



Prediction of L2 speech proficiency based on multi-level linguistic features

Verdiana De Fino^{1,2}, Lionel Fontan², Julien Pinquier¹, Isabelle Ferrané¹, Sylvain Detey³

¹IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France

²Archean Labs, Montauban, France

³SILS & GSICCS, Waseda University, Tokyo, Japan

verdiana.defino@irit.fr, lfontan@archean.tech, julien.pinquier@irit.fr,
isabelle.ferrane@irit.fr, detey@waseda.jp

Abstract

This study investigates the possibility to use automatic, multi-level features for the prediction of L2 speech proficiency. The method was applied on a corpus containing audio recordings and transcripts for 38 Japanese learners of French who participated in a semi-spontaneous oral production task. Each learner's speech proficiency level was assessed by three experienced French teachers. Audio recordings were processed to extract features related to the pronunciation skills and phonetic fluency of the learners, while the transcripts were used to measure their lexical, syntactic, and discursive abilities in French.

A Lasso regression using a leave-one-out cross-validation procedure was used to select relevant features and to accurately predict speech proficiency scores. The results show that five features related to the phonetic fluency (speech rate), lexical abilities (lexical density), discourse planning and elaboration skills (number of hesitation and false starts, mean utterance length) of the learners can be used to predict speech proficiency ratings ($r = 0.71$, mean absolute error on a 5-point scale: 0.53).

Index Terms: automatic assessment, non-native speech, semi-spontaneous speech, linguistic levels, prediction

1. Introduction

Assessment is a necessary component of the learning, teaching and certification process. In foreign language learning, all learners, as well as their teachers, could therefore benefit from objective, automatic tools to assess oral production, in particular with regards to the widely used descriptors of the Common European Framework of Reference for Languages (CEFR) [1]. The difficulties of assessing oral production differ from the difficulties of assessing written production, in that the former also relies on the phonetic-phonological level. However, most studies addressing the automatic evaluation of L2 oral skills focus on a single linguistic level, such as pronunciation [2], fluency [3] or lexical richness [4], and uses relatively constrained production tasks, such as reading out loud [5] or repetition [2].

This study aims at predicting the speech proficiency of L2 learners of French who were recorded during a semi-spontaneous speech task. This prediction is based on a set of automatic, multi-level linguistic measures. Phonetic-phonological skills, such as segmental pronunciation and phonetic fluency, are assessed based on the audio signal, while lexical, syntactic and discourse abilities are assessed based on the orthographic transcripts of the recordings.

This paper is organized as follows. Section 2 presents a state-of-the-art of the automatic measures used to assess L2 learners' oral and written skills, at each linguistic level. Section 3 describes the set of multi-level features used in the present study for the prediction of speech proficiency. Sec-

tion 4 presents the experimental context on Japanese learners of French while section 5 is dedicated to the prediction of the learners' speech proficiency. Results are discussed in Section 6.

2. Automatic assessment of non-native productions

Currently, many tools are available to process spoken or written productions and extract features at different linguistic levels. Segmental and suprasegmental levels can contribute to assess learners' pronunciation, while lexical, syntactic and discursive levels can help to analyse speech transcripts and assess each learner's linguistic proficiency.

Regarding pronunciation, two aspects may be considered: extracting phonemic and prosodic features from audio recordings on one side, and using an automatic speech recognition (ASR) system on the other side. ASR systems are based on free or forced alignment techniques, to extract phoneme recognition scores [6].

Regarding fluency, two approaches can also be considered. In the context of foreign languages, fluency is defined as "the degree to which speech flows easily without pauses and other disfluency markers" [7, p. 5]. One approach may consist in using speech recognition as such tools also provide temporal features, such as speech rate and average pause length [3]. Since the use of such systems has limitations, as they are dependent on the language for which the models have been trained, new methods have recently been developed to measure fluency more automatically and independently of the target language. For example, the algorithm presented by [8] can be mentioned, resulting from pilot work on the automatic assessment of phonetic fluency of Japanese learners of French in reading task [5, 9]. This work relies on the forward-backward divergence segmentation method [10] based on the detection of breaks in the energy trajectory of the speech signal over time and allows, in addition, to compute variables from pseudo-syllables [11] and silent pauses, such as speech rate or percentage of speech.

Regarding learners' lexical ability, the dimension referred to as "richness" is generally used for assessments [12, 13]. Lexical richness is divided into three sub-parts, commonly referred to as *lexical diversity*, *lexical density* and *lexical sophistication*. Lexical diversity captures the size of the vocabulary in a text or statement, defining the number of different words produced by a speaker. Two widely used methods for measuring lexical diversity are the Guiraud index [14] and the Type-Token Ratio [15]. Lexical density, unlike lexical diversity, focuses essentially on the production of lexical words, and calculates their proportion in an utterance. A lexical word, or full-meaning word, is a verb, a noun, an adjective or an adverb. On the same principle, lexical sophistication is defined by the number of lexical words pro-

duced by a speaker that reflect a more advanced knowledge or practice of the language through the use of less frequent words (*racing car* or *tourism car* instead of *car* for example). The notion of word frequency in the language, and particularly in speech, is therefore an essential information for this measure. Moreover, there is a difference in the use of specialized or rare words according to the certified levels of non-native speakers. Lexical databases designed for specific languages usually integrate frequency and lexical information. In this work on French as a target language (L2), the information collected in the Lexique3 database [16] are used. This database represents more than 135,000 lexical entries, except proper names.

The term syntactic analysis, on the other hand, often refers to the analysis of the syntactic complexity of an utterance. The most common unit on which syntactic complexity is objectively measured is the sentence, even if the reality of oral syntax is more complex to deal with than that of written syntax [17]. One way to measure it is to count the average number of words per sentence, the average number of coordinated and subordinated clauses, and the mean depth of syntactic trees [18, 19].

Finally, assessing discourse structure in language learning is often a question of assessing discourse cohesion. The use of linking elements, or connectors (e.g. *then*, *moreover*), contributes to the structuring and articulation of discourse, and allows the propositions that make it up to be linked [20, p. 171].

3. Multi-level linguistic features

The automatic measures described below are obtained from the audio signals and transcripts of the learners' oral productions.

3.1. Phonetic-phonological assessment

3.1.1. Assessment of pronunciation at the segmental level

The assessment of French learners' segmental production is performed using an automatic speech recognition (ASR) system. The acoustic models of this system have been trained on 340h of data from different French audio corpora (ESTER1, ESTER2, EPAC, BREF and French Librivox), a data augmentation phase having also been performed (rhythm perturbation, addition of chattering noise and Time Domain SpecAugment) [21]. This system was used via the Paty platform¹ (*Plateforme de Parole Atypique* [Atypical Speech Platform]). This ASR system can be used to recognize French phones or words, and outputs not only a text corresponding to the units that were recognized, but also a confidence index in the form of a probability. In the present study, the Paty ASR system was used to compute confidence indexes for all the words pronounced by each learner. An average confidence index was then processed for each learner.

3.1.2. Assessment of phonetic fluency

The algorithm described in [8] was used to assess phonetic fluency. From the boundaries of the detected audio segments and their energy, the algorithm identifies pseudo-syllables and silent pauses. Based on these units, and for each of the learners' oral productions, four features were computed. The first three features are functions of the total duration of the processed audio file. These are the speech rate (number of pseudo-syllables per second), the percentage of speech, the standard deviation of the duration of pseudo-syllables and the normalized number of silent pauses.

¹<https://paty.irit.fr/demo/>

3.2. Assessment of lexical richness, syntactic complexity and discourse cohesion and planning

3.2.1. Lexical richness

Lexical diversity: Guiraud index was implemented to measure the lexical diversity of each learner's oral productions, as it is known to be more stable than the Type-Token Ratio when faced to the varying length of the utterances [22]. Its calculation corresponds to the formula:

$$\text{Lexical diversity} = \frac{V}{\sqrt{N_w}} \quad (1)$$

with V the number of distinct words (both lexical and grammatical words) used, and N_w the total number of words.

Lexical density: the percentage of lexical words produced was computed according to the formula:

$$\text{Lexical density} = \frac{V_{lw}}{N_w} \times 100 \quad (2)$$

with V_{lw} the number of distinct lexical words.

Lexical Sophistication: lexical sophistication was measured based on the word frequencies in oral French, as provided by the Lexique3 database [16]. The Lexique3 database provides the frequencies of French word lemmas and forms, as measured from written (books) and oral (movie subtitles) corpora. In the present study, the frequencies of the words correspond to the frequencies of their lemmas in movie subtitles, as specified by Lexique3. The proportion of lexical words whose lemma frequency is lower than 10 per million, which is considered as rare in the Lexique3² database, was computed according to the formula:

$$\text{Lexical sophistication} = \frac{V_{10}}{N_w} \times 100 \quad (3)$$

with V_{10} the number of lexical words with a lemma frequency lower than 10 in Lexique3.

3.2.2. Syntactic complexity

Two measures of syntactic complexity were used in the study. The first one corresponds to the mean length of the learners' oral productions, in words. The second measure is the mean depth of the syntactic trees of the learners' oral productions. For each learner, the mean length of the oral production was measured with the formula:

$$\text{Mean word length} = \frac{1}{N_u} \sum_{i=0}^{N_u} \text{len}(U_i) \quad (4)$$

with $\text{len}(U_i)$ the word length of the utterance i , and N_u the total number of utterances in the processed oral production. Disfluencies, such as hesitation words, were also included.

The same was done for the mean depth of syntactic trees, which was measured with the formula:

$$\text{Mean depth} = \frac{1}{N_u} \sum_{i=0}^{N_u} \text{depth}(T_i) \quad (5)$$

with $\text{depth}(T_i)$ the mean depth of the syntactic tree of the utterance i . Note that the mean depth of a tree was computed by first summing the depths of the leaves, then dividing this sum by the number of leaves in the tree.

²<https://groups.google.com/g/lexiqueorg/c/C2fJ6JLQPK8/m/ydKYm2E9BAAJ>

3.2.3. Discourse cohesion and planning

The assessment of discourse cohesion was performed through the number of discourse connectors used by each learner. With the LEXCONN list [23], providing 431 discourse connectors for French, a proportion of discourse connectors used by each learner was computed. To get closer to the criteria used in the CEFRL framework concerning discourse cohesion, the diversity of connectors was also considered, by adapting the lexical diversity formula described in equation (1).

Concerning discourse planning, disfluencies such as hesitation words (for example, “*eah*” or “*mh*” for the French language) and false starts (or unfinished words) were counted for each learner. For each learner, two disfluency measures were computed: the ratio of hesitations and the ratio of false starts.

4. Application to the CLIJAF corpus

4.1. Corpus description

The method described in the previous section was applied to the CLIJAF corpus, collected within the general methodological framework of the IPFC project [24, 25]. Only a subset of the CLIJAF corpus was used, corresponding to the recording of Japanese learners of French during a semi-directive interview in French (answering to questions about their personal background and linguistic and cultural experiences), conducted by a native speaker. More precisely, in the present study, the answers to four questions were used:

“*How old are you and what is your nationality?*”, “*Which languages do you speak?*”, “*What are your biggest difficulties when learning French?*”, and “*What are the main cultural or social differences between France and Japan?*”.

This subset represents about 1 hour and 20 minutes of audio recorded from 38 Japanese students learning French (30 female), from four different Japanese universities. The data were recorded either in a recording studio or in quiet classrooms. The CLIJAF corpus also provides manual orthographic transcripts of the speakers’ oral productions.

For each learner, four recordings (and their associated transcripts) were used, that is, one per question. For the assessment of all linguistic skills but syntax, for each learner, the automatic measures were computed across questions (i.e., one measure for the whole set of four answers). As syntactic skills are usually measured at the sentence level, the answer to each of the four questions was considered independently for computing measures of syntactic complexity.

4.2. Human assessment of speech proficiency

Three experienced French as a Foreign Language (FFL) teachers (one female) participated in the assessment task. All three teachers were official assessors for the DELF and DALF examinations [1], which are designed to evaluate both written and oral proficiency in beginner-to-fully-independent FFL learners. They also had a FFL teaching experience in Japan. The four answers of the 38 students were concatenated, resulting in only one recording per student. The 38 recordings were presented once to each FFL teacher, in random order, using the Prodigy software³ (ExplosionAI GmbH, Berlin, Germany) version 1.11.7. The FFL teachers were asked to listen to the stimuli, and to assess the speech proficiency by attributing a single CEFRL level to each recording.

³<https://prodi.gy/>

Each student was therefore assigned three CEFRL levels (one per teacher). For computational purposes, the CEFRL levels were translated into a discrete numeric scale ranging from 1 (corresponding to the A1 level) to 6 (corresponding to the C2 level). To check the inter-rater agreement, Spearman correlations were computed for each pair of teachers.

Table 1 presents the Spearman correlations for each pair of teachers. As the correlation coefficients were strong (all $\rho \geq 0.70$), indicating a strong inter-rater agreement, for each learner speech proficiency scores were averaged across teachers.

Table 1: Spearman correlation coefficient (ρ) for each pair of teachers, and associated p -value.

Pair of teachers	ρ	p -value
1, 2	0.78	<.001
1, 3	0.79	<.001
2, 3	0.70	<.001

5. Predicting speech proficiency

For each Japanese student, 14 features were computed: lexical diversity, lexical density, lexical sophistication, mean answer word length, mean depth of syntactic trees, discourse connectors, diversity of connectors, hesitation words, false starts, speech rate, percentage of speech, standard deviation of the duration of pseudo-syllables, normalized number of silent pauses and word-level confidence index.

As this number of features is rather important for such a small dataset, a Lasso regression was used to determine the features that might contribute the most to the prediction of speech proficiency. As the dataset is too limited to use a separate validation set, the α regularization hyperparameter of the Lasso regression was tuned using a leave-one-out nested cross-validation procedure. α varied from 0.001 to 1. The α value yielding the lowest Mean Absolute Error (MAE) was 0.1.

Five features were selected by the Lasso regression (Table 2).

Table 2: Standardized (Std.) coefficients associated to the five features selected by the Lasso regression, sorted in descending order of absolute value.

Feature	Std. coefficient
Speech rate	0.160
Proportion of false starts	-0.108
Proportion of hesitation words	0.037
Mean length of answers (words)	0.019
Lexical density	-0.007

Figure 1 shows the result of the linear regression performed to predict speech proficiency. In this figure, each student is represented by a dot, and the regression line between the predictions and the ground truth was computed. The result of the prediction was satisfactory enough, as a $r = 0.71$ Pearson correlation as well as a MAE of 0.53 were obtained between the predicted speech proficiency and the ground truth. Moreover, a strong Pearson correlation ($r = 0.73$) was obtained between the predicted speech proficiency and the scores given by Teacher 3, as shown in Table 3.

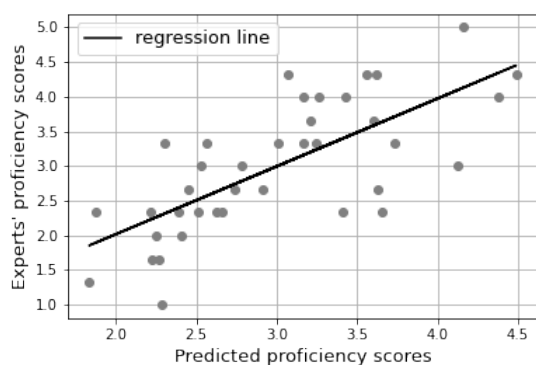


Figure 1: Scatterplot relating the average speech proficiency scores provided by the teachers to the predicted proficiency scores achieved by the Lasso regression.

Table 3: Pearson correlation coefficient (r) between actual and predicted speech proficiency scores as a function of teacher, and associated p -value.

Teacher	r	p -value
1	0.60	< .001
2	0.59	< .001
3	0.73	< .001

As shown in Table 2, the strongest coefficient is found for speech rate, indicating that this feature contributes the most for the prediction of the speech proficiency of the Japanese learners of French. Figure 2 illustrates the relationship between speech proficiency and speech rates⁴. It is interesting to note that the higher the speech rate, the highest the proficiency scores given by the teachers.

6. Discussion and conclusion

The results of this study demonstrate that French L2 speech proficiency, as evaluated by language teachers, can be predicted by using automatic measures of the learners' skills at different linguistic levels: measures based on the pronunciation (speech rate) and lexical (lexical density) abilities of the learners, and measures based on their ability to plan their discourse (number of hesitations and false starts) and to elaborate their answers (mean length of answer). The selection of these five features was automatically done by a Lasso regression as a function on their predictive power. It is possible that some other features, also relevant for the prediction of speech proficiency, were discarded by the Lasso regression because they were already correlated with one of the five best features (e.g., mean length of answer and mean depth of syntactic trees). The use of five automatic features to predict the speech proficiency of Japanese learners of French is highly successful at the scale of this study, with a 0.53 MAE and a $r = 0.71$ Pearson correlation coefficient. The high correlation ($r = 0.73$) between the results and the scores given by Teacher 3 shows that the automatic prediction could also be considered as a fourth rater. The speech proficiency prediction results are even more interesting when

⁴Scatterplots for the 13 other features can be accessed at: https://github.com/vdefino31/linguistic_features

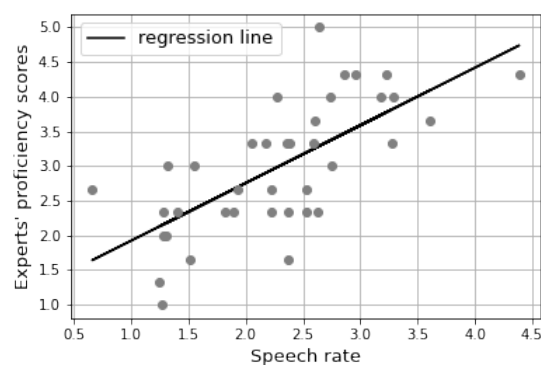


Figure 2: Scatterplot relating average speech proficiency scores to speech rate measures ($r=0.69$, $MAE=0.78$).

considering that this method based on level features could be adapted for other L1-L2 language pairs.

In future studies, it might be interesting to consider other constraints for the pre-selection of relevant features, such as pedagogical constraints (e.g., having at least one feature per linguistic level, which could be useful for teachers). Also, as some of the features used in the present study were based on a manual transcript of the learners' oral productions, future work is warranted to check if this manual step could be replaced by the use of an automatic speech recognition system.

7. Acknowledgements

This study is part of a joint research project between the ALAIA⁵ (*Foreign Language Learning Assisted by Artificial Intelligence*) laboratory and the project "From corpus to target data as steps for automatic assessment of L2 speech: L2 French phonological lexicon of Japanese learners"⁶, concerning the automatic assessment of the oral production of Japanese learners of French, partly relying on the use of the CLIJAF corpus⁷.

The Paty platform used for the assessment of segmental pronunciation skills was developed in partnership by the IRIT and LPL research labs, Toulouse and Aix-en-Provence (France).

8. References

- [1] Conseil De l'Europe, *Cadre européen commun de référence pour les langues: apprendre, enseigner, évaluer (CECR)*. Paris: Didier, 2001.
- [2] V. Laborde, T. Pellegrini, L. Fontan, J. Mauclair, H. Sahraoui, and J. Farinas, "Pronunciation assessment of japanese learners of french with gop scores and phonetic information," in *Annual conference Interspeech (INTERSPEECH 2016)*. San Francisco, CA, US: International Speech Communication Association (ISCA), 2016, pp. 2686–2690. [Online]. Available: <https://oatao.univ-toulouse.fr/17159/>
- [3] C. Cucchiari, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *The Journal of the Acoustical Society of America*, vol. 107, pp. 989–99, 03 2000.

⁵ANR-18-LCV3-0001-01, <https://www.irit.fr/SAMOVA/site/projects/current/labcom-alaia/>

⁶Detey, S. (dir.) (2020-2024). JSPS: Grant-in-Aid for Scientific Research (B) 20H01291.

⁷Detey, S. (dir.) (2011-2018). JSPS: Grant-in-Aid for Scientific Research (B) 23320121 & 15H03227

- [4] C. Lindqvist, C. Bardel, and A. Gudmundson, "Lexical richness in the advanced learner's oral production of french and italian L2," vol. 49, no. 3, pp. 221–240, 2011. [Online]. Available: <https://doi.org/10.1515/iral.2011.013>
- [5] L. Fontan, M. Le Coz, and S. Detey, "Automatically Measuring L2 Speech Fluency without the Need of ASR: A Proof-of-concept Study with Japanese Learners of French," in *Proc. Interspeech 2018*, 2018, pp. 2544–2548.
- [6] S. Detey, L. Fontan, and T. Pellegrini, "Traitement de la prononciation en langue étrangère: approches didactiques, méthodes automatiques et enjeux pour l'apprentissage," in *Revue TAL*, vol. 57, no. 3, 2016, pp. 15–39.
- [7] T. M. Derwing and M. J. Munro, *Pronunciation Fundamentals: Evidence-based perspectives for L2 teaching and research*. Amsterdam: John Benjamins, 2015.
- [8] L. Fontan, M. Le Coz, and C. Alazard, "Using the forward-backward divergence segmentation algorithm and a neural network to predict L2 speech fluency," in *Proc. 10th International Conference on Speech Prosody 2020*, 2020, pp. 925–929.
- [9] S. Detey, L. Fontan, M. Le Coz, and S. Jmel, "Computer-assisted assessment of phonetic fluency in a second language: a longitudinal study of Japanese learners of French," *Speech Communication*, vol. 125, pp. 69–79, 2020.
- [10] R. André-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 1, pp. 29–40, 1988.
- [11] J. Farinas and F. Pellegrino, "Automatic rhythm modeling for language identification," in *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, 2001, pp. 2539–2542.
- [12] B. Laufer and P. Nation, "Vocabulary size and use: Lexical richness in L2 written production," *Applied linguistics*, vol. 16, no. 3, pp. 307–322, 1995.
- [13] A. Bonvin and A. Lambelet, "Exploration empirique de la richesse lexicale: la perception humaine," *Linguistik Online*, vol. 100, no. 77, p. 65–94, Dec 2019.
- [14] P. Guiraud, *Problèmes et méthodes de la statistique linguistique*. Dordrecht: D. Reidel, 1959, vol. 2.
- [15] M. Daller, R. Hout, and J. Treffers-Daller, "Lexical richness in the spontaneous speech of bilinguals," *Applied Linguistics*, vol. 24, pp. 197–222, 06 2003.
- [16] B. New, C. Pallier, and L. Ferrand, "Manuel de Lexique 3," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 3, pp. 516–524, 2005.
- [17] N. Rossi-Gensane, "Oralité, syntaxe et discours," in *Les variétés du français parlé dans l'espace francophone : ressources pour l'enseignement*, S. Detey, J. Durand, B. Laks, and C. Lyche, Eds. Paris: Ophrys, 2010, pp. 83–106. [Online]. Available: <https://halshs.archives-ouvertes.fr/halshs-01684333>
- [18] A. C. Lahuerta Martínez, "Analysis of syntactic complexity in secondary education EFL writers at different proficiency levels," *Assessing Writing*, vol. 35, pp. 1–11, 2018.
- [19] P. Blache, "Un modèle de caractérisation de la complexité syntaxique," in *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles*. Montréal, Canada: ATALA, Jul. 2010, pp. 81–90. [Online]. Available: <https://aclanthology.org/2010.jeptalnrecital-long.9>
- [20] J. C. Beacco, S. Bouquet, and R. Porquier, *Niveau B2 pour le français*. Paris: Didier, 2004.
- [21] A. Heba, "Reconnaissance automatique de la parole à large vocabulaire : des approches hybrides aux approches End-to-End," Thèse de doctorat, Université Toulouse 3 Paul Sabatier, Mar. 2021. [Online]. Available: <https://hal.archives-ouvertes.fr/tel-03269807>
- [22] R. van Hout and A. Vermeer, "Comparing measures of lexical richness," in *Modelling and assessing vocabulary knowledge*, H. Daller, J. Milton, and J. Treffers-Daller, Eds. Cambridge: Cambridge University Press, 2007, pp. 93–116.
- [23] C. Roze, L. Danlos, and P. Muller, "LEXCONN: a French lexicon of discourse connectives," *Discours - Revue de linguistique, psycholinguistique et informatique*, 2012. [Online]. Available: <https://hal.inria.fr/hal-00702542>
- [24] S. Detey and Y. Kawaguchi, "Interphonologie du Français Contemporain (IPFC): récolte automatisée des données et apprenants japonais," in *Journées PFC: Phonologie du français contemporain: variation, interfaces, cognition*, Paris: MSH, Dec. 2008.
- [25] I. Racine, F. Zay, S. Detey, and Y. Kawaguchi, "Des atouts d'un corpus multitâches pour l'étude de la phonologie en L2: l'exemple du projet "Interphonologie du français contemporain" (IPFC)," in *Recherches récentes en FLE*, A. Kamber and C. Skupien, Eds. Bern: Peter Lang, 2012, pp. 1–19.