The following paper was presented at The 9th Workshop on Disfluency in Spontaneous Speech (DiSS 2019) held at ELTE Eötvös Loránd University in Budapest, Hungary on 12–13 September, 2019.

Title: Empathetic hearers perceive repetitions as less disfluent, especially in non-broadcast situations

Author(s): Iulia Grosman, Anne Catherine Simon and Liesbeth Degand

Abstract: This experiment measures the impact of the communicative situation on perceived fluency in French speech. We consider three dimensions of fluency: grammatical, discursive and socio-interpersonal. We first hypothesise that grammatical fluency is less influenced by contextual constraints than the other two dimensions. Furthermore, taking into account the Interpersonal Reactivity Index of each participant, we hypothesise that hearers with higher interpersonal capacities will be more tolerant in their fluency evaluation, because of their ability to project into the speaker’s mind. The strength of the design rests on the proposal to test natural stimuli and integrate social and individual variables in a perception experiment.

DOI: https://doi.org/10.21862/diss-09-007-gros-etal


The complete proceedings for DiSS 2019 are available as follows.

ISBN: 978-963-489-063-8

DOI: https://doi.org/10.21862/diss-09

DiSS 2019 was sponsored by The Faculty of Humanities, ELTE Eötvös Loránd University and the International Speech Communication Association (ISCA).
Empathetic hearers perceive repetitions as less disfluent, especially in non-broadcast situations

Iulia Grosman, Anne Catherine Simon and Liesbeth Degand
Institute for Language and Communication, University of Louvain, Louvain-la-Neuve, Belgium

Abstract
This experiment measures the impact of the communicative situation on perceived fluency in French speech. We consider three dimensions of fluency: grammatical, discursive and socio-interpersonal. We first hypothesise that grammatical fluency is less influenced by contextual constraints than the other two dimensions. Furthermore, taking into account the Interpersonal Reactivity Index of each participant, we hypothesise that hearers with higher interpersonal capacities will be more tolerant in their fluency evaluation, because of their ability to project into the speaker’s mind. The strength of the design rests on the proposal to test natural stimuli and integrate social and individual variables in a perception experiment.

Introduction
Evaluation of fluency
This study evaluates three types of fluency: the grammatical dimension of fluency, the discourse-level one and the socio-interpersonal one. This tri-dimensional model of fluency, proposed in Grosman (2018), is an expansion of the socio-cognitive framework on saliency proposed by Schmid and Günther (2016).

In this model, a stretch of speech is evaluated as being disfluent when the (grammatical, discursive, socio-interpersonal) discourse expectations of the hearer are over-confirmed or over-deceived. This evaluation depends on the degree of convergence of the discourse with the hearers’ expectations.

The first hypothesis is that grammatical fluency is less dependent on contextual constraints than discursive and socio-interpersonal fluency. The second hypothesis is that the higher the ability to project into a speaker’s mind, the more lenient hearers are in their evaluation.

We modulate the production of repetitions to evaluate the perceived dimension of fluency. We focus solely on identical repetitions structure that interrupts the syntactic, grammatical speech flow, and might cause socio-cognitive dis/fluency. Four conditions for repeated segments were considered: contiguous, with a filled pause, with a silent pause and no repetition.

Method and design
For each participant, the collected data comprises the results of a pretest (Interpersonal Reactivity Index questionnaire) and the degree of agreement with 6 assertions related to 40 speech stimuli. Furthermore, the experiment tests the impact of the broadcasting orientation on fluency evaluation and on the different dimensions of fluency.

Participants
The 202 native francophone participants were allowed to use hi-fi speakers or headphones (43% and 57% respectively). Their age ranged from 19 to 72 ($\mu = 30.62, \sigma = 11.62$) and 69% of them were women. Almost half of the participants were bilingual (47%) and only a quarter declared themselves as monolingual or trilingual (26% and 23% respectively).

Interpersonal Reactivity Index (F-IRI)
The participants completed the F-IRI test (Davis, 1983; Gilet et al., 2013). Through a 28-item self-report, the index evaluates aspects of empathy: the capacity to put oneself in someone else’s shoes, to imagine his/her thoughts and feelings.

The test is structured around four subscales: (1) Perspective Taking: tendency to adopt the point of view of other people in everyday life; (2) Fantasy: tendency to transpose oneself into the feelings and actions of fictitious characters in books, movies, and plays; (3) Empathic Concern: tendency to experience feelings of warmth, compassion and concern for other people; (4) Personal Distress: assessment of typical emotional reaction of personal unease and discomfort in reaction to the emotions of others.

The mean score of F-IRI is 95.6 ($\sigma = 12.3$, max = 123, min = 60). The subscale Empathic Concern reaches the highest mean score ($\mu = 26, \sigma = 4.7$). It is followed by the subscales fantasy ($\mu = 25.9, \sigma = 5.2$), Perspective Taking ($\mu = 24.5, \sigma = 4.6$) and Personal Distress ($\mu = 19.24, \sigma = 5.4$). We categorised the participants’ scores into 5 quartiles for each subscale (e.g. lowest, low, average, high and highest scores).

Audio stimuli selection and processing
The experiment tests the impact of repetitions on perceived fluency (see examples below). We consider four conditions for repeated segments: contiguous,
with a filled pause, with a silent pause and no repetition. The 40 natural stimuli were extracted from C-HUMOUR (Grosman, 2016), LOCAS-F (Degand, Martin & Simon, 2014), C-Phonogenre (Goldman, Prsir & Auchlin, 2014), Rhapsodie (Kahane & Peitrandrea, 2019) and DRIVE (Christodoulides, 2016).

We selected semi-automatically two sets of original speech samples: 20 original items including a contiguous identical repetition (REP-C) and 20 original items including an identical repetition with a filled pause in the editing phase (REP-FP). We transformed both sets respectively into 20 modified repetition-free items, by removing the first segment (NO-REP), and 20 modified repetition items with a silent pause (REP-UP). A participant listening to a speech sample in the REP-C or REP-FP condition would never hear its counterpart in the NO-REP or REP-UP condition, and vice versa.

Examples:
- REP-C to No-REP: ...la ville est assez haute et {(la) la} Loire est en bas... …the city is quite high and {the} the Loire is below...\ (1521-2-2NM-Rhap-D0003)
- REP-FP to REP-UP: ...ces nobles vont en général chanter entre guillemets {leur (euh)|(//) leur} nostalgie... …the nobles will generally sing {their (uh)|(//) their} nostalgia...\ (386-1-2NM-conv-f-2)

Based on the most frequent occurrences in the distribution of repetitions in the corpus data (Grosman et al., 2017), we selected monosyllabic, single repetitions of grammatical determiners. We excluded repetitions occurring at the very beginning or very end of a sample, avoiding primacy or recency effect during evaluation. Speech sample duration varies from 7 to 19 sec. (µ = 11.3, σ = 3.1).

Impact of broadcasting orientation

Each stimulus was contextualised by an image indicating the communicative situation: broadcast oriented or non-broadcast-oriented speech (see Figure 1). The speech samples were either presented in their original production setting or in a pretended one. Each participant evaluated 40 stimuli, 10 stimuli by fluency condition, with half of them presented in their original context of production and the other half in the pretended modified context.

Each stimulus was paired with a short nominal sentence of contextualisation (e.g. Writing a book on Jimmy Hendricks, Choosing a job, Spider reproduction). The experiment fosters hearers to focus on the speech form rather than on its content.

Tridimensional evaluation of fluency

For each of the 40 stimuli listened to, participants expressed their degree of agreement with 6 assertions (40 stimuli × 6 assertions = 240 answers per participant). Each pair of assertions corresponded to an evaluated dimension of fluency described above:

- Grammatical:
  - A1. The sentence is well-formed.
  - A2. The sentence includes hesitations.
- Discourse-level:
  - A3. The speech is fluid.
  - A4. The speech is nice to listen to.
- Socio-interpersonal:
  - A5. The speech in this context appeals to me.
  - A6. The speech used in this context is improper.

For the evaluation of the results, the scale of A2 and A6 assertions were inverted as they were negatively formulated. The distance between each ordinal point was considered equidistant as there was a significant information value added compared to a model with symmetric thresholds (around the central one, here “average”) (LR_{test} = 535.25, p < 0.001).

Hypothesis

In this framework, disfluency corresponds to an oversaliency by converging or diverging expectations of the hearer regarding context. Depending on the context considered, distinct (dis)fluency evaluations emerge from different types of expectations. Regarding fluency evaluation, the main hypothesis is that the evaluation of grammatical fluency is less conditioned by contextual constraints than the evaluation of discursive fluency. Some specific hypotheses (A–C) were also tested.

A. Hypothesis on Internal Structure of the Repetition: The internal structure of a repetition influences its evaluation, independently of the discursive context. More precisely, non-contiguous repetitions with an editing term (the uh the) are evaluated less leniently than contiguous repetition (the the). This hypothesis supports the following continuum: (fluent) > no repetition > repetition with filled pause > repetition with unfilled pause > contiguous repetition > (disfluent).

The orientation of this hypothesis is motivated in twofold way. Results observed in production reveal
the relative importance of repetition involving a silent pause, even in prepared discourse (Grosman, 2018: 252–265). Results observed in perception reveal the ease of processing and the saliency accordingly for each structure (Grosman, 2015).

B. Hypothesis on Broadcast Speech: The stimuli presented in a formal broadcast-oriented speech context are rated more severely than those in non-broadcast-oriented speech. Additionally, a same repetition in one context of production does not necessarily induce the same evaluation for every hearer. We do not formulate any hypothesis linking the type of repetition and the context of production (i.e. one structure is not more accepted than another in a certain context), as this latter feature was non-relevant in our corpus data (Grosman, 2015: 252–265).

C. Hypothesis of F-IRI: The higher the interpersonal capacities of the hearers are, the more lenient they may be in their evaluation, due to the ability to project into the speaker’s mind. Furthermore, we think the F-IRI is more relevant when evaluating discursive or socio-interpersonal fluency than grammatical fluency.

Finally, the perception and evaluation of disfluency depend on one’s ability to anticipate content. The better the subject’s projection capacity is, the more s/he can process (dis)fluency efficiently. Hence, subjects getting high or very high scores in the dimension of Empathic Concern, Perspective Taking and Imagination would rate fluency less harshly.

Main Results

Importance of stimuli, participant and thresholds

The results were processed with an ordinal mixed model with 2 random effects recognizing the fact that (1) the same audio stimulus was evaluated according to 6 fluency measures and (2) the same participant evaluated 40 stimuli. The compared results between the ordinal model and the ordinal mixed model showed that 22.44% of the variance is explained by these random effects (Pseudo-$R^2 = 0.224, p < 0.001$).

Internal structure of the repetition

The cumulative mixed effect model reveals that the overall degree of perceived fluency tends to lower when a repetition appears, no matter the internal structure of the repetition (Table 1). A stimulus with a repetition has 88% probability to be evaluated as more disfluent that the repetition-free ones.

Results show that the force of the devaluation is stronger when evaluating the socio-interpersonal dimension of fluency ($OR = 0.97, p < 0.001$) than for the discursive or grammatical dimensions (respectively $OR = 0.86$ and $OR = 0.88$ at $p < 0.001$). Hence, a speech sample including a repetition was not considered as including more hesitation or being inappropriate but rather as less pleasant to listen to and less fluid.

Table 1. Fixed effect results on the fluency score depending on the type of repetition (NO-REP as intercept).

<table>
<thead>
<tr>
<th></th>
<th>Estimates</th>
<th>OR</th>
<th>ES</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REP-C</td>
<td>−0.126</td>
<td>0.89</td>
<td>0.024</td>
<td>−5.29</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>REP-FP</td>
<td>−0.166</td>
<td>0.848</td>
<td>0.218</td>
<td>−0.76</td>
<td>0.446</td>
</tr>
<tr>
<td>REP-UP</td>
<td>−0.138</td>
<td>0.872</td>
<td>0.218</td>
<td>−0.63</td>
<td>0.526</td>
</tr>
</tbody>
</table>

Impact of broadcast speech situation

Overall, there is a significant difference in the appreciation of perceived fluency depending on the situation in which the speech is produced ($est. = 0.33, OR = 1.03, p < 0.05$). The degree of perceived fluency increases in broadcast situation. The participants rate less severely non-broadcast situations than broadcast ones. Nevertheless, no interaction was found with any dimension of fluency, the supposed situation impacting every dimension equally.

Impact of interpersonal reactivity

Results show that F-IRI scores do not have an impact on overall fluency perception ($est. = 0.006, p < 0.15$), but that it does influence each dimension of fluency separately. In fact, the higher the F-IRI scores of participants are, the higher they evaluate the samples on the discursive and socio-interpersonal dimensions of fluency (respectively $est. = 0.39, p < 0.03$ and $est. = 0.27, p < 0.005$), but not on the grammatical one ($est. = −0.04, p < 0.7$).

In addition, each subscale of the F-IRI test does have an impact on overall fluency evaluation and on its different dimensions. Overall, the more participants show empathic concern, the more they rate items as fluent ($est. = 0.33, p < 0.03$). Similarly, the more one shows personal distress, the more one judges the items as fluent ($est. = 0.29, p < 0.02$).

The evaluation of grammatical fluency is mainly influenced by the participant’s imagination (Fantasy) ($est. = −0.34, p < 0.03$). The more participants have imagination, the more fluent they judge the grammatical aspect of a speech sample.

The evaluation of discourse-level fluency is mainly influenced by the empathetic concern of a participant ($est. = 0.41, p < 0.01$). Confirming our hypothesis, the more participants show empathetic concerns, the more they rate a speech sample in its context as fluent.

The evaluation of socio-interpersonal fluency is mainly impacted by the participants’ personal distress
(est. = 0.38, \(p < 0.02\)). The more one has personal distress capacity, the more one evaluates a speech sample as fluid and enjoyable to listen to.

Discussion and conclusion

This study unveils three main results:

Regarding the first hypothesis, the internal structure of repetitions does not impact the overall perceived fluency. The presence of a repetition favours by itself the perception of a speech sample as less fluent, especially when evaluating the socio-interpersonal dimension (more than the grammatical or discourse-level ones). A speech sample is not perceived as including more hesitation or being inappropriate, but rather as less pleasant to listen to and less fluid.

Regarding the second hypothesis, participants overall rate less severely repetitions in non-broadcast situations than in broadcast ones, supporting the idea that the type of speech evaluated by the participants matters greatly.

Regarding the third hypothesis, the F-IRI test shows that the higher the scores of participants are, the more lenient they are to rate items as fluent on the discursive and socio-interpersonal dimensions. The overall F-IRI does not have an impact on the grammatical evaluation of fluency, but each subscale of the F-IRI brings refinements to the analysis. The grammatical fluency is particularly impacted by the participants’ imagination, while the discursive fluency evaluation is mainly influenced by their empathetic concern, while the socio-interpersonal fluency is impacted by their level of personal distress.

This paper broadly questions (dis)fluency perception and evaluation: the type of discourse that is under evaluation, the empathetic participants profile and their personal discourse expectations, as well as the type of fluency evaluated. It reveals that certain types of evaluation (e.g. interpersonal fluency) require to consider individual profiles more than others (e.g. grammatical fluency).

In this sense, it addresses several challenges: (1) the benefits of coupling production and perception data for an integrated study of repetitions; (2) the advantage of using natural aligned speech data, selected on the basis of corpus analysis, for perceptive experiments; (3) the importance of integrating individual traits into fluency evaluation; (4) the impact of the speech situation on fluency evaluation.

To address these challenges, we proposed an integrated model of fluency according to which the evaluation of fluency depends on the degree of convergence/divergence between the discourse and the hearer expectations.

Acknowledgements

This research benefits from the support of the ARC-project “A Multi-Modal Approach to Fluency and Disfluency Markers” granted by the Fédération Wallonie-Bruxelles (grant nr. 12/17-044). We thank S. Roekhaut, L. Rousier-Vercruyssen and G. Christodoulides for technical support.

References


26