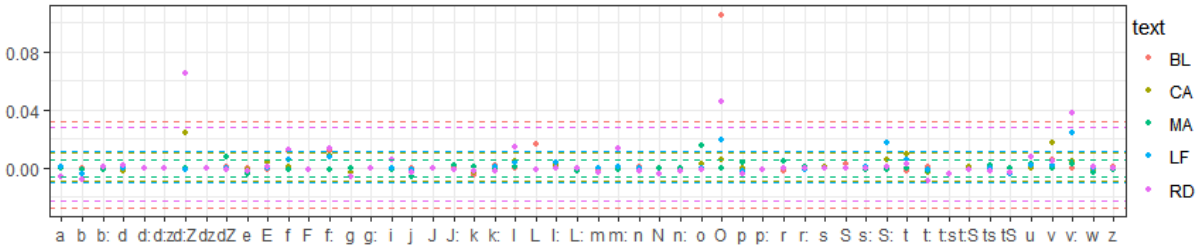
Phonetically-balanced vs phonetically-rich: clarifying common misunderstandings and evaluating Italian texts

Paolo Mairano, Enrico Zovato (Nuance Communications)

Introduction. Phonetically-balanced and/or phonetically dense texts are widely used in the fields of phonetics, audiometry / speech therapy, and speech technology. In general, these texts are relatively short passages (100-300 words) that are supposed to be somehow representative of the phonetic/phonological properties for a given language. However, phoneticians, speech therapists, audiometrists and speech scientists have different requirements, and what exactly is meant by 'representative of phonetic/phonological properties' may vary, causing a certain amount of misuses and misevaluations. In particular, such texts can be evaluated in terms of how much phonetic/phonological diversity they contain (we will call this: *phonetic richness*), or in terms of how phonetic/phonological units reflect their distribution in the language (we will call this: *phonetic balance*). Some authors (e.g. Gibbon, Rogers & Winski, 1997) call for texts that fulfil both requirements and even add lexical criteria. Jesus, Valente & Hall (2015) claimed that phoneme counts are a sufficient metric. Instead, we argue that phonetic richness and phonetic balance are distinct (often incompatible) properties which need to be evaluated with specific metrics. We here illustrate this by measuring the phonetic richness and the phonetic balance for 5 different Italian texts which are used in phonetic research: 3 versions of Aesop's fable *Il vento di tramontana e il sole* ([RD], [CA], [LF], see bibliography), *La Borea e il Favonio* ([BL]), and *Lo scherzo del pastore* ([MA]; possibly on the spur of Deterding, 2006).

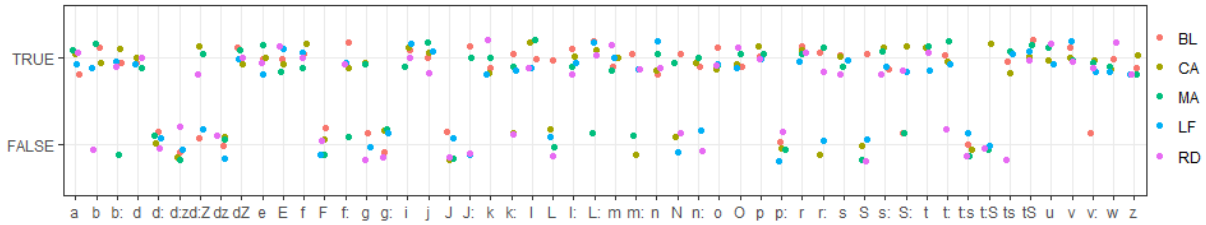
Data and methodology. According to the mainstream procedure in this field, we compared the 5 texts above with a large corpus of Italian taken as a reference (we used CELI corpus, 2001, 30.000.000 words). All 5 texts and the corpus were transcribed phonetically with the Italian component of MARY-TTS (Tesser et al., 2013), whose output transcriptions include the most relevant allophones for standard Italian, such as [ɱ] and [ŋ]. We then computed phoneme, diphone and triphone frequencies for each text.

Evaluating phonetic balance: results. We used Pearson's chi-squared test to compare phoneme frequencies for each of the 5 texts with the reference (ref) corpus. Only the distribution of phoneme frequencies for MA do not differ significantly from those found in the ref corpus (p=.17), while frequencies for all others present significant differences. Visual inspection with a modified Bland-Altman plot (figure below) reveals that some phones are overrepresented in the texts (above/below the 95% confidence interval): [ɔ] is in BL (due to several repetitions of *Borea* and *Favonio*), [dːʒ] and [v] in CA, [o] [dʒ] in MA, [ɔ] [ʃː] [v] in LF, [dːʒ] [ɔ] [vː] in RD.



Although phone(me) frequencies are the most often used metric for phonetic balance, they tell us little about the possible phonotactic combinations in a language. So, we turned our analysis to diphone and triphone frequencies. As far as diphones are concerned, frequencies found in CA, MA, LF seem to reflect those in the ref corpus (p=1, p=1, p=.68 respectively), while not so for BL and RD (p<.001 for both). In the case of triphones, the only text which does not reflect frequencies in the ref corpus is BL (p<.001), with most overrepresented triphones attributable to the words 'La Borea' and 'Favonio' (namely [abɔ, bɔr, ɔre, vɔn, ɔnj]).

Evaluating phonetic richness: results. We then evaluated phonetic richness of the 5 texts via a phone presence test, type counts and type/token ratios (TTR) of phones, diphones, and triphones. The phone presence test surprisingly revealed that all texts lack some phones of standard Italian, and even that some phones (e.g. [dː] [dːz] [dz] [ɱ] [gː] [pː] [tːs]) are missing entirely from all texts (see figure below – singletons and geminates are considered as separate items). BL is the richest text by number of represented phones (39), RD the poorest (33).



Type counts suggest that BL is the richest in terms of diphones (241), triphones (488) and syllable types (129), while RD is the poorest (respectively: 168, 271, 87). Conversely TTRs suggest that diphone, triphone, syllable *density* is highest for RD (5.2, 80, 107.6) and lowest for BL (3.3, 51.1, 67.7). Since these figures are affected by text length, normalization procedures are currently being considered. Additionally, we are investigating the possibility of using combinations of pertinent phonological traits, beyond diphones and triphones.

Discussion. Many considerations arise from these results and will be discussed at the conference. Most importantly, it has been shown that short texts evaluated as phonetically balanced by standard metrics may even lack phones, thereby proving that phonetic balance and phonetic richness are distinct properties that need to be evaluated independently, as already suggested by Gibbon et al. (1997).

References

[BL] Bertinetto, P. M., & Loporcaro, M. (2005). The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome. *JIPA*, *35*(2), 131-151.

[CA] Canepari, L. (2004) *Manuale di pronuncia italiana*, Bologna: Zanichelli.

Deterding, D. (2006). The North Wind versus a Wolf: short texts for the description and measurement of English pronunciation. *JIPA*, *36*(2), 187-196.

Gibbon, D., Moore, R., & Winski, R. (Eds) (1997). *Handbook of standards and resources for spoken language systems*. Walter de Gruyter.

Jesus, L.M.T., Valente, A.R.S. & Hall, A. (2015) Is the Portuguese version of the passage 'The North Wind and the Sun' phonetically balanced? *JIPA*, 45(1), 1-11.

[LF] *LFSAG archive*, retrieved from http://www.lfsag.unito.it/ark/tramontane.html

[MA] Matticchio, I. (2016). Italian in Istria. Prosodic differences in read Italian by Italians and Istrians. *Proc. of Spoken Communication 2016*, 108-109.

[RD] Rogers, D., & D'Arcangeli, L. (2004). Italian. *JIPA*, *34*(1), 117-121.

Tesser, F., Paci, G., Sommavilla, G., & Cosi, P. (2013). A new language and a new voice for MARY-TTS. *Proc. of the 9th AISV Congress*, 21-23 January 2013, Venice.