DEVELOPMENT OF THE FRBIO SENTENCE TEST

François Bergeron, Dominique Demers  
*Université Laval, Québec, Québec, Canada*  
*email: francois.bergeron@fmed.ulaval.ca*

Aurore Berland  
*Université de Toulouse II*

Elizabeth M. Fitzpatrick  
*Université d’Ottawa, Ottawa, Ontario, Canada*

Christophe Vincent  
*CHRU de Lille, Lille, France*

Annie Giasson  
*Hôpital Régional Chaleur Regional Hospital, Bathurst, Nouveau Brunswick, Canada*

Kevin Leung Kam, Walid Chafiq, Thibaut Fanouillère  
*Université de Montpellier, Montpellier, France*

The AzBio is a sentence test that has been widely adopted by clinical teams to evaluate speech recognition performance in adults with hearing impairment. The goal of this work was to develop and normalize an international French version of the AzBio sentence test following the same steps used in creating the original test. The initial FrBio corpus included 1000 sentences generated by international French speaking collaborators and recorded with two female and two male talkers. The mean intelligibility of each sentence was estimated by processing each sentence through a four-channel cochlear implant (CI) simulation and calculating the mean percent correct score achieved by 16 normal-hearing listeners in France and Canada. Sentences from each talker were sorted by percent correct score, and 165 sentences were selected from each talker and then sequentially assigned to 33 lists, each containing 20 sentences (5 sentences from each talker). List equivalency was validated by presenting all lists, in random order, to 30 listeners with normal hearing and 25 listeners with varying degrees of hearing impairment in France and Canada. The 33 lists of sentences had a mean intelligibility score of 85% with normal hearing listeners. The results of the validation study with normal hearing and hearing-impaired listeners revealed no significant differences in percent correct scores for 30 of the 33 sentence lists.

Keywords: French AzBio sentence recognition test
1. Introduction

The AzBio was developed at the Arizona State University in 2012 to 1) provide an unbiased evaluation of individuals with hearing impairment who had extensive exposure to traditional sentence materials, 2) allow for evaluation of performance in a large number of listening conditions, 3) create lists of sentences with similar levels of difficulty for within-subject comparisons, and 4) provide an estimate of performance that was consistent with the patient’s perception of their performance in everyday listening environments [1].

Since its introduction, the AzBio has been widely adopted by English speaking cochlear implant teams, especially because of the complexity of the sentence material that prevents the ceiling effect observed with the use of common clinical materials when assessing contemporary hearing devices.

The availability of a complex sentence material is also highly desirable for assessing hearing impaired individuals in other languages. Thus, this project aimed to develop and normalize an international French version of the test.

2. Materials and methods

2.1 Sentence construction

Similarly to the original AzBio, the initial FrBio sentence corpus included 1000 sentences based on up-to-date, adult topics and current social ideas. Sentence length was limited from 3 to 12 words (mean = 8.0, SD = 2.1) and proper nouns were avoided. No other restrictions were placed on complexity, vocabulary, or phonemic content.

As an international use of the French version is expected, the corpus was generated through input from collaborators located in different regions of France (Lille, Toulouse) and Canada (Ontario, Quebec, New-Brunswick); these collaborators provided written sentences in accordance to the criteria defined above. All sentences were shared among collaborators for review; unfamiliar sentences in any region were excluded. The final list then constitutes a consensus on the familiarity of the lexicon and expressions in all regions.

2.2 Sentence recordings

Four adult talkers, two males and two females, were selected to each record 250 sentences. These talkers were selected following a survey based on audio samples among our international collaborators. In order to select the best intelligibility independently of regional accent, collaborators had to rate a variety of speakers reading some sentences. At the end, the selected talkers were characterized by a composite accent; they were either natives from France who had lived in Canada for many years, or native Canadians educated in a European French environment.

During recording, talkers were seated in a sound-treated booth. The 1000 sentences (250 per talker) were recorded using a RODENT2 condenser microphone connected to a Zoom H4N recorder. All recordings were made with a sample frequency of 44.1 Hz and 24-bit resolution. The microphone was positioned approximately 6 to 10 inches from the talker. Each talker was instructed to speak at a normal conversational pace and volume and to avoid using overly enunciated speech. Sentence production was monitored by an examiner. In the event of mispronunciations, misread words, or any other unintended
disruptions, the talker was prompted to repeat the sentence. The final production of each sentence was isolated from the recorded block and saved as a unique sound file. A global adjustment was made to the 250 recorded sentences of each talker (e.g., all recordings from a single talker were attenuated by 2 dB) to control for slight differences in recording levels across talkers. The final sound files were rendered in a 22,050-sample rate /16 bits sampling format.

Across talkers, the average speaking rate ranged from 4.3 to 6.1 syllables per second, consistent with normal speaking rates [2]

2.3 Sentence intelligibility estimation

The mean intelligibility rating for each sentence was estimated by processing each sentence through a four-channel cochlear implant (CI) simulation (Dorman et al. 1998) and calculating the mean percent correct score achieved by 16 normal-hearing listeners recruited in the same regions of France and Canada, where the sentences were initially generated. The CI simulation was introduced to make the sentences more difficult, therefore limiting a potential ceiling effect for normal-hearing listeners. Listeners were seated in a sound-treated booth, instructed to repeat each sentence and to guess when unsure about any word. Sentences were presented from a laptop computer at a comfortable level using high quality headphones. The initial sentences were interleaved by a step of 100 between each listener in order to diffuse any learning effect. Sentences were scored as the number of words correctly repeated by each listener. The mean percent correct score for each sentence (total words repeated correctly/total words presented) was used as the estimate of intelligibility. The mean score for all listeners and sentences was 85%, (SD = 9.2) with a range of sentences scores between 11 to 100%.

2.4 Sentence selection and list formation

Similarly to the original AzBio development, sentences from each talker were sorted by percent correct score. In order to generate lists with a similar level of difficulty to the original AzBio, the consecutive 165 sentences that generated a mean score of 85% were selected from each talker; these sentences were sequentially assigned to 33 lists, each containing 20 sentences, that is 5 sentences from each talker. This sentence-to-list assignment produced lists with a global mean score of 85% correct, which is equal to the original AzBio test. Lists had an average of 159 words; the English version had an average of 142 words. Average intelligibility of individual talkers across lists was 85.7% (SD = 0.10) and 84.9% (SD = 0.10) for the two female talkers and 83.5% (SD = 0.08) and 85.8% (SD = 0.08) for the two male talkers.

2.5 List equivalency validation

The FrBio list equivalency validation followed a two-step procedure. First, in order to confirm that the lists are equivalent regardless of linguistic factors, all lists were presented to 30 normal-hearing listeners equally recruited in France and Canada. During testing, subjects were seated in a sound-treated booth. Sentences were presented at a comfortable level, typically 60 dB SPL, in the sound field from a single loudspeaker at 0° azimuth on the horizontal axis. Subjects were instructed to repeat each sentence and to guess when unsure of any word. Each sentence was scored as the number of words repeated correctly, and a percent correct score was calculated for each list.

Second, in order to confirm the equivalence of lists for the target clientele, all lists were presented to a group of 25 mild to severe hearing-impaired listeners recruited in France and Canada. The same testing conditions were applied with this group.
3. Results

Student-t tests were computed on each list in order to verify the similarity of performance between Canada-French and France-French listeners. No significant difference was found on most lists (p>0.05), except for lists 6 and 28. While performances for list 6 were clearly different between Canada and France with average scores of 95.0% and 99.3% respectively, performances with list 28 were almost perfect (98.9 vs 99.9% respectively); thus, a ceiling effect potentially biased the analysis on latter list. As performance for the two groups then appears similar, data from both groups were congregated for subsequent analysis.

Figure 1 shows the scores for the 30 normal hearing listeners recruited for the first stage of the validation study involving linguistic factors. Scores are normalized per participant to isolate the list effect. For each listener, the individual list scores were reduced by that listener’s mean percent correct score on all 33 lists (list score - mean score). This transformation retains the distribution characteristics of the original list scores but normalizes the mean score for each listener to zero. An analysis of variance revealed a significant main effect for the list number (p<0.00). A post hoc Tukey test revealed that list 6 rendered lower results than all the other lists.

Fig. 1. Normalized scores for 30 normal-hearing listeners on all 33 sentence lists. Symbols represent an individual listener’s list score relative to their overall mean level of performance. Positive values indicate better than average performance and negative values indicate below average performance.
The mean level of performance achieved by the 25 hearing-impaired listeners recruited for the second stage of the validation study ranged from 52 to 100% correct (mean = 93%, SD = 11). Scores shown in figure 2 are also normalized by participant to isolate the list effect. An analysis of variance revealed a significant main effect for the list number (p<0.00). A post hoc Tukey test revealed that lists 6, 9 and 33 produced lower results than the other lists.

![Figure 2](image)

**Fig. 2.** Normalized scores for 25 hearing impaired listeners on all 33 sentence lists. Symbols represent an individual listener’s list score relative to their overall mean level of performance. Positive values indicate better than average performance and negative values indicate below average performance.

On the basis of these analysis, lists identified as significantly different than the other lists (6, 9, and 33) were removed from the set, resulting in 30 lists for the final FrBio test.

### 4. Conclusion

The goal of this work was to develop and normalize a French version of the AzBio sentence test. It was considered important that the FrBio be applicable to international French speaking clients. Thus the 1000 sentences written and recorded for inclusion in the FrBio sentence corpus were generated from a consensus of many contributors form different parts of France and Canada. From this corpus 660 sentences were used to create 33 lists of 20 sentences, and 30 of the 33 lists were found to be of equivalent intelligibility based on the scores obtained from 30 normal-hearing and 25 impaired-hearing listeners in France and Canada.

With a set of 30 lists, researchers and clinicians can use the FrBio to evaluate a large number of experimental conditions; changes in performance over time or across conditions can then be tracked.

Future research will focus on exploring the reliability of this material when applied to diverse populations and diverse listening conditions. The development of a paediatric specific corpus is planned.
REFERENCES
