'Broken Finnish': Speaker L1 and its recognition affecting rating in National Certificates of Language Proficiency test in Finnish

Sari Ahola
University of Jyväskylä, Finland

Mia Halonen
University of Jyväskylä, Finland

Abstract

As many European countries have language proficiency requirements for obtaining citizenship, language testing is a possible source of social inequality. The 'Broken Finnish' project has been set up to ensure test fairness by addressing the proficiency rating in the National Certificates of Language Proficiency (NCLP) test for Finnish in Finland, with a special focus on perceptions of pronunciation and 'accent' in relation to the examinee's L1 and how the raters recognise them. We explore if and how these perceptions affect the proficiency ratings. We are also interested in studying where the perceptions might arise from. In this paper, we present results from one L1 group: Thai speakers.

Introduction: Accent perceptions in societal gatekeeping

Like many European countries, Finland uses language proficiency requirements as one of the gatekeepers for citizenship. Consequently, language testing is a possible source of social inequality. To ensure the fairness of a test system, in the project *'Broken Finnish': Accent perceptions in societal gatekeeping* (Academy of Finland; 2018–2022, www.jyu.fi/hytk/fi/laitokset/solki/broken-finnish/in-english), we study the rating process in the National Certificates of Language Proficiency (NCLP) test in Finland, with a focus on perceptions of pronunciation in relation to the examinee's L1 and how the raters recognize it. We study whether there is any bias towards test-takers, and whether the correct recognition of the speakers' L1 affects the ratings of oral proficiency. We will focus on the ratings of Thai L1 speakers and show 1) how the correct recognition of the Thai speakers' L1 influenced their rating, and 2) how the raters described the performances of the speakers.

Context of the study: Finnish National Certificates of Language Proficiency (NCLP) test

The NCLP is a test system for adult second and foreign language learners overseen by the Finnish government's official system for language proficiency testing in a total of nine languages. Since 1994, there have been approximately 130,000 test-takers, and there are approximately 8,000 L2 Finnish test-takers per year.

The organizations responsible for the test system are Finnish National Agency for Education (EDUFI) and University of Jyväskylä (Centre for Applied Language Studies). It is independent of any syllabus or curriculum, but applies the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). NCLP has its own rating scale varying from 1–6 in four skills. Levels 3 and 4 form the so-called intermediate level which equals CEFR Levels B1and B2, which are the citizenship threshold in Finland; since 2003, CEFR Level B1 in Finnish or Swedish has been the language requirement for Finnish citizenship.

The focus of our project is the speaking test of Finnish at the intermediate level. The performances are rated along seven criteria for speaking (aligning with the CEFR): 1) general criteria (a holistic criterion covering six other, more analytical criteria), 2) fluency, 3) flexibility, 4) coherence and cohesion, 5) vocabulary (range, accuracy, idiomaticity), 6) pronunciation and phonological control, and 7) grammatical accuracy.

Data of the project

In order to explore whether the test-takers face any biased rating, we designed a data set focusing on test-takers of five different L1s: Arabic, Estonian, Russian, Thai, and Finland Swedish. Informed by previous research, we knew that the speakers of these languages face negative stereotyping in Finland (e.g. Jaakkola, 2009; McRae, Bennett, & Miljan, 1988; Reuter & Kyntäjä, 2005; Sirkkilä, 2005.) Our hypothesis was that, because of these general stereotypes, these groups might also face biased rating.

The data were collected through an online platform in 2015 and 2016. Altogether 50 (10 for each L1) examinees' speaking performances were rated by 44 certified raters. The performances were rated both holistically (the general criterion) and separately for the six analytical criteria presented above. The raters were also asked to write down the speaker's L1, to indicate how certain they were about the assumption, and to describe the speaker, the performance and the bases for their assumptions of the L1.

The three-fold interaction between L1, the various criteria and the L1 assumption (whether it was accurate, that is, 'recognized', or not) was studied using multi-faceted Rasch analyses. In addition to these statistical analyses done by the project's statisticians, we analyzed how the performances were described in the open answers.

Findings: L1 recognition affects on the rating of Thai L1 speakers

The results confirm the hypothesis in that the correct recognition of the L1 affected the rating in the case of Thai speakers. If the L1 was recognized, there was a significant decrease in the rating of pronunciation as well as a significant increase in the ratings of fluency and the general criterion. Furthermore, the other criteria seem to be affected along with recognition, but these changes were only tendencies, not reaching statistical significance. In this section, we will present some of the raters' reasoning for the recognition as well as descriptions of the affected skills, pronunciation, fluency and general holistic impression, drawing on Ahola's previous research (in press).

On one hand, the L1 Thai female speakers (n=8) were very well recognised by the raters. The recognition seemed to be based on the raters' experience in teaching and testing L1 female Thai speakers. On the other, L1 Thai male speakers (n=2) were not recognised by the raters but the suggested L1s included Russian, Somali and Arab speakers. The most probable reason for this bias in gender-related (non)recognition is simply that the raters do not have experience in teaching or rating L1 Thai male speakers because there are not many in Finland.

Negatively rated skill: Pronunciation

The criterion that was negatively rated, and thus rated lower when L1 was recognized, was pronunciation. The pronunciation was heavily criticized and the learning of comprehensible pronunciation was described as difficult or even impossible:

> For the speakers of these languages it (pronunciation) is so difficult to learn that they will never manage to raise their level any higher. One feels sorry for them.

> Source language can be strongly heard in the pronunciation. It requires attentiveness and patience from the listener and certainly repetition from the speaker to be understood in everyday situations.

The challenges in comprehension were described, for example, on the sound level. The raters described problems in producing consonants and consonant clusters (e.g. *r = l*, *-ts-* , *-st-*) which are typical problems of Thai language speakers and, more generally, speakers of tonal languages in Finnish (e.g. Aho, Toivola, Karlsson, & Lennes, 2016). In addition to sounds, reasons for comprehension challenges were put down to prosodic features, like high pitch or speech rhythm. Deviant prosody is known to easily reveal learners' L1 (Anderson-Hsieh, Johnson, & Koehler, 1992; Giles & Rakić, 2014).

> The Thai appear to find it difficult to produce speech. Stress is often on each word or individual words and there are pauses between words.

> Speech rhythm, intonation and phonology sound Asian throughout.

> Pronunciation also sounds naive, like little children's speech.

> A girly way to speak and intonation remind me of Thai speakers.

As can be seen in the last two extracts, the prosodic features were connected to the perceptions and stereotypes of Thai women. In these descriptions the raters describe more qualities of an imagined speaker than the performance itself.

Positively rated skills: Fluency and holistic impression (general criterion)

Fluency is connected to the amount of speech and the speakers' ability to fill the given time with speech as well as to the speech rate (e.g. Fillmore, 1979; Kormos & Dénes, 2004). This 'productivity' leads naturally to more output to be assessed. The Thai L1 speakers produced speech rather actively and there was more speech production compared to speakers of other L1s. This means that they were able to show their production fluency even though they had shortcomings in other language skills (Ang-Aw & Goh, 2011; May, 2006; Pollitt & Murray, 1996). The raters perceived and described this productivity in terms of fluency, which they also rated higher when they recognized the speaker as a Thai L1 speaker.

> Produces a lot of content. Difficult to assess – appearance of fluency but a lot of vagueness.
>
> Speech is fluent and there is a relative amount of speech production. Searches for words to an extent and the speech is partly listing items.
>
> Speaks nonstop without pausing to make things into logical wholes.
>
> The verb is often missing and the expressions are constructed by putting words together.
>
> A lot of speech together, nonstop. The speech is a little choppy structurally.

As we can see, perception of fluency does not depend on, for example, the perception of the level of proficiency of pronunciation, proven by the fact that fluency was rated higher while pronunciation was rated lower. Fluency is often compared to general proficiency or they are even used as synonyms for each other. This is seen in the behavior of the raters: they granted the Thai L1 speakers higher ratings in the general criterion, in their holistic impression of proficiency, than the ratings of the analytical criteria would have predicted.

Summary and implications

Our data showed that the recognition of the L1 of the speakers and the raters' perceptions of the speakers as a group had an effect on rating in the case of Thai L1 speakers. This is not an unexpected or surprising result as there is an extensive pool of research on rater bias based on speakers' background (see, e.g. Brennan & Brennan, 1981; Carey, Mannel, & Dunn, 2011; Cargile & Giles, 1997; Kang & Rubin, 2009; Lev-Ari & Keysar, 2010; Lindemann, 2005; Munro, 2003; Reid, Trofimovich, & O'Brien, 2019; Toivola, 2011).

Previous studies have often addressed so-called 'primed' ratings where the samples are preceded with, for example, real and fake photos of the (disguised) speakers and the raters are nonprofessionals in relation to language proficiency assessment. Our research differs from most of those in that it is done in a high-stakes test context and with trained raters. However, despite the education and experience in rating, the raters of our research are vulnerable to (possibly unconscious) stereotyping and the differences in their degree of linguistic awareness (Niedzielski & Preston, 2000; Preston, 1996).

Comprehensibility as a phenomenon is both a listener-specific and speaker-specific feature, and, unlike intelligibility, sensitive to various aspects independent of language or proficiency, like stereotyping and prejudices (see, e.g., Isaacs & Trofimovich, 2012; Munro & Derwing, 1995; Riney, Tagaki, & Inutsuka, 2005). In general, our results show that in addition to the formal criteria, there are some hidden, implicit or unconscious criteria: 'the raters' own' criteria, such as 'grammatical accuracy equals high proficiency', 'you cannot expect more of them', or a 'pity factor'. The rating seems to be dependent on the expectations of proficiency level, which seem to be set much lower than for many other L1s, e.g. for Estonians (see Ahola, 2020). These lower expectations, then, apparently connect to the image of the Thai L1 speakers living in rural peripheral areas as stay-home mothers and wives given less opportunities to learn Finnish. This is naturally a stereotyped view on the speaker group and does not apply to everyone, but according to statistics and previous research, this seems to be the case with this particular speaker group (e.g. Lumio, 2014; Shinyella, 2012; SVT, 2018)

No test is perfectly reliable, but it is obvious that in this kind of high-stakes test, there should not be any kind of bias involved in the rating process. However, it is also obvious that, where there are humans involved, there will also be emotions and attitudes. Now that we have more understanding of the bias, we are able to develop rater training further, also in relation to these more unconscious and sensitive areas, such as stereotyping. This will be implemented by giving the raters even more feedback on rater bias and raising awareness of rating behavior and hidden criteria, as well as training with more samples of different L1 speakers.

References

Aho, E., Toivola, M., Karlsson, F., & Lennes, M. (2016). Aikuisten maahanmuuttajien suomen ääntämisestä. *Puhe ja kieli, 32*(2), 77–96.

Ahola, S. (2020). Sujuvaa mutta viron kielen vaikutusta. *Virittäjä*, *124*(2), 217–242.

Ahola, S. (in press). Yleisten kielitutkintojen arvioijien käsityksiä thainkieliseksi tunnistettujen suomenoppijoiden suullisesta kielitaidosta, *Puhe ja Kieli*, *40*(4), 203–224.

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, *42*(4), 529–555.

Ang-Aw, H. T., & Goh, C. C. M. (2011). Understanding discrepancies in rater judgement on national-level oral examination tasks. *RELC Journal, 42,* 31–51.

Brennan, R. L., & Brennan, D. J. (1981). Measurements of accent and attitude towards Mexican-American speech. *Journal of Psycholinguistic Research 10,* 487–501.

Carey, M. D., Mannel, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing 28*(2), 201– 219.

Cargile, A. C., & Giles, H. (1997). Understanding language attitudes: Exploring listener affect and identity. *Language & Communication, 17*(3), 195–217.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Fillmore, C. J. (1979). On fluency. In C. J. Fillmore, D. Kempler & W.S-Y. Wang (Eds.), *Individual Differences in Language Ability and Language Behavior* (pp. 85–102). London: Academic Press.

Giles, H., & Rakić, T. (2014). Language attitudes: Social determinants and consequences of language variation. In T. M. Holtgraves (Ed.), *The Oxford Handbook of Language and Social Psychology* (pp. 11–26). Oxford: Oxford University Press.

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition, 34*(3), 475–505.

Jaakkola, M. (2009). *Maahanmuuttajat suomalaisten näkökulmasta. Asennemuutokset 1987–2007*. Helsinki: The City of Helsinki.

Kang, O., & Rubin, D. L. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*, *28*(4), 441–456.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*, 145–164.

Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, *46*, 1,093–1,096.

Lindemann, S. (2005). Who speaks 'broken English'? US undergraduates' perceptions of non-native English. *International Journal of Applied Linguistics*, *15*, 187–212.

Lumio, M. (2014). Hymyn takana – thainmaalaiset maahanmuuttajat ja suomalais-thainmaalaiset avioliitot. In E. Heikkilä, P. Oksi-Walter & M. Säävälä (Eds.) *Monikulttuuriset avioliitot sillanrakentajina* (s. 36–51). Turku: Siirtolaisinstituutti.

May, L. A. (2006). An examination of rater orientation on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Language Testing, 11*(1), 29–51.

McRae, K. D., Bennett, S. E., & Miljan, T. (1988). *Intergroup Sympathies and Language Patterns in Finland: Results from a Survey*. Helsinki: Suomen Gallupin julkaisusarja.

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

Munro, M. J., & Derwing, T. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*(1), 73–97.

Niedzielski, N. A., & Preston, D. R. (2000). *Folk Linguistics*. Berlin: De Gruyter Mouton.

Pollitt, A., & Murray, N. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance Testing, Cognition and Assessment. Selected Papers from the 15th Language Research Testing Colloquium* (pp. 74–91). Studies in Language Testing volume 3. Cambridge: UCLES/Cambridge University Press.

Preston, D. R. (1996). Whaddayaknow – modes of folk linguistic awareness. *Language Awareness*, *5*, 40–74.

Reid, K. T., Trofimovich, P., & O'Brien, M. G. (2019). Social attitudes and speech ratings: Effects of positive and negative bias on multi-age listeners' judgments of second language speech. *Studies in Second Language Acquisition 41*(2), 419–442.

Reuter, A., & Kyntäjä, E. (2005). Kansainvälinen avioliitto ja stigma. In T. Martikainen (Ed.), *Etnisyys Suomessa 2000-luvulla* (pp. 104–125). Helsinki: Finnish Literature Society.

Riney T. J., Tagaki, N., & Inutsuka, K. (2005). Phonetic parameters and perceptual judgements of accent in English by American and Japanese listeners. *TESOL Quarterly, 39*(3), 441–466.

Shinyella, T. (2012). *Kaksikulttuurista arkea suomalais-thainmaalaisissa lapsiperheissä. Selvitys suomalais-thainmaalaisten lapsiperheiden tilanteesta sekä erityispiirteistä, -tarpeista ja -haasteista.* Helsinki: Monikulttuuriyhdistys Familia Club ry.

Sirkkilä, H. (2005). *Elättäjyyttä vai erotiikkaa. Miten suomalaiset miehet legitimoivat parisuhteensa thainmaalaisen naisen kanssa?.* Jyväskylä Studies in Education, Psychology and Social Research 268. Jyväskylä: University of Jyväskylä.

SVT (2018). *Suomen virallinen tilasto: Väestörakenne*. Helsinki: Tilastokeskus.

Toivola, M. (2011). *Vieraan aksentin arviointi ja mittaaminen suomessa*. Helsinki: The University of Helsinki.