Typological tendencies in verse and their cognitive grounding

De Castro Arrazola, V.

7 Experimental testing of sensitivity to textsetting rules

7.1 Introduction

7.1.1 The problem of textsetting

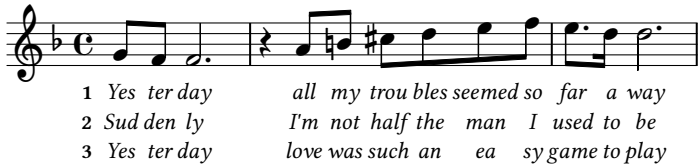
Songs can be analysed as composite objects consisting of a tier of words set to a tune (Dell & Halle 2009). Evidence for the independence of these two levels of structure can be found in strophic songs, where the same tune is repeated several times with different lyrics set to it (like in Example 7.1). The alignment of these two tiers has been shown to be non-random in a number of languages, and its systematicity is captured by so-called *textsetting constraints* (see Chapter 6, and Proto 2015 for an overview).

In general, two broad types of constraints are proposed in the literature: (1) constituent alignment, and (2) feature alignment. To illustrate the first type of constraint, in Example 7.1 we can observe that the beginning of the musical phrase always coincides with the beginning of a linguistic phrase, rather than starting in the middle of a word, for example.

The second class of constraints usually describe a correlation between a prosodic feature (like stress or tone) and a musical feature (like metrical prominence or pitch contour). For instance, the opening lyrics of the song *Yesterday* (Example 7.1), musically, start on the beat, i.e. the first note has greater metrical prominence than the following two notes; and linguistically we observe a similar pattern, where the syllable *yes-* has greater stress than the following *-terday*. We can describe this pattern as a decreasing stress contour (s–), matched to a decreasing metrical prominence contour (p–).

This kind of alignment, summarised as s–p–, involves *parallel* contours, whereas a pattern like s+p– involves *opposing* contours. We would obtain the latter alignment by keeping the same melody but replacing the word *yésterday* with a word like *tomórrow*, which shows an increasing stress contour (s+) in its first two syllables. In languages employing linguistic stress (like Italian or English),

Example 7.1: The opening melody of the song *Yesterday* by The Beatles, with three different lyrics set to it.



parallel contours (s+p+, s–p–) are preferred over opposing contours (s+p–, s–p+) (Hayes 2009; Proto & Dell 2013).

From a methodological point of view, most analysis of textsetting rely on corpus data. Combinations of linguistic and musical features which are statistically rare or fully absent from a given corpus are considered ill-formed. Nevertheless, a well-known limitation of corpora is that they can only provide positive data; that is, absent or under-represented patterns are hard to interpret (Schütze 2011; 2016). This is particularly evident in smaller datasets, where it is likely to find accidental gaps with no statistical implications.

7.1.2 Experimental approaches

Despite these shortcomings, corpus analyses, combined with authors' judgements, provide precise hypotheses about which alignments may be perceived as (un)acceptable by native speakers. The current study and a few preceding ones attempt to substantiate and refine these hypotheses using experimental methods.

Hayes & Kaun (1996) developed a production task in order to address the textsetting intuitions of native speakers of English and compare them to corpus-derived rules. Participants were shown the words of 670 lines taken from a corpus of folk songs and asked to chant them as set to a binary template. The proposed settings showed a high degree of between-subject agreement, and further supported the textsetting patterns found in the original corpus.

More recently, Gordon, Magne & Large (2011) used a combination of behavioural and brain imaging data to test the sensitivity of participants to different textsetting patterns. The stimuli consisted of short sentences sung to newly-composed melodies paired with an isochronous beat. In some trials, the beat matched the stressed syllables; in others, the beat was displaced so that it matched unstressed syllables instead. Immediately after each trial, subjects performed a lexical decision task, which was executed more quickly and more accurately in the well-aligned trials (i.e. beats match stresses). Regarding the EEG

Table 7.1: Sample textsetting alignments with parallel and opposing contours. Each pattern is a combination of a stress (s) contour and a prominence (p) contour. The syllable with the greatest stress or prominence within the word is marked with an accent or an underline respectively.

	Contour	Pattern	Example
1	Opposing	s–p+	*méisje* 'girl'
2	Opposing	s+p–	*pastóor* 'priest'
3	Parallel	s–p–	*méisje* 'girl'
4	Parallel	s+p+	*pastóor* 'priest'

recordings, an analysis at the alpha (8–12 Hz), beta (13–29 Hz), and low-gamma (30–50 Hz) bands further showed significant differences between the aligned and misaligned settings.

Unlike corpus-based studies, the materials used in experiments can be designed so as to cover an exhaustive range of linguistic and musical patterns. This gives the researcher finer-grained control over the hypotheses to test. Hence, experiments can potentially provide positive *and* negative data, narrowing down the characterisation of textsetting intuitions.

7.1.3 The present study

The purpose of the chapter is twofold: (1) to further our understanding of Dutch textsetting, (2) to describe a simple yet effective methodology which can be employed to uncover the textsetting intuitions of a community of speakers of a given language.

To the best of our knowledge, the preceding chapter constitutes the first dedicated study of textsetting in Dutch. Based on a corpus of ca. 3,700 songs, we described a number of under-represented text-to-tune alignments. Among content words (i.e. nouns and verbs), the most avoided patterns involve opposing contours of stress and metrical prominence, as illustrated in the first two examples of Table 7.1. Patterns with parallel contours were, unsurprisingly, very common (rows 3 and 4 of Table 7.1). As a notational shorthand, we indicate linguistic stress with an acute accent, and underline the syllable which receives the highest metrical prominence in the word.

Based on the results from Chapter 6, the present study tests two hypotheses. First, we predict that native speakers of Dutch will be sensitive to the marked

distributional contrast found in the corpus between parallel and opposing alignments. Second, we test whether they also show a difference in preference between the two opposing patterns, favouring s+p– cases like *pastóor* over s–p+ cases like *méisje* (see Table 6.2b and Example 6.5, Supplementary Information of Chapter 6).

The study also provides three methodological novelties. First, the musical dimension has been simplified by excluding pitch, yielding more controlled stimuli. It is known that the melodic contour in musical tunes induces variations in perceived prominence, e.g. via so-called *melodic accents* (Thomassen 1982; Huron & Royal 1996; Müllensiefen, Pfleiderer & Frieler 2009). Our stimuli, hence, employ speech-like pitch contours instead of musical melodies, thereby eliminating a potential confound which remains under-studied within the textsetting of non-tonal languages (cf. Särg & Ambrazevičius 2007).

Second, the differences between the trials from the different conditions is minimal: it affects the alignment of a single word, keeping a constant frame. By using these localised contrasts, we can pinpoint more accurately the variables driving the different judgements provided by the subjects, unlike other procedures where alignment differences affect all the words contained in a trial (Gordon, Magne & Large 2011:4).

Third, we assume that textsetting intuitions are gradual rather than binary, as it is claimed for metrics more generally (Ryan 2011). Hence, we apply a suitable methodology to derive a ranking from the most preferred to the least preferred patterns: a two-alternative forced-choice task (Thurstone 1927). Thus, instead of asking to rate the well-formedness of individual settings, subjects are asked to choose one out of two minimally-differing trials. Though mainly used in the field of psychophysics, this approach to uncovering gradual acceptability intuitions has been successfully applied in linguistic studies too (Cowart 1997; Stadthagen-González et al. 2017).

7.2 Method

7.2.1 Participants

We used the Meertens Panel database[1] to recruit 135 participants via the internet (74 females, 61 males; mean age = 60.43). All participants are native speakers of Dutch, and most (93 %) were born and currently live in The Netherlands.

[1] The database is managed by the Meertens Institute (Amsterdam, The Netherlands). More information can be found here: www.meertens.knaw.nl/meertenspanel/.

Table 7.2: Drum pattern to which the experimental sentences were aligned. The labels B1–4 provide a reference for the four beats contained in the pattern.

	B1		B2		B3		B4	
Bass drum	●				●	●		
Snare drum			●				●	
Closed hi-hat	●	●	●	●	●	●	●	●
	*							
Relative	*				*			
prominence	*		*		*		*	
	*	*	*	*	*	*	*	*

7.2.2 Procedure

The whole experimental procedure was conducted online, so each participant performed the task in the environment of their choice. Subjects were presented a screen with instructions, followed by 36 pages with the experimental task, i.e. a two-alternative forced-choice task per page. In each of these pages, they were asked to listen to two separate audio recordings. Each recording consisted of a sentence (4-5 words) and a simultaneous drum sequence in the background. Participants were asked to select one of the two recordings based on how well the words fitted the background rhythm. Both the order of the 36 comparisons and the order of the two recordings within each comparison were randomised for each participant.

7.2.3 Stimuli

Each stimulus had the same structure: a sentence of four or five words is spoken by a female native speaker of Dutch while a metrical drum sequence is heard in the background. The drum sequence was the same in every recording, but the string of words and the word-to-drum alignment varied.

The drum sequence (Table 7.2) is expected to provide strong metrical cues to Western listeners (Bouwer, Van Zuijen & Honing 2014). The sequence consists of a concatenation of eight sound events; the time points labelled as B1, B2, B3, B4 are equally spaced, with a time-span of 666 ms between adjacent positions. Listeners are likely to perceive a sense of beat at this rate (Honing 2013), which would yield a tempo of 90 beats per minute. The pattern in Table 7.2 is repeated four times in

Table 7.3: Complete set of base sentences used to create the 24 sound stimuli.

1	*willem* willem	*lévert / bestélt* supplies / orders	*mooie kleren* nice clothes
2	*marloes* marloes	*óefent / studéert* practises / studies	*een vreemde taal* a foreign language
3	*sandra* sandra	*ántwoordt / notéert* answers / notes down	*het laatste punt* the last point
4	*jeroen* jeroen	*tékent / verbéeldt* draws / imagines	*een bonte herfst* a colourful autumn
5	*femke* femke	*schíldert / verlícht* paints / illuminates	*onze kamer* our room
6	*matthijs* matthijs	*flúistert / verzínt* whispers / makes up	*lieve woordjes* sweet words

each stimulus, with a total duration of approximately eleven seconds. We used the open-source Hydrogen drum machine with the TR707 sample-set (Cominu, Wolkstein & Moors 2015) in order to generate the drum sounds. The lower part of Table 7.2 shows the relative metrical prominence which the drum pattern is expected to elicit (Bouwer, Van Zuijen & Honing 2014). Following the method described by Lerdahl & Jackendoff (1983), a larger amount of asterisks indicate greater metrical prominence.

In each recording, the participants were presented a sentence parallel to the drum pattern; these sentences, however, were recorded independently and then aligned with the drums in a controlled way. We used a total of six base sentences (Table 7.3), all following the same structure: subject - verb - object. For each base sentence, the subject and the object were kept constant, as well as their alignment with respect to the drum pattern. Hence, the critical section of the stimuli is the verb.

All the verbs here chosen are bisyllabic, and belong to two rhythmic categories: iambic or trochaic. Iambic verbs show an increasing stress contour (first syllable is unstressed, second is stressed), while trochaic verbs show a decreasing stress contour. In our manipulation, each verb contour can be aligned to an increasing or decreasing metrical context; this 2×2 design yields the four experimental conditions listed in Table 7.4. We refer to the conditions with the following acronyms,

Table 7.4: The four experimental conditions, produced by combining two contours of metrical prominence with two contours of linguistic stress. Underlined syllables indicate a position with greater metrical prominence than non-underlined ones.

	prominence−	prominence+
stress−	1. *lé-vert*	2. *lé-vert*
stress+	3. *be-stélt*	4. *be-stélt*

Table 7.5: The four experimental conditions using one of the six frame sentences. The critical section being manipulated (i.e. the verb) has been visually framed here. The sequence of starts depict the metrical structure of the background drum sequence, the number of starts being correlated with metrical prominence.

	B1		B2	B3			B4		B1		B2	
	*								*			
	*				*				*			
	*		*		*		*		*		*	
	*	*	*	*	*	*	*	*	*	*	*	*
1					*lé-*	*vert*						
2	*wil-*	*lem*		*lé-*	*vert*				*moo-*	*ie*	*kle-*	*ren*
3					*be-*	*stélt*						
4				*be-*	*stélt*							

where *s* stands for *stress*, and *p* for *metrical prominence*: (1) `s–p–`, (2) `s–p+`, (3) `s+p–`, (4) `s+p+`. Hereby, we highlight the fact that conditions 1 and 4 represent a match between the stress and the prominence contours, whereas conditions 2 and 3 combine divergent contours.

Table 7.5 exemplifies this by depicting the four stimuli produced by manipulating the first base sentence. The relative metrical prominence elicited by the drum pattern is represented with asterisks, as explained before. The first two events of the pattern, for instance, show a decrease in metrical prominence, going from 4 to 1 asterisk. In the framed section we highlight the critical part of the trials, where the verb displays the four possible alignments of linguistic and metrical prominence contours.

A native speaker of Dutch recorded the complete set of subjects, verbs and objects separately. She was instructed to pronounce the items as in speaking (not singing), but with an isochronous timing of the syllables, unlike in everyday speech. To ensure this, she produced the items while listening to a metronome

Table 7.6: Number of times (and proportion) the condition on the row is preferred over the condition on the column.

	1. s–p–	2. s–p+	3. s+p–	4. s+p+
1. s–p–	–	594 (.73)	605 (.75)	372 (.46)
2. s–p+	216 (.27)	–	465 (.57)	229 (.28)
3. s+p–	205 (.25)	345 (.43)	–	167 (.21)
4. s+p+	438 (.54)	581 (.72)	643 (.79)	–

track set at 180 beats per minute. Later, the relevant recordings were concatenated to form sentences, and aligned with the drum pattern according to one of the four conditions.

In order to test the textsetting intuitions of the participants, we presented them the four stimuli from each base sentence in a pairwise manner. This produces six comparisons per base sentence and a total of thirty-six comparisons using all the base sentences in Table 7.3.

7.2.4 Statistical analyses

We used two different models to assess which textsetting alignments were regarded as better-formed. Each judgement made by a subject yielded a winner (i.e. a preferred setting), and the relative number of wins and losses for each stimulus is used to obtain a ranking from the most preferred to the least preferred condition.

We first applied the Thurstone-Mosteller model (Thurstone's Law of Comparative Judgement, Case V, Thurstone 1927), and further refined the analysis with the Bradley-Terry model (Bradley & Terry 1952). Thurstone's model is far more widespread, but Bradley-Terry's offers a number of advantages, like the flexibility to incorporate additional (random) predictors (Handley 2001). Both models were implemented in `R` (R Core Team 2017) with the packages `psych` (Revelle 2017) and `BradleyTerry2` (Turner & Firth 2012).

Table 7.7: Results of the Thurstone-Mosteller test. Given the pairwise results in Table 7.6, the test produces values between 0 and 1 indicating the relative preference for each condition, the least favoured option given a 0 by convention.

Pattern	Example	Condition	Thurstone value
s+p+	*be-stélt*	4	0.79
s–p–	*lé-vert*	1	0.71
s–p+	*lé-vert*	2	0.17
s+p–	*be-stélt*	3	0.00

7.3 Results

Table 7.6 presents the number of times (and proportion) the conditions on the rows are preferred over the conditions on the columns. On the second column of row 1, for instance, we observe that the first condition (e.g. *lé-vert*) is favoured over the second (e.g. *lé-vert*) in 73% of the trials.

By applying the Thurstone-Mosteller model to these data, we obtain the relative ranking of preferences shown in Table 7.7. By convention, the least preferred option receives a value of 0, and the other options receive higher values, with a theoretical maximum of 1. As predicted from the literature, participants display a sharp division between parallel contours (conditions 1 and 4), and opposing contours (conditions 2 and 3).

In addition, an iambic verb set to a trochaic pattern (*be-stélt*) is more heavily dispreferred than its mirror setting (*lé-vert*). A smaller difference is found between the two parallel settings, where iambic verbs (*be-stélt*) are preferred over trochaic ones (*lé-vert*).

In order to assess the statistical significance of these preference differences between conditions, we fit the data to a Bradley-Terry model. The most dispreferred condition (3, *be-stélt*) is taken as a baseline, and the model tests what the relative increase in preference is provided by each of the other three conditions, and the robustness of the difference. The results are displayed in Table 7.8, where we verify that the preference shown by the remaining three conditions is statistically robust ($\alpha = 0.05$). By setting condition 1 (*lé-vert*) as a baseline, we can further confirm that the difference between the two parallel conditions is also statistically significant ($z = 2.43$, $p = 0.015$).

Additionally, we controlled for the potential effect of the position of a sentence on the screen. In each page, subjects were asked to play first one sentence, then

Table 7.8: Results of the Bradley-Terry test. Each row indicates the extent to which the predictor shown in the first column increases the likelihood of a trial being preferred over the baseline condition (3).

Predictor	Estimate	Std. Error	z value	Pr ($> z$)
condition1	1.17	0.056	20.9	4.22e-97
condition2	0.276	0.0535	5.16	2.51e-07
condition4	1.3	0.0568	22.9	9.43e-116
first.displayed	0.118	0.0316	3.73	0.000193

the other, and finally make a decision about the preferred one. It is likely that subjects listened first to the sentence displayed at the left side of the screen, and that they systematically showed a tendency to prefer or dispreferr a sentence based on the display order of the stimuli. Hence, we added the predictor `first.displayed` to the model. Indeed, being shown first increases the likelihood of a sentence being preferred by a factor of 0.118 (z = 3.73, p = 0.000193, last row of Table 7.8). Still, the relative differences in preference ranking remain robust after correcting for order of display.

7.4 Discussion

Using a two-alternative forced-choice task we show that native speakers of Dutch prefer parallel rather than opposing contours. This preference for a congruent alignment is unsurprising and predicted by previous literature; yet, it provides support for the validity of the methodology employed. More interestingly, we also show that the two types of opposing pattern are not equally dispreferred. The reason why participants judged patterns like *be-stélt* as being worse than patterns like *lé-vert* may be grounded on phonological properties of Dutch, or stem from more general principles of how prominence is parsed.

First, there is a strong trochaic bias in the Dutch language (and, more generally, in Germanic), both synchronically and in acquisition (Fikkert 1994). Regarding the subset of the lexicon critical to our study, over 75 % of Dutch bisyllabic conjugated verb forms are trochaic (Baayen, Piepenbrock & van Rijn 1995). Even if the lexical items chosen as stimuli are comparable in terms of frequency of use, speakers could still be sensitive to the lexical prevalence of trochaic (i.e. `s–`) verb forms, and exhibit a `s–p+` > `s+p–` preference. Nevertheless, if that general preference for trochaic forms was in place, we should expect it to affect parallel

contours too. Still, subjects show a preference for iambic forms within the context of parallel contours (i.e. s+p+ is preferred over s–p–). The effect of a trochaic bias remains a possibility, but, given the limited number of lexical items here tested, the available evidence is inconclusive.

Second, an alternative account can be that an increase in stress within a decreasing prominence environment (s+p–) is more salient, and hence more readily dispreferred, than a decrease in stress in an increasing prominence environment (s–p+). This is somehow reflected in the Stress Maximum Constraint proposed for English poetry by Halle & Keyser (1971), although evidence from other unrelated (non-trochaic) languages would be required in order to generalise the principle. A stress maximum is defined as a stressed syllable between two unstressed syllables within a given constituent; according to the constraint, maxima are only allowed in metrically strong positions.

Under our analysis, prominence is always parsed left-to-right, in a strict temporal order, so that the relative prominence of a syllable n is defined only by the prominence of syllable $n - 1$, and not affected by the prominence of $n + 1$, as in the definition by Halle & Keyser (1971). Hence, among the stress contours of our critical bisyllabic verbs, only the second syllable of iambic verbs (*bestélt*) can be equated to a stress maximum.

According to Halle & Keyser (1971), stress maxima must occur in strong positions (p+ in our notation). That means that p– positions are more constrained, because they do not allow stress maxima (s+), but p+ positions in turn are free to receive any kind of syllable, whether s+ or s–. In our results we do observe a preference for p+ contexts to be aligned with s+ words (*be-stélt* > *lé-vert*), but this difference is smaller than the preference for s–p– (*lé-vert*) over s+p– (*be-stélt*). This suggests that, indeed, p– contours are more stringent, as argued by Halle & Keyser (1971).

Additional evidence for this view that s+p– is perceptually more salient than s–p+ comes from experiments testing the ease with which we process deviant tones in metrical contexts. Bouwer & Honing (2015) found that subjects were faster and more accurate in detecting unexpected amplitude *increments* in a metrical drum sequence, compared to unexpected *decrements* in amplitude.

In contrast with our study, this kind of approach provides higher temporal resolution, since subjects do not make judgements over a whole stimulus lasting several seconds, but react directly to deviant events. In order to address the issue of how unexpected a particular syllable is in a given metrical context, the offline preference judgements we have employed can be followed up with complementary online tasks.

On the behavioural side, phoneme monitoring can provide a measure of how disruptive different text-to-metre alignments are (Quené & Port 2005; Connine & Titone 1996; Finney, Protopapas & Eimas 1996). This bears resemblance with the lexical decision task used by Gordon, Magne & Large (2011), but with the advantage of recording the dependent variable closer to the required moment of the stimulus. Specifically, subjects can be asked to detect a target phoneme occurring at or just after the critical textsetting item (i.e. the verb in our case).

Finally, brain imaging techniques such as EEG can provide very fine-grained measures of how strongly different textsetting patterns violate the listener's expectations. To be sure, the most dispreferred patterns in our study are likely to produce noticeable mismatch negativities (Honing, Bouwer & Háden 2014; Winkler 2007). All in all, corpora remain invaluable sources to extract underrepresented patterns, and define the musical and linguistic features which can then be systematically manipulated to conduct experiments like the one here reported.

7.5 Conclusion

We have shown that participants are sensitive to four different alignments of linguistic stress and metrical prominence. This validates the two-alternative forced-choice task for textsetting purposes, and its use to complement corpus-based analyses; here, indeed, we have replicated the corpus findings described in Chapter 6. The extent to which the judgements depend on the phonological properties of the Dutch language or on general auditory cognition is to be determined by further research.

References

Baayen, R. H., R Piepenbrock & H van Rijn. 1995. *The CELEX lexical database (release 2)*. Web version (last consulted: 2016.06.24). Philadelphia: Linguistic Data Consortium.

Bouwer, F. L. & H. Honing. 2015. Temporal attending and prediction influence the perception of metrical rhythm: evidence from reaction times and ERPs. *Frontiers in psychology* 6.

Bouwer, F. L., T. L. Van Zuijen & H. Honing. 2014. Beat processing is pre-attentive for metrically simple rhythms with clear accents: an ERP study. *PloS one* 9(5). e97467.

Bradley, R. A. & M. E. Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39(3/4). 324–345.

Cominu, A., M. Wolkstein & S. Moors. 2015. *Hydrogen drum machine. Version 0.9.6.1*. Computer program. URL: <www.hydrogen-music.org>.

Connine, C. M. & D. Titone. 1996. Phoneme monitoring. *Language and Cognitive Processes* 11(6). 635–645.

Cowart, W. 1997. *Experimental syntax: applying objective methods to sentence judgments*. California: Sage publications.

Dell, F. & J. Halle. 2009. Comparing musical textsetting in French and English songs. In J.-L. Aroui & A Arleo (eds.), *Towards a typology of poetic forms*, 63–78. Benjamins.

Fikkert, P. 1994. *On the acquisition of prosodic structure*. Holland Institute of Generative Linguistics.

Finney, S. A., A. Protopapas & P. D. Eimas. 1996. Attentional allocation to syllables in American English. *Journal of Memory and Language* 35(6). 893–909.

Gordon, R. L., C. L. Magne & E. W. Large. 2011. EEG correlates of song prosody: A new look at the relationship between linguistic and musical rhythm. *Frontiers in Psychology* 2(352).

Halle, M & S.-J. Keyser. 1971. A theory of meter. In *English stress: its form, its growth, and its role in verse*, 139–180. Harper & Row.

Handley, J. C. 2001. Comparative analysis of Bradley-Terry and Thurstone-Mosteller paired comparison models for image quality assessment. In *Pics*, vol. 1, 108–112.

Hayes, B. 2009. Textsetting as constraint conflict. In J.-L. Aroui & A Arleo (eds.), *Towards a typology of poetic forms*, 43–61. Benjamins.

Hayes, B. & A. Kaun. 1996. The role of phonological phrasing in sung and chanted verse. *The Linguistic Review* 13. 243–303.

Honing, H. 2013. Structure and interpretation of rhythm in music. In D Deutsch (ed.), *The psychology of music*, 369–404. Elsevier.

Honing, H., F. L. Bouwer & G. P. Háden. 2014. Perceiving temporal regularity in music: the role of auditory event-related potentials (ERPs) in probing beat perception. In H Merchant & V de Lafuente (eds.), *Neurobiology of interval timing*, 305–323. Springer.

Huron, D & M Royal. 1996. What is melodic accent? Converging evidence from musical practice. *Music Perception* 13(4). 489–516.

Lerdahl, F & R Jackendoff. 1983. *A generative theory of tonal music*. Massachusetts Institute of Technology.

Müllensiefen, D, M Pfleiderer & K Frieler. 2009. The perception of accents in pop music melodies. *Journal of New Music Research* 38(1). 19–44.

Proto, T & F Dell. 2013. The structure of metrical patterns in tunes and in literary verse. Evidence from discrepancies between musical and linguistic rhythm in Italian songs. *Probus* 25(1). 105–138.

Proto, T. 2015. Prosody, melody and rhythm in vocal music: the problem of textsetting in a linguistic perspective. *Linguistics in the Netherlands* 32(1). 116–129.

Quené, H. & R Port. 2005. Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica* 62(1). 1–13.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Revelle, W. 2017. *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 1.7.8. Northwestern University. Evanston, Illinois.

Ryan, K. M. 2011. Gradient syllable weight and weight universals in quantitative metrics. *Phonology* 28(03). 413–454.

Särg, T & R Ambrazevičius. 2007. Melodic accent in Estonian and Lithuanian folk songs. In K Maimets-Volt, R Parncutt, M Marin & J Ross (eds.), *Proceedings of the 3rd conference on interdisciplinary musicology*. Published at: http://www-gewi.uni-graz.at/cim07. Tallinn (Estonia): Estonian Academy of Music & Theatre.

Schütze, C. T. 2011. Linguistic evidence and grammatical theory. *Wiley Interdisciplinary Reviews: Cognitive Science* 2(2). 206–221.

Schütze, C. T. 2016. *The empirical base of linguistics: grammaticality judgments and linguistic methodology*. Berlin: Language Science Press.

Stadthagen-González, H., L. López, M. C. Parafita Couto & C. A. Párraga. 2017. Using two-alternative forced choice tasks and Thurstone's law of comparative judgments for code-switching research. *Linguistic Approaches to Bilingualism*.

Thomassen, J.-M. 1982. Melodic accent: experiments and a tentative model. *Journal of the Acoustical Society of America* 71. 1596–1605.

Thurstone, L. L. 1927. A law of comparative judgment. *Psychological review* 34(4). 273–286.

Turner, H. & D. Firth. 2012. Bradley-terry models in R: the BradleyTerry2 package. *Journal of Statistical Software* 48(9). 1–21.

Winkler, I. 2007. Interpreting the mismatch negativity. *Journal of Psychophysiology* 21(3-4). 147–163.