores than the mixed interlanguage: in the total of 7 × 6 = 42 cases, the matched interlanguage yields better results than the mixed interlanguage in 28 against 14 comparisons.

The mean intelligibility scores for native speaker, mixed interlanguage and shared interlanguage are 59, 66 and 80%, respectively. The differences are not significant by a one-way Analysis of Variance, $F(2, 53) = 1.4$ ($p = .225$, $\eta^2 = .050$).

Let us now look at the same results in relative terms, applying the concept of R-ISIB. The results are as in Table 3.

Speakers	**Listeners**							
	HK	In	Ja	Ko	Ma	Ne	Ph	**Mean**
Hong Kong	**16.5**	–4.9	.4	.4	6.5	–6.8	–12.1	**.0**
India	–6.8	**–3.2**	–7.9	–7.9	8.2	7.0	10.6	**.0**
Japan	–1.9	–4.3	**14.0**	–.1	1.1	–6.2	–2.6	**.0**
Korea	–3.6	1.0	9.2	**17.2**	–4.6	–5. 9	–13.3	**.0**
Malaysia	3.5	–.9	–1.6	–2.6	**1.5**	–1.8	1.9	**.0**
Nepal	–8.2	.4	–11.3	4.7	–20.2	**30.5**	4.2	**.0**
Philippines	1.1	7.7	–1.1	–14.1	9.1	–9.2	**6. 5**	**.0**
USA	–.5	4.1	–1.6	2.4	–1.5	–7.8	4.9	**.0**
Mean	**.0**	**.0**	**.0**	**.0**	**.0**	**.0**	**.0**	**.0**

Table 3. *Results of Table 2 expressed in relative terms using R-ISIB. Further see Table 2.*

The mean R-ISIB scores for the same three conditions as above are 0% (native speaker), –2% (mixed interlanguage) and 12% (shared interlanguage). The one-way Analysis of Variance is highly significant, $F(2, 53) = 10.6$ ($p < .001$, $\eta^2 = .286$). Post-hoc tests

(Bonferroni correction, α = .05) of differences between means indicate that the R-ISIB due to shared interlanguage is significantly better than that obtained by the other two conditions, which do not differ from each other.

It would appear from this exercise that the relative measure yields clearer results also in statistical terms. Unlike other studies, however, the Smith and Rafiqzad data do not differentiate between the mixed interlanguage benefit and the native speaker handicap we identified before.

4. (R-)ISIB in Bent and Bradlow (2003)

Bent & Bradlow (2003) examined the interlanguage benefit in a database with mutual intelligibility scores in English obtained for five types of speakers: one high-proficiency and one low-proficiency Korean L2 speaker of English, one high-proficiency and one low-proficiency Chinese L2 speaker of English, and one native speaker of American English. Sentences produced by these five (female) speakers were presented to four groups of listeners with Chinese (N = 21), Korean (N = 10), American (N = 21) and mixed-foreign (N = 12) backgrounds. Intelligibility scores were determined for all 5 × 4 = 20 combinations of speaker and hearer L1 backgrounds.

Table 4 shows the results in absolute terms. The scores are not in percentages but in Rationalised Arcsine Units (RAUs). The arsine transform was applied by Bent and Bradlow to unwarp the bottom and top ranges of the percentage scale in order to compensate for bottom and ceiling effects. After ‘rationalisation’ the transformed scale extends between −17 and +117 RAU; 50 RAU = 50 per cent (Studebaker, 1985).

Speakers	Listeners				
	NN Chin	NN Kor	NN Mixed	L1 Am	**Mean**
Chinese_High	64	60	62	77	**66**
Chinese_Low	30	22	19	38	**27**
Korean_High	66	74	70	91	**75**
Korean_Low	41	53	41	60	**49**
USA	56	60	67	109	**73**
Mean	**51**	**54**	**52**	**75**	**58**

Table 4. *Intelligibility scores (in RAU) for five selected speakers (female Chinese and Korean speakers of English with high and low proficiency, and one American native speaker) as perceived by four groups of listeners. NN: Non-native. Adapted from Bent and Bradlow (2003: Table 3). Further see text*

In absolute terms, the results of this experiment do not consistently support the interlanguage benefit hypothesis. The proficient Korean speaker is most intelligible on average, even more so than the American native speaker. This Korean speaker is even more intelligible to Chinese listeners than the high-proficiency Chinese speaker is. Most damning for the interlanguage intelligibility hypothesis is that American listeners understand any speaker best, irrespective of the speaker's language background.

When we apply the relative notion of the interlanguage benefit, the results are much more interpretable. Table 5 presents the R-ISIB values, analogous to Table 4. Now the results are much more in line with the predictions. The greatest benefit is observed between native speakers and native listeners. Shared interlanguage (indicated by bold numbers in shaded cells in Table 5) has consistently positive values; also, the benefit is larger for poor L2 speakers than for good L2 speakers. This makes sense, since the poor speakers will exhibit the phonology of the native language more strongly than the good L2 speakers (the interlanguage of the latter group will be closer to the norms of the target language). American native speakers are exceptionally difficult to understand for Chinese and Korean listeners, and less so for non-native listeners of other language backgrounds. The condition with mixed interlanguage (i.e. NN mix as well as Chinese speakers with Korean listeners and vice versa) generally has neither a positive nor a negative R-ISIB: this condition assumes an intermediate position with R-ISIB values close to zero.

Speakers	Listeners				
	NN Chin	NN Kor	NN Mix	L1 Am	**Mean**
Chinese_High	**4.9**	−1.5	2.5	−5.8	**.0**
Chinese_Low	**9.4**	−1.0	−2.0	−6.3	**.0**
Korean_High	−2.7	**3.0**	1.0	−1.3	**.0**
Korean_Low	−1.1	**8.5**	−1.5	−5.8	**.0**
USA	−10.4	−8.8	.2	**19.0**	**.0**
Mean	**.0**	**.0**	**.0**	**.0**	**.0**

Table 5. *Results of Table 4 expressed in relative terms using R-ISIB. Further see Table 4.*

The number of conditions is too small to allow meaningful statistics to be computed. We will defer statistical testing to a later section, in which data of several studies will be aggregated.

5. (R-)ISIB in Wang (2007)

In this section I will present the results of all six mutual intelligibility tests described Wang's (2007) doctoral thesis, which I will briefly summarized in the next few paragraphs; see also section 2 and references given there for experimental detail).

Wang's main experiment contained the complete sets of materials for all five test parts (see section 2), but only those spoken by the six optimally representative speakers, as identified in the earlier speaker selection test. Part 1 included the 19 /hVd/ words of all six speakers in random order (across speakers) and preceded by six practice stimuli, which yielded a total of 120 items. The /hVd/ frame, such as in the words *heed*, *hid*, *head*, *had*, etc., is fully productive in English so that all English vowels may occur as a word or short phrase (Peterson and Barney, 1952). Therefore, the consonant environment does not provide the listener with any useful information about the identity of the vowel. Part 2 contained the 24 /ɑCɑ/ (all intervocalic consonants in a non-word) items in random order (across speakers) at a total of 150 items (including six preceding practice stimuli). Part 3 contained the six (speakers) × 21 /ɑCC(C)ɑ/ (a selection of intervocalic clusters in a non-word) items in random order, preceded by four practice items (130 all together). In Part 4, a selection of SUS sentences (semantically unpredictable sentences such as *The state sang by the long week,* see also Benoît, Grice

& Hazan, 1998) was presented such that each speaker contributed a single, lexically different, sentence in each syntactic frame so that the test contained 5 (frames) × 6 (speakers) = 30 sentences (with a total of 112 content words) in random order across frames and speakers (and preceded by five practice sentences, one for each different syntactic frame). Because Part 4 is a word recognition task, in which a word that has been recognized earlier would have an advantage when presented the second time due to learning effects (so-called 'priming'), it was necessary to block sentences over the speakers such that the same content word was never presented twice to the same listener. Part 5, finally, contained SPIN (Speech in Noise) sentences (Kalikov, Stevens & Elliott, 1974). Each of the six speakers contributed eight different sentences. The same sentence was never presented more than once to the same listener (blocking). The set of 48 sentences was preceded by two practice sentences (one high predictable, one low predictable), which yields a total of 50 SPIN sentences.

The materials were presented to 36 native listeners of Dutch (Leiden, from the City Belt in the West of the country), 36 Chinese listeners (Mandarin-speakers in Changchun) and to 36 American listeners (South Californian English-speaking, tested at the University of California at Los Angeles, USA). Each group of listeners comprised 18 men and 18 women. Listeners participated in the experiment on a voluntarily basis, had no self-reported hearing deficiencies, and received (the equivalent of) € 10 for their services.

The stimuli were presented in small lecture rooms over headphones. In Parts 1, 2 and 3 were the listeners were instructed to make a forced choice from the 19 (part 1), 24 (part 2) or 21 (part 3) response alternatives, which were printed on their answer sheets. Listeners had to make a single choice at all times or gamble in case of doubt. Each item was offered only once, with a pause of 7 seconds in between items in the first half of every part of the test and of 5 seconds pause in the second half of every part (because the listeners could then find their way on the answer sheet more quickly). In Part 4, the entire sentence was made audible just once. Then the sentences were repeated incrementally such that the sentence was truncated after the first content word during the first repetition and after the second content word on the second repetition and so on, until at last even the final content word was made audible. Listener had answer sheets in front of them with the function words printed per sentence while the content words had

been replaced by a line of uniform length, as in *Why does the _____ _____ the _____?* After each repetition listeners were given 3 seconds to fill in the next content word in the sentence. Then the whole sentence was repeated one more time to allow the listeners a final opportunity to make changes.[5] In Part 5, the listeners' task was just to write down the final word of each following sentence. The subjects did not receive a printed version of the spoken sentences.[6] The entire experiment took about 90 minutes, with a pause in the middle.

Table 6 lists the raw means for each of nine combinations of Chinese (Mandarin), Dutch and American speakers of English obtained in six tests, i.e. vowel identification, single consonant identification, cluster identification, semantically unpredictable sentences, as well as high and low predictability keywords in meaningful sentences. Each test was done by 36 listeners, the same individuals for each language group. The correlations (for details see Wang, 2007: chapter 10) between the results of the six tests were so low that I will consider these tests to constitute statistically independent data.

Speakers	Listeners	Tests					
		Vowels	Consonants	Clusters	SUS	SPIN_LP	SPIN_HP
Chinese	Chinese	29.7	57.2	52.8	39.3	19.4	16.7
	Dutch	40.3	66.6	78.8	57.1	26.9	33.1
	USA	**44.9**	**72.5**	**82.5**	**59.5**	**39.4**	**57.8**
Dutch	Chinese	33.5	46.8	36.9	39.0	38.9	37.8
	Dutch	59.3	73.7	**87.8**	**86.2**	**81.3**	76.1
	USA	**61.0**	**76.1**	85.7	83.0	67.7	**99.4**
USA	Chinese	33.1	58.2	56.0	44.2	17.9	31.8
	Dutch	58.6	80.6	89.1	90.5	77.8	84.9
	USA	**75.3**	**85.7**	**89.3**	**95.5**	**95.2**	**99.1**

Table 6. *Summary of test results. Percent correct on each of six tests broken down by language background of speaker and broken down further by native language of listener Each mean is based on 36 listeners. The listener group with the absolute best performance is represented in bold face*

[5] Since the semantically unpredictable sentences are basically meaningless, the SUS test measures on the listeners' speech recognition ability rather than speech understanding.

[6] No formal check was performed to ascertain whether the listeners understood the sentences they heard. However, the very nature of the task presupposes that the sentence-final keyword should be easier to supplete if the listener graps the contents of the preceding part of the sentence.

The rows in Table 6 where an absolute ISIB is predicted, are shaded. When the listeners are Chinese, the effect never happens; without a single exception, the American speakers are most intelligible to the Chinese listeners, Dutch speakers are always second, and the Chinese speakers are always least intelligible. When the listeners are Dutch, an absolute ISIB is found in three out of six tests. When the listeners are American, the fellow native speakers are always most intelligible, although the difference with the Dutch speakers is negligible in the consonant cluster identification test. So, if we follow our earlier reasoning and accept the absence of any interlanguage as a valid case of shared interlanguage (in this case shared absence), the ISIB hypothesis makes the right prediction in 9 out of 18 test cases (50%). If we omit the all-American speaker/listener combination, the ISIB is found in 3 out of 12 comparisons (25%).

Let us now analyse the results after conversion to relative ISIB values (i.e. R-ISIB). Table 7 lists the results.

Speakers	Listeners	Tests						Mean
		Vowels	Consonants	Clusters	SUS	SPIN_LP	SPIN_HP	
Chinese	Chinese	**7.7**	**6.3**	**6.1**	**12.5**	**17.0**	**11.7**	10.2
	Dutch	−2.3	−3.9	−4.6	−6.8	−12.1	−7.8	−6.3
	USA	−5.4	−2.4	−1.5	−5.8	−5.0	−3.9	−4.0
Dutch	Chinese	−1.5	−4.2	−8.7	−5.2	2.5	−2.4	−3.3
	Dutch	**3.7**	**3.1**	**5.6**	**4.9**	**8.3**	−.1	4.3
	USA	−2.3	1.1	2.9	.3	−10.8	**2.5**	−1.1
USA	Chinese	−6.3	−2.1	**2.5**	−7.3	−19.5	−9.3	−7.0
	Dutch	−1.4	.7	−1.1	1.9	3.8	**7.9**	2.0
	USA	**7.6**	**1.4**	−1.5	**5.5**	**15.7**	1.4	5.0
Mean		**.0**	**.0**	**.0**	**.0**	**.0**	**.0**	**.0**

Table 7. *Same as Table 6 but values are relative ISIB (R-ISIB) scores*

If we omit the all-American speaker/listener combination, the R-ISIB is found in 11 out of 12 comparisons (92%). If we accept the absence of any interlanguage as a valid case of shared interlanguage, the R-ISIB hypothesis makes the right prediction in 15 out of 18 test cases (83%). This is a large improvement over testing the interlanguage benefit in absolute terms.

6. Aggregated data

As a last exercise I will now perform a statistical analysis across all data that were discussed above. We will specifically test two related hypotheses. The first is that (1a) there will be a strong interlanguage intelligibility benefit such that two non-natives with the same mother tongue will understand each other best when speaking a foreign language (shared interlanguage), (1b) two non-natives with different native language backgrounds will understand each other more poorly (mixed interlanguage), and (1c) the poorest intelligibility will be observed when a non-native communicates with a native speaker (whether as speaker or as listener). The second hypothesis is that these predictions will be borne out more clearly when using the relative measure of the ISIB than when looking at absolute intelligibility scores.

The aggregate data contain 130 cases, i.e., the total number of speaker-listener group combinations in either Tables 2-4-6 (for absolute ISIB scores) or Tables 3-5-7 (for relative R-ISIB scores). The six tests in Tables 6 and 7 will be treated as uncorrelated, so that these tables contribute 9 (speaker-listener group combinations) × 6 (independent tests) = 54 cases to the dataset. Table 8 presents the mean ISIB and R-ISIB values for four types of speaker-listener group combinations, i.e. combinations yielding (i) shared interlanguage, (ii) mixed interlanguage, (iii) native/non-native pair and (iv) native-native pairs (as a control condition).

In terms of absolute interlanguage benefit, the results indicate that all-native speaker-listener pairs yield near-ceiling intelligibility scores (93%), which is ca. 30 percentage points better than any of the three combinations involving one or two non-native interactants; these three speaker-listener combinations do not differ from each other by a post-hoc comparison of means (Bonferroni-corrected, after one-way ANOVA, see Table 8). The results obviously contradict the hypothesis that there is any benefit to be gained by non-natives, whether they do or do not share an interlanguage: all non-natives are equally handicapped, whether communicating with a native or with each other.

Interlocutor pair	Absolute ISIB			Relative ISIB		
	Mean	SD	N	Mean	SD	N
1. Shared interlanguage	63.1	21.9	26	6.8	8.5	26
2. Mixed interlanguage	59.4	24.7	62	−2.4	6.1	62
3. One native	63.8	21.5	35	−2.3	5.3	35
4. All native	92.7	10.7	7	7.0	7.7	7
ANOVA	$F(3, 126) = 4.5$, $p = .005$, $\eta^2 = .096$			$F(3, 126) = 16.2$, $p << .001$, $\eta^2 = .279$		
Posthoc (Bonferroni, $\alpha = .05$)	{2, 1, 3} < {4}			{3, 2} < {1, 4}		

Table 8. *Mean (absolute) ISIB and R-ISIB broken down by four types of speaker-listener group combinations, aggregated over all experiments reviewed in this paper*

In relative terms, however, the situation is much more as predicted. First of all, non-natives with a shared interlanguage enjoy the same intelligibility benefit as two natives, with positive R-ISIB values of 6.8 and 7.0, respectively. Moreover, when speaker and listener have a non-matched (mixed) interlanguage, the R-ISIB is negative (−2.4). There is a clear difference, then, between the matched and the non-matched interlanguage pairs to the effect that no benefit remains when speaker and listener have different native languages. The idea of a native speaker handicap, however, is not supported by the aggregate data. It is not the case that a non-native listener is at a greater disadvantage when communicating with a native speaker than when communicating with a non-native with whom he does not share the native language background. Not only are the R-ISIB results more in line with the hypotheses formulated in the literature, they are also statistically more reliable, given that the effect size (η^2) of the speaker-listener combination is roughly three times larger in relative (R-ISIB) than absolute (ISIB) scores (see Table 8).

7. Conclusion and discussion

On the basis of the literature I formulated two hypotheses with respect to the effect of the specific composition of a speaker-listener pair involving different combinations of native and non-native interactants. The first hypothesis predicted (1a) that two non-natives will understand each other in English best when they have the same native-

language background and (1b) will perform better than when they have different native language backgrounds. These subhypotheses proved false when tested in absolute terms but were clearly supported by the data when evaluated in relative (R-ISIB) terms. In fact, in relative terms, non-native speaker-listener pairs enjoy the same interlanguage benefit as all-native speaker-listener pairs. One more subhypothesis, which was formulated on the basis of earlier analyses of Wang's (2007) data, cannot be upheld in the meta-analysis: (1c) it is not the case that communication between native and non-native interactants is poorer than between two non-natives with different language backgrounds. On the strength of this latter finding, hypothesis (1c) has to be rejected. Not only were the results more germane to these predictions, also the second hypothesis was upheld by the data, namely that the effects would be stronger when evaluated in relative rather than in absolute scores. The ANOVA indicated an effect size in R-ISIB that was three times larger than when analysed in absolute scores.

In terms of substance, then, the meta-analysis boils down to a very simple and clear-cut binary division in intelligibility between native and nonnative speakers of a language. When two interactants share the same native language, they enjoy the advantage of a shared phonology (as well as a shared morpho-syntax). In this sense native speakers communicating with native listeners also share a common interlanguage, namely the ideal (near-)perfect grammar/phonology of the native speaker/ listener. When two interactants do not have the same mother tongue, their mutual intelligibility is poorer. Here, it does not matter whether both interactants are non-native or whether a foreigner communicates with a native – the point is that they do not share any interlanguage.

Finally, there is a methodological conclusion to be drawn. As I pointed out in the introduction, it has been observed that non-native listeners of English often have the intuition that they understand a fellow non-native talker, i.e., one with whom they share a common mother tongue, better than a native speaker of English. This intuition is supported by the experimental data, but only when the intelligibility scores are expressed in relative terms, i.e., in terms of the R-ISIB measure that was explained in section 2. I conclude, therefore, that the proper way of evaluating the concept of the interlanguage speech intelligibility benefit, as formulated by Bent and Bradlow (2003), is in relative rather than in absolute terms.

References

Benoît, C., Grice, M. and Hazan, V. (1996). The SUS test: A method for the assessment of text-to speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Comm.*, 18, 381–392.

Bradlow, A. R. and Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors, *J. Acoust. Soc. Am.*, 106, 2074–2085.

Gooskens, C. S. and Bezooijen, R. van (2014). The effect of pause insertion on the intelligibility of Danish among Swedes. In Caspers, J., Chen, Y., Heeren, W., Pacilly, J., Schiller, N. O. and Zanten, E. van (eds) *Above and beyond the segments. Experimental linguistics and phonetics*. Amsterdam: John Benjamins. 96–108.

Heuven, V. J. van and Scharpff, P. J. (1991). Acceptability of several speech pausing strategies in low quality speech synthesis: interaction with intelligibility. *Proc. 12th Int. Cong. Phonetic Sc.*, Aix-en-Provence. 458–461.

Heuven, V. J. van and Wang, H. (2007). Quantifying the interlanguage speech intelligibility benefit. In: Barry, W. and Trouvain, J. (eds) *Proc. 16th Int. Cong. Phonetic Sc.* Saarbrücken: Universität des Saarlandes. 1729–1732.

Imai, S., Flege, J. E., and Walley, A. (2003). Spoken word recognition of accented and unaccented speech: Lexical factors affecting native and nonnative listeners, in *Proc. Int. Cong. Phonetic Sc.*, Barcelona, Spain.

Kalikow, D. N., Stevens, K. N. and Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *J. Acoust. Soc. Am.*, 61, 1337–1351.

Lane, H. (1963). Foreign accent and speech distortion. *J. Acoust. Soc. Am.*, 35, 451–453.

Munro, M. J. and Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners, *Lang. Learn*., 45, 73–97.

Munro, M. J. (1998). The effects of noise on the intelligibility of foreign accented speech, *Stud. Second Lang. Acquisit.*, 20, 139–153.

Nash, R. (1969). Intonational interference in the speech of Puerto Rican bilinguals, *J. English*, 4, 1–42.

Nooteboom, S. G. and Truin, P. G. M. (1980). Word recognition from fragments of spoken words by native and non-native listeners. *IPO An. Progr. Rep.*, 15, 42–47.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *J. Acoust. Soc. Am.*, 24, 175–184.

Scharpff, P. J. and Heuven, V. J. van (1988). Effects of pause insertion on the intelligibility of low quality speech. In: Ainsworth, W. A., Holmes, J. N. (eds) *Proc. 7th FASE/Speech-88 Symposium*, Edinburgh: Institute of Acoustics. 261–268.

Scharpff, P. J. (1994). *Het effect van spreekpauzes op de herkenning van woorden in voorgelezen zinnen [The effect of speech pauses on the recognition of words in read-out sentences]*, Diss. Leiden University.

Sebastian-Galles, N. and Soto-Faraco S. (1999). Online processing of native and non-native phonemic contrasts in early bilinguals. *Cognition*, 72, 111–123.

Smith, L. E. and Rafiqzad, K. (1979). English for cross-cultural communication: The question of intelligibility, *TESOL Quarterly*, 13, 371–380.

Studebaker, G. A. (1985). A 'rationalized' arcsine transform. *J. Speech, Lang, Hear. Res.*, 28, 455–462.

Wang, H. (2007). *English as a lingua franca: Mutual intelligibility of Chinese, Dutch and American speakers of English.* LOT dissertation series 147. Utrecht: LOT.

Wang, H. and Heuven, V. J. van (2003). Mutual intelligibility of Chinese, Dutch and American speakers of English. In: Fikkert, P. and Cornips, L. (eds) *Linguistics in the Netherlands 2003*. Amsterdam: John Benjamins. 213–224.

Wang, H. and Heuven, V. J. van (2004). Cross-linguistic confusion of vowels produced and perceived by Chinese, Dutch and American speakers of English. In Cornips, L. and Doetjes, J. (eds) *Linguistics in the Netherlands 2004.* Amsterdam: John Benjamins. 205–216.

Wang, H. and Heuven, V. J. van (2006). Acoustical analysis of English vowels produced by Chinese, Dutch and American speakers. In Weijer, J.M. van de and Los, B. (eds) *Linguistics in the Netherlands 2006*. Amsterdam: John Benjamins. 237–248.

Wang, H. and Heuven, V. J. van (2014). Is a shared interlanguage beneficial? Mutual intelligibility of American, Dutch and Mandarin speakers of English. In: Rupp, L. and Doel, R. van den (eds) *Pronunciation Matters. Accents of English in the Netherlands and elsewhere*. Amsterdam: VU University Press. 175–194.

Wijngaarden, S. J. van (2001). Intelligibility of native and non-native Dutch speech. *Speech Commun*., 35, 103–113.

Wijngaarden, S. J. van, Steeneken, H. J. M., and Houtgast, T. (2002). Quantifying the intelligibility of speech in noise for non-native listeners. *J. Acoust. Soc. Am.*, 111, 1906–1916.