A corpus of spontaneous L2 English speech for real-situation speaking assessment

Sylvain Coulange[1,2], **Marie-Hélène Fries**[3], **Monica Masperi**[1], **Solange Rossato**[2]
[1] *Laboratory of Linguistics and Didactics of Foreign and Mother Tongues (LIDILEM)*
[2] *CNRS, Institute of Engineering, Grenoble Computer Science Laboratory (LIG)*
[3] *National Coordination for the Certificate of language skills in French higher education (CLES)*
Université Grenoble Alpes, 38000 Grenoble, France
{sylvain.coulange, marie-helene.fries, monica.masperi, solange.rossato}@univ-grenoble-alpes.fr

Abstract

When assessing second language proficiency (L2), evaluation of spontaneous speech performance is crucial. This paper presents a corpus of spontaneous L2 English speech, focusing on the speech performance of B1 and B2 proficiency speakers. Two hundred and sixty university students were recorded during a speaking task as part of a French national certificate in English. This task entailed a 10-minute role-play among 2 or 3 candidates, arguing about a controversial topic, in order to reach a negotiated compromise. Each student's performance was evaluated by two experts, categorizing them into B2, B1 or below B1 speaking proficiency levels. Automatic diarization, transcription, and alignment at the word level were performed on the recorded conversations, in order to analyse lexical stress realisation in polysyllabic plain words of B1 and B2 proficiency students. Results showed that only 35.4% of the 6,350 targeted words had stress detected on the expected syllable, revealing a common stress shift to the final syllable. Besides a substantial inter-speaker variability (0% to 68.4%), B2 speakers demonstrated a slightly higher stress accuracy (36%) compared to B1 speakers (29.6%). Those with accurate stress placement utilized F0 and intensity to make syllable prominence, while speakers with lower accuracy tended to lengthen words on their last syllables, with minimal changes in other dimensions.

Keywords: Spontaneous speech, corpus, lexical stress, assessment, rhythm, comprehensibility

1. Introduction

In today's globalized world, the demand for English proficiency among learners has never been higher. To cater to this need, providing learners with effective materials and tools to enhance their speaking ability is crucial. A multitude of automated scoring systems has been designed to assist human teachers and raters in coping with the rising demand for speaking practice and evaluation. However, existing automated scoring methods often rely on highly controlled elicitation protocols, such as reading aloud isolated words or short sentences, limiting their ability to evaluate spontaneous speech (Saito et al., 2022).

To address this gap and develop a digital, partially automated version of its tests, the CLES[1] (Certification de Compétences en Langues de l'Enseignement Supérieur), a state language certificate established by the French Ministry of Higher Education and Research, initiated the collection of spontaneous L2 speech recordings elicited by university students during exam sessions. This data, accompanied by high-quality certification-level proficiency ratings, forms a valuable resource, a significant portion of which is publicly available for research purposes.

This paper presents the data collection procedure and describes the complete corpus, followed by an analysis of lexical stress patterns by French learners of English, which is itself part of a larger PhD research project on automated evaluation of L2 speech rhythm. Indeed, while some high and low-stakes scoring systems have recently incorporated features for spontaneous speech assessment (Zhang, 2020; Coulange, 2023), score prediction predominantly depends on phenomena such as utterance length, frequency of pauses, percentage of phonation, lexical diversity, or syntactic complexity (Evanini and Zechner, 2019), but rarely rhythm-related features, such as lexical stress pattern, though its impact on comprehensibility in L2 English has been frequently highlighted (Cutler, 2015; Isaacs et al., 2018; Tortel, 2021).

A subset of the corpus presented here was utilized to explore lexical stress patterns in French learners' spontaneous L2 English speech among B1 and B2 proficiency speakers. These levels are indeed characterized by differences in speaker fluency and appropriateness of stress and rhythm (Council of Europe, 2020, p.134). The main research questions are the following: do B2 speakers produce a more accurate lexical stress (i.e. position and quality) compared to B1 speakers? and how do French-L1 speakers tend to stress words? The authors developed an automated processing pipeline for measurement of lexical stress realisation in polysyllabic plain words. This pipeline was employed to compare stress patterns among 176 B1 and B2 speakers in this evaluation situation.

[1] https://www.certification-cles.fr/english/

LREC-COLING 2024, pages 293–297
20-25 May, 2024.

Following a presentation of the CLES certification speaking tasks and the data collection recorded so far, Section 3 will delve into the data processing and speech annotation conducted for the analysis of lexical stress patterns, ending with a concise overview of the preliminary results.

2. The CLES corpus of spontaneous L2 English

2.1. Tasks of the CLES certification test

The CLES certificates, designed for university-level language proficiency assessment, evaluate each CEFR level independently. The corpus presented here primarily comprises recordings from the B2 level test. In this test, participants engage in a 10-minute role-play where two or three candidates delve into an argumentative discussion on contentious topics, like security cameras, animal testing, or e-cigarettes. Each candidate assumes a specific role, either advocating for or against the subject, and has a two-minute preparation period before the conversation begins. Although participants are allowed to take notes, reading during the conversation is prohibited. Their objective is to negotiate, exchange viewpoints, and work towards a compromise.

Professional onsite raters evaluate each candidate on eight dimensions related to oral production at the B2 level: positioning and negotiation skills, relevance and variety of arguments, interaction aptitude, fluency, phonetic accuracy, coherence, grammatical precision, and lexical diversity and appropriateness. Failure to meet any of these criteria results in a validation at the B1 level, or no validation if proficiency falls below the threshold. Candidates are ultimately classified as B2, B1, or non-validated based on their performance.

Given the smaller proportion of B1 proficiency students during the recording sessions, recordings from the B1 level test were also included. This test requires candidates to deliver a monologue-type prepared spontaneous speech and engage in a role-play where they record two vocal messages about a home exchange program. In the first message, candidates introduce themselves, inquire about the house they wish to book, and leave a contact number for callbacks. In the second message, they describe their own house and its accommodations, following the instruction provided. Candidates have 5 minutes to prepare before each recording, during which they can take notes, but must refrain from reading them during the message recording. An onsite rater assesses the candidate's performance based on compliance with instructions, presence of key information, overall intelligibility, pronunciation of key words, coherence, grammar, and vocabulary. Only B1 level can be validated in this task.

Conversation type	Nb. of speakers	Duration
3-speaker	15	1h03'44"
2-speaker	232	18h16'50"
1-speaker	13	39'28"
Total	260	20h00'02"

Table 1: Number of speakers per conversation.

Proficiency	Nb. of speakers	%
B2	151	58%
B1	75	29%
non-validated	34	13%

Table 2: Number of speakers per speaking proficiency level.

2.2. Corpus description

The corpus comprises recordings of 134 groups of students. Among them, 116 feature 2-speaker conversations, 5 involve 3 speakers, and 13 consist of monologues from the B1 level test. These recordings were conducted in empty classrooms at University Grenoble Alpes during sessions held in January 2020, May 2022, December 2022, and January 2023. To ensure consistency, recordings were edited to exclude explanation and preparation time, retaining only the candidate conversations.

Each candidate's metadata includes proficiency ratings in listening, reading, writing, and speaking, overall proficiency, test level (B1 or B2), assigned role in the role-play scenario, number of speakers in the conversation, gender, and mother tongue. Notably, 83% of the speakers (n=215) declared French as their mother tongue, while the remaining 17% represented 14 diverse mother tongues.

Among the recorded speakers, 58% (n=151) were rated B2 speaking proficiency by the raters, 29% (n=75) B1 proficiency, and 13% (n=34) failed to validate the task. Among the B1 speakers, 11 attempted the B1 level test, along with 2 speakers who received no validation. The total corpus duration is 20 hours, with B2 level test conversations averaging 9'35" (ranging from 5'12" to 14'30") and B1 level test monologues averaging 3'2" (ranging from 1'46" to 5'4", cf. Tables 1 and 2).

3. Lexical stress analysis

Rhythm significantly impacts speaker intelligibility, particularly in L2 English, aiding listeners in segmenting and processing speech (Cutler, 2015). However, acquiring a new rhythm can be challenging for EFL learners, especially if their native language lacks lexical stress (Tortel, 2021). French, for instance, tends to exhibit a fixed stress on the last syllable of content words, mainly characterized by a longer duration (Astesano, 2001), whereas its position varies in English and generally combines a rise of fundamental frequency (F0) and intensity, longer duration, and vowel reduction in adjacent syllables (Cutler, 2015). Consequently, from a suprasegmental point of view, we can expect French speakers of English to lengthen final syllables of most words, with limited use of F0 and intensity.

There has been considerable research into automated lexical stress classification in recent decades. Commonly, these systems employ F0, intensity, and duration measures, and sometimes segmental information, to predict word stress patterns (Li et al., 2018; Johnson and Kang, 2015; Ferrer et al., 2015). However, these tools require extensive training with annotated data, and they hardly allow us to understand how stress is produced by the speakers.

This section presents the analysis of a subset of the previously outlined corpus, focusing on B1 and B2 speakers with French as their mother tongue (n=176; B1=59, B2=117). The analysis of lexical stress involved comparing the acoustic stress pattern of polysyllabic content words with their expected lexical stress patterns, according to the CMU Pronouncing Dictionary[2], and quantifying the contrast between stressed and unstressed syllables. Notably, only primary stress was considered in this study.

3.1. Methodology

To analyse the lexical stress patterns of individual speakers, recordings first underwent speaker diarization using Pyannote (Bredin and Laurent, 2021). Mono-speaker continuous speech segments were then extracted for independent processing. Each segment was transcribed and aligned at the word level using WhisperX (Bain et al., 2023). Subsequently, a complementary part-of-speech tagging was added following morphosyntactic analysis conducted with SpaCy (Honnibal et al., 2020) in order to take into consideration word category. Syllable nuclei were identified using the Praat script described in (de Jong et al., 2021), with a bandpass filter at 300–3300Hz to minimize the influence of non-voice-related events.

For each word, the expected number of syllables was derived from the dictionary and compared with the number of syllable nuclei detected within the word boundaries. In order to avoid analysing incorrectly aligned words, only words with the correct number of syllables were included in the analysis. We will call them target words.

Stress was measured along three dimensions: F0 and intensity at syllable nucleus, and syllable duration, estimated from the midpoints of neighboring syllable nuclei and/or word boundaries. Each measure was speaker-normalised to mitigate variation of speech rate, voice height and microphone distance. In subsequent sections, prosodic values are expressed in percentiles, calculated from each speaker's overall value distribution: 50 signifies the speaker's median value, while 90 represents very high (or loud, or long) values, and 10 indicates very low values for that specific speaker.

Acoustic stress was determined for the most prominent syllable within each word for all three dimensions. These dimensions were given equal weight to derive a unified global stress pattern.

3.2. Results

Out of the 6,350 target words analysed, a mere 35.4% exhibited correct acoustic stress placement. For 2-syllable words, where 85% were expected to have the stress on the first syllable, only 31% exhibited this pattern; instead, 69% received stress on the final syllable (see Figure 1). This trend persisted for 3- and 4-syllable words, where most words appeared stressed on the final syllable, contrary to the expected first or second syllable stress.

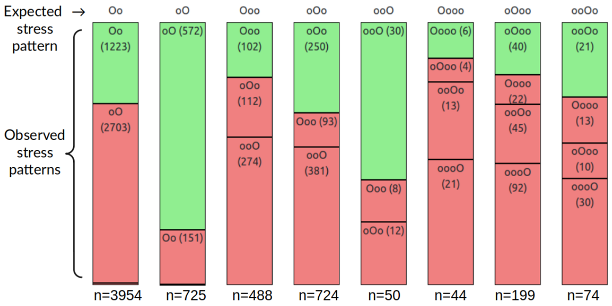

Figure 1: For each expected stress pattern in columns, the number of words for each observed pattern is shown. "O" represents the stress syllable, "o" represents other syllables.

Individual accuracy in stress position varied widely, ranging from 0% to 68.4%. Proficiency-based analysis revealed substantial overlap between B1 and B2 speakers, albeit with B2 speakers outperforming their B1 counterparts on average (median at 36% vs. 29.6%, $p < .0001$). Figure 2 illustrates stress position accuracy for each speaker, highlighting a stark contrast: only 3% of B1 speakers achieved accuracy above 50%, whereas 22% of B2 speakers surpassed this threshold.

Figure 3 illustrates the mean acoustic contrast between stressed and unstressed syllables for each speaker, revealing a wide range of speaker profiles within both B1 and B2 groups. Notably, only B2 speakers consistently produced appropriate, contrasted stress on the correct syllable (top right of the chart), except for one B1 speaker. The difference in mean contrast between the two groups was significant ($p < .0001$).

[2] The dictionary is available at: `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

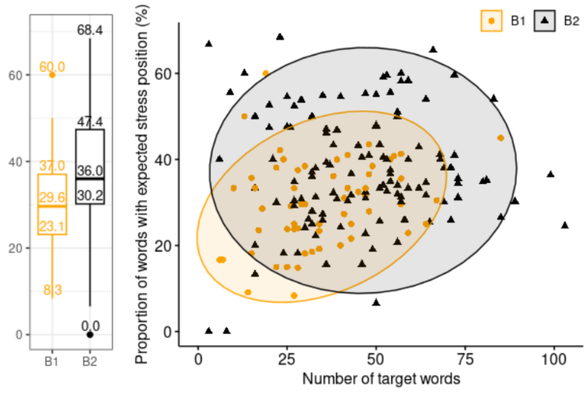

Figure 2: Proportion of target words with expected stress position per speaker.

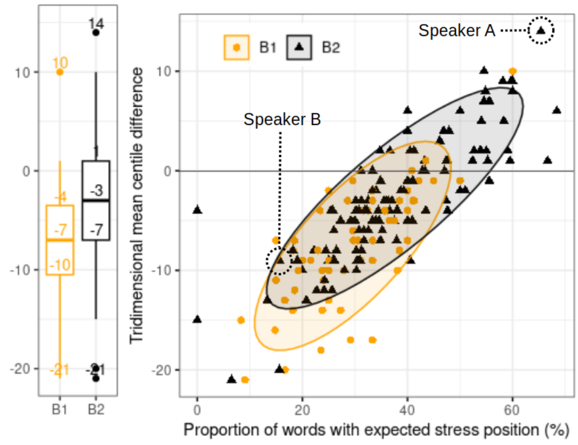

Figure 3: Mean acoustic difference between expected stressed and reduced syllables per speaker. Both variables are strongly correlated ($R = .82$, $p < .0001$).

In Figure 4, the contrast between expected stressed and unstressed syllables is demonstrated for two B2 speakers. Speaker A exhibited 65% accuracy in stress placement, emphasizing the expected stressed syllable significantly in F0 (30 points higher) and intensity (+17). In contrast, speaker B, with only 16% accuracy, stressed the wrong syllable (one of the unstressed ones), displaying a notable increase in duration (21 points longer) and F0 (+11), with no substantial change in intensity. Similar patterns were observed in other speakers with high or low stress placement accuracy. Those proficient in stress placement accentuated the expected stressed syllable primarily through increased F0 and intensity, while speakers adhering to French stress patterns accentuated the duration and F0 of the last syllables, contrary to English norms, with no change in intensity.

4. Discussion

Systematic prosodic measures of lexical stress revealed an expected trend among French speakers

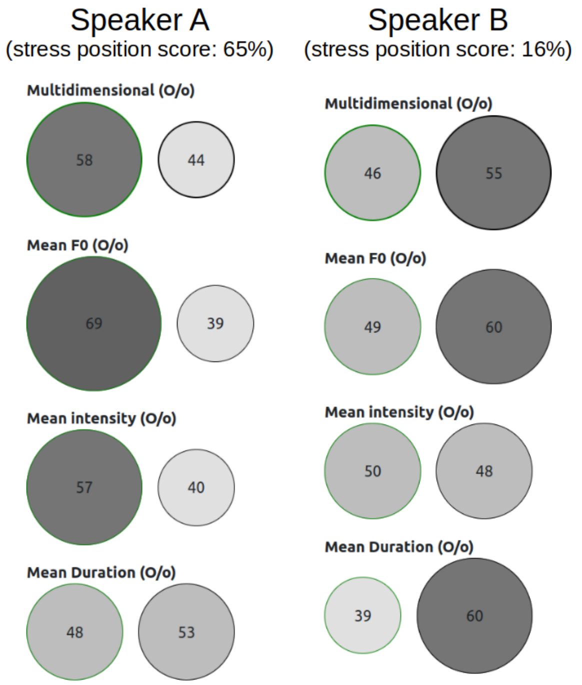

Figure 4: Mean centile value of prominence for expected stressed (first circle) and reduced (second circle) syllables in each dimension for speaker A and speaker B.

of English to lengthen final syllables. Besides a natural tendency to lengthen and rise last syllables of words in conversational contexts, and moreover in a foreign language, this common stress shift to the last syllable might be a consequence of speakers' L1. The analysis revealed a wide variation among speakers, B2 speakers performing generally better than B1 speakers, with a large overlap between the two groups. The same pattern is observed with prosodic contrast between stressed and unstressed syllables, with a strong – and unsurprising – correlation with stress placement accuracy. Finally, it is interesting to observe that speakers with low stress placement accuracy tend to make prominence mostly through duration change, while high accuracy speakers tend to neutralize duration and focus on F0 and intensity variation.

These results corroborate those of Tortel (2021), who noted that French speakers often stress final syllables by lengthening them, with limited or no reduction of neighboring vowels. Tortel's observations were based on readings by 20 high school English-level speakers and 20 English language bachelor-level speakers. Additionally, Tortel and Herment (2018) demonstrated, in readings by B and C level students, that vowel duration of unstressed syllables approached that of native speakers as proficiency levels increase, therefore resulting in a better contrast with the stressed vowel. It is now important to investigate stress production within spontaneous contexts, and the expansive CLES corpus enabled us to replicate and extend these studies on a larger scale, analyzing spontaneous speech.

While the present study includes a diverse set of speakers, the mean speech duration per speaker is rather limited, and variations among speakers may impact results (median at 3'39" for B1 and 3'52" for B2 speakers, ranging from 32" to 6'52"), particularly for speakers with limited speech data. The accuracy of speech recognition and word-level alignment was satisfactory: 92 out of 100 randomly checked words were correctly recognised by WhisperX, and 95 were correctly aligned. In some cases, however, word boundaries appeared to shorten either the initial or final syllable of the word, which may impact the results and needs further investigation. Finally, the methodology relies on syllable nuclei points to determine syllable prominence, neglecting pitch variation within syllables and potential impact of syllable structure on duration. To address these methodological issues, we are currently developing an enhanced pipeline that incorporates phoneme-level alignment, allowing for F0 and duration measures based on vowel intervals instead of whole syllables.

To further explore the influence of native languages, a similar study could encompass speakers with different L1s, exhibiting varied stress patterns. Additionally, conducting similar analysis at broader proficiency levels, and further investigating the use of F0, intensity and duration at those levels, could offer deeper insights into stress patterns across proficiency stages.

Acknowledgements

This work was supported by JSPS KAKENHI n°23H00648. The authors thank the IDEX for funding S.C. mobility grant. The authors express their gratitude to the Direction of CLES for collecting and sharing these recordings.

Supplementary Materials

The complete processing pipeline is open-source and freely available here: https://gricad-gitlab.univ-grenoble-alpes.fr/lidilem/plspp.

Ethical considerations and limitations

Informed consent was obtained from all subjects involved in the study. The public portion of the corpus contains recordings from 128 students, and is accessible at https://hdl.handle.net/11403/cles-spontaneous-english.

References

Corine Astesano. 2001. *Rythme et accentuation en français: invariance et variabilité stylistique*. Collection Langue & parole. L'Harmattan.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *Interspeech*.

Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In *Interspeech*.

Sylvain Coulange. 2023. Computer-aided pronunciation training in 2022: When pedagogy struggles to catch up. In Alice Henderson and Anastazija Kirkova-Naskova, editors, *Proc. of the 7th International Conf. on English Pronunciation: Issues and Practices*, pages 11–22.

Council of Europe. 2020. *Common European framework of reference for languages*. Council of Europe, Strasbourg, France.

Anne Cutler. 2015. Lexical stress in english pronunciation. In *The Handbook of English Pronunciation*, pages 106–124. John Wiley & Sons, Inc, Hoboken, NJ.

Nivja H. de Jong, Jos Pacilly, and Willemijn Heeren. 2021. Praat scripts to measure speed fluency and breakdown fluency in speech automatically. *Assessment in Education: Principles, Policy & Practice*, 28:456–476.

Keelan Evanini and Klaus Zechner. 2019. *Overview of automated speech scoring*, Innovations in Language Learning and Assessment at ETS, pages 3–20. Routledge, London, England.

Luciana Ferrer, Harry Bratt, Colleen Richey, Horacio Franco, Victor Abrash, and Kristin Precoda. 2015. Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems. *Speech Comm.*, 69:31–45.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Talia Isaacs, Pavel Trofimovich, and Jennifer Ann Foote. 2018. Developing a user-oriented second language comprehensibility scale for english-medium universities. *Language Testing*, 35(2):193–216.

David O. Johnson and Okim Kang. 2015. Automatic prominent syllable detection with machine learning classifiers. *Int. J. Speech Technol.*, 18(4):583–592.

Kun Li, Shaoguang Mao, Xu Li, Zhiyong Wu, and Helen Meng. 2018. Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks. *Speech Comm.*, 96:28–36.

Kazuya Saito, Konstantinos Macmillan, Magdalena Kachlicka, Takuya Kunihara, and Nobuaki Minematsu. 2022. Automated assessment of second language comprehensibility: Review, training, validation, and generalization studies. *Second Lang. Acquis.*, pages 1–30.

Anne Tortel. 2021. Le rythme en anglais oral : considérations théoriques et illustrations sur corpus. *Recherche et pratiques pédagogiques en langues - Cahiers de l'APLIUT*, (Vol. 40 N°1).

Anne Tortel and Sophie Herment. 2018. La voyelle inaccentuée e en position initiale : analyses acoustiques et enjeux pédagogiques pour l'anglais L2. In *XXXIIe Journées d'Études sur la Parole*, ISCA. ISCA.

Emily Di Zhang. 2020. Automated speaking assessment: Using language technologies to score spontaneous speech. *Lang. Assessment Quarterly*, 17(3):327–330.