



HAL
open science

Est-il possible d'annoter la naturalité des pauses lors de la lecture d'un texte à haute voix ?

Vincent P. Martin, Jean-Luc Rouas, Agathe Basse, Benoît Caudron, Marie Huillet, Pierre Philip

► To cite this version:

Vincent P. Martin, Jean-Luc Rouas, Agathe Basse, Benoît Caudron, Marie Huillet, et al.. Est-il possible d'annoter la naturalité des pauses lors de la lecture d'un texte à haute voix?. Journées d'Étude de la Parole - JEP 2022, Jun 2022, Noirmoutier, France. hal-03835281

HAL Id: hal-03835281

<https://hal.science/hal-03835281>

Submitted on 31 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/359486128>

Est-il possible d'annoter la naturalité des pauses lors de la lecture d'un texte à haute voix ?

Conference Paper · June 2022

CITATION

1

READS

42

6 authors, including:



Vincent P. Martin

University of Bordeaux

24 PUBLICATIONS 47 CITATIONS

[SEE PROFILE](#)



Jean-Luc Rouas

Laboratoire Bordelais de Recherche en Informatique

75 PUBLICATIONS 710 CITATIONS

[SEE PROFILE](#)



Pierre Philip

Université Victor Segalen Bordeaux 2

401 PUBLICATIONS 10,456 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Phenovirt [View project](#)



Upper Airway Resistance Syndrome [View project](#)

Est-il possible d'annoter la naturalité des pauses lors de la lecture d'un texte à haute voix ?

Vincent P. Martin¹ Jean-Luc Rouas¹ Agathe Basse² Benoît Caudron²
Marie Huillet² Pierre Philip³

(1) LaBRI, Univ. Bordeaux, CNRS UMR 5800, Bordeaux INP, 256 cours de la Libération, 33405 Talence, France

(2) CFUOB, Univ. Bordeaux, 146 Rue Léo Saignat, 33076 Bordeaux, France

(3) SANPSY, Univ. Bordeaux, CNRS UMR 6033, CHU de Bordeaux, 13ème étage, aile 3, Place Amélie-Raba-Léon, 33076 Bordeaux, France

{vincent.martin,rouas}@labri.fr, {agathe.basse, benoit.caudron, marie.huillet}@etu.u-bordeaux.fr, pierre.philip@u-bordeaux.fr

RÉSUMÉ

L'utilisation de biomarqueurs vocaux est une des technologies les plus prometteuses pour l'implémentation de systèmes de santé numérique en conditions écologiques. En effet, de nombreuses et diverses pathologies sont maintenant diagnostiquées automatiquement de manière fiable grâce à des marqueurs vocaux. Cet article étudie et discute la faisabilité d'annoter des textes – qui seront ensuite lus à voix haute par des patients hypersomniaques – afin de concevoir des descripteurs basés sur la naturalité des pauses faites par ceux-ci. Pour cela, trois spécialistes ont annoté six textes extraits du Petit prince. Nous étudions à la fois à travers des mesures statistiques, mais aussi sous le prisme du rapport des annotateurs, deux axes : les différences entre les méthodes d'annotation ; et les lieux de désaccord dans les textes. Enfin, nous concluons quand à la fiabilité d'utiliser de telles annotations comme vérité terrain dans un système automatique.

ABSTRACT

Is it possible to annotate the naturalness of pauses made during reading out loud ?

The use of voice and speech biomarkers is one of the most promising technologies to implement digital health in ecological conditions. Indeed, numerous and diversified pathologies are now accurately diagnosed by autonomous systems based on voice features. This article investigates and discusses the feasibility of annotating texts – that are then read out loud by hypersomniac patients – in order to design new features based on the naturalness of the pauses made by the patients. To do so, three specialists annotated six texts extracted from Le Petit Prince. We investigate two axes through statistics but also from the perceptive of annotators' report : differences in annotation methods ; and disagreement location in texts. Finally, we conclude about the robustness of using such annotations as ground truth in an automated pipeline.

MOTS-CLÉS : pauses de lecture, biomarqueurs vocaux, accord interannotateurs, lecture oralisée.

KEYWORDS: reading pauses, voice biomarkers, inter-annotator agreement, reading out loud.

1 Introduction

1.1 La voix : un outil de diagnostic à l'ère de la médecine numérique

L'utilisation de biomarqueurs vocaux pour le diagnostic ou le suivi de pathologies chroniques est un des champs les plus prometteurs de la santé numérique (Fagherazzi *et al.*, 2021). En effet, contrairement aux instruments de mesure classiques comme la vidéo ou l'électroencéphalographie, les enregistrements de voix sont peu chers (à la fois en termes humains, calculatoires et monétaires) et sont robustes aux environnements bruités, facilitant leur implémentation en milieu écologique. De plus, avec l'arrivée récente des agents de conversation virtuel dans le domaine de la santé [comme par ex. *KANOPEE* (Dupuy *et al.*, 2021)] et l'utilisation grandissante des technologies vocales dans la vie courante (Sorensen, 2019), l'implémentation de la voix dans ces systèmes promet un diagnostic et un suivi non-invasif, centré sur le patient (médecine de précision) et déployable à large échelle.

Puisque la parole nécessite à la fois la coordination de plus de 100 muscles (Denes & Pinson, 1963) et de nombreux processus psycho et neurolinguistiques (Brown & Hagoort, 2000), les pathologies neurodégénératives, mais aussi les troubles mentaux sont maintenant détectés grâce à des biomarqueurs vocaux, avec des performances permettant leur utilisation en conditions cliniques [pour des revues de la détection de pathologies dans la voix, voir (Fagherazzi *et al.*, 2021) et (Low *et al.*, 2020)].

Dans la littérature, nous avons ainsi pu identifier que ces systèmes sont basés sur des combinaisons des descripteurs suivants :

- des descripteurs acoustiques, extraits par exemple avec les boîtes à outils `openSMILE` (Eyben & Schuller, 2015), ou `AUDEEP` (Freitag *et al.*, 2017);
- l'organisation des mots dans les phrases, comme par exemple pour la maladie d'Alzheimer (Shah *et al.*, 2021);
- les erreurs de systèmes de transcription automatique, proposés par exemple pour la maladie de Parkinson (Romana *et al.*, 2021);
- ou encore les pauses de lecture, très utilisées par exemple dans la détection de la maladie d'Alzheimer (Nasreen *et al.*, 2021).

1.2 Objectifs et structure de l'article

À notre connaissance, les études utilisant les pauses de lecture n'utilisent que l'estimation de la longueur des pauses (généralement grâce à un détecteur d'activité vocale). Notre objectif est d'étudier, sur de la parole lue et dans le cadre de la somnolence pathologique, la pertinence de nouveaux marqueurs : l'emplacement des pauses de lecture. Les lecteurs s'arrêtent-ils/elles plutôt à des endroits où la pause est 'naturelle' (comme par exemple au niveau d'un signe de ponctuation) ou non (comme par exemple en milieu d'un mot), et dans quelle proportion ?

Dans un premier temps, nous avons élaboré un système permettant l'extraction automatique des durées et des emplacements des pauses de lecture dans le texte de référence (Martin *et al.*, 2021a). Nous désirons maintenant construire une grille de notation de référence de ces pauses, afin de quantifier leur naturalité. Pour cela, nous avons demandé à trois étudiants en cursus d'orthophonie d'annoter en triple aveugle les pauses entre les mots de chaque texte de notre corpus avec un score indiquant si une pause de lecture à cet emplacement leur semble naturelle ou non (Section 2).

Comme tout processus d'annotation, évaluer la naturalité des pauses de lecture d'un texte reflète la subjectivité et la sensibilité propre des annotateurs à ce qu'est une pause « naturelle ». Afin de faire émerger les points communs et différences d'appréhension de ce concept, nous leur avons donc demandé, après avoir finalisé leurs annotations, de rédiger un rapport sur leur façon d'annoter. L'originalité de cette étude est de proposer une comparaison entre, d'une part l'évaluation subjective de chaque annotateur concernant sa façon d'annoter, et d'autre part des statistiques sur les annotations résultantes. Cette comparaison permet d'objectiver – ou non – les ressentis des annotateurs et d'ainsi évaluer la fiabilité de leur production pour une utilisation en tant que vérité terrain dans un système automatisé.

Nous proposons d'évaluer les annotations selon deux axes :

- l'analyse des différences entre les trois annotateurs, présentée dans la Section 3 ;
- l'analyse des lieux de désaccords, présentée dans la Section 4.

L'utilisation ultérieure de ces annotations dans un système automatisé est cependant conditionnée à un bon agrément interannotateurs sur les textes proposés. La corrélation intraclasse pour chaque combinaison d'annotateurs et chaque texte est présentée dans la Section 5 et permet de conclure quant aux observations faites dans cette étude.

2 Corpus et consignes d'annotation

2.1 Corpus

Le corpus utilisé dans notre étude est le corpus TILE (Martin *et al.*, 2021b, *MSLTc* en anglais). Il contient les enregistrements de 125 patients du pôle universitaire de médecine du sommeil de Bordeaux, passant un Test Itératif de Latence d'Endormissement (Arand *et al.*, 2005). Durant ce test, les patients font cinq siestes toutes les deux heures à partir de 9h, et leur temps d'endormissement est mesuré par des mesures polysomnographiques. Notre corpus contient des enregistrements vocaux correspondant à la lecture à voix haute d'un texte avant chaque sieste, plus un enregistré la veille du test (Texte n°0), servant d'enregistrement de référence. Il y a donc un total de six textes à annoter. Les textes sont différents, mais tous les patients les lisent dans le même ordre. Ceux-ci sont des extraits du Petit prince (A. de Saint-Exupéry), déjà utilisés lors de précédentes études phonétiques (Goldman *et al.*, 2016) et ayant à la fois un vocabulaire et une grammaire simple, et peu d'inférence émotionnelle qui pourrait interférer avec la somnolence de la personne (ni activation ni endormissement induit).

2.2 Consigne d'annotation

Trois des auteurs – AB, BC et MH – ont annoté les textes précédents avec la consigne suivante : « Annotez à quel point il est naturel de faire une pause dans la lecture au milieu de chaque paire de mots, -10 étant très peu naturel, 10 étant très naturel, 0 étant neutre ». Il a été convenu de noter de 2 en 2, les niveaux de notation disponibles étant les suivants : -10, -8, -6, -4, -2, 0, 2, 4, 6, 8, 10. Les trois annotations ont été faites en triple aveugle, afin de pouvoir prendre en compte la définition propre à chaque annotateur du caractère "naturel" de chaque pause. Les annotations proposées ainsi que leur moyenne et leur écart-type sont disponibles en ligne ¹.

1. <https://zenodo.org/record/5813261>

3 Stratégies d'annotation

3.1 Évaluation subjective des annotateurs

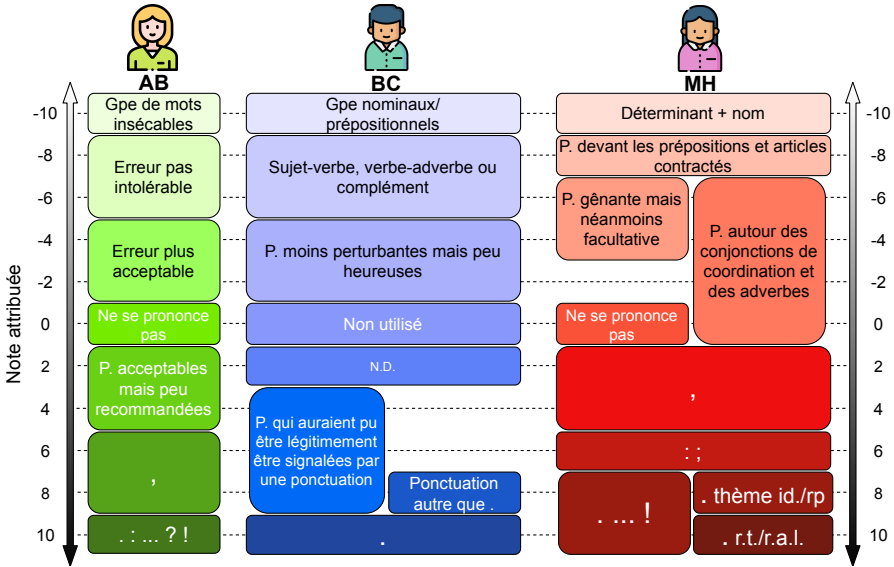


FIGURE 1 – Schéma illustrant les stratégies de chaque annotateur pour l'annotation de la naturalité des pauses de lecture dans les textes proposés. *Gpe* : Groupe; *P* : Pauses; *N.D.* : Non décrit; *thème id.* : thème identique; *rp* : reprise pronominale; *r.t.* : rupture temporelle; *r.a.l.* : retour à la ligne

Puisque l'annotation des pauses est un processus comprenant une grande part d'évaluation subjective, nous avons demandé à AB, BC et MH de décrire, a posteriori, leur façon d'annoter les pauses. Un condensé de ce rapport est proposé dans la Figure 1.

Une première remarque concerne un accord sur la sanction par une note de -10 de la séparation des groupes nominaux, insécables. Ensuite, d'un point de vue global, les notes positives semblent favoriser les arrêts aux signes de ponctuation, tandis que les notes négatives reflètent des erreurs qui gênent le rythme de lecture ou qui conduisent à des erreurs de sens. Par ailleurs, de nombreux niveaux sont utilisés non pas de manière discrète, mais de manière groupée, en utilisant différentes nuances pour le même type de pause suivant le contexte dans lequel elle se fait. Par exemple, BC nous renseigne qu'« en fonction des proportions des différents groupes [il a] marqué d'un -8 ou d'un -6 une potentielle pause entre le sujet et son verbe, comme entre le verbe et son adverbe ou complément : de telles pauses introduisent des ruptures inégalement inappropriées. »Concernant les pauses liées à la ponctuation, BC et AB sont globalement d'accord sur la favorisation des pauses par les signes de ponctuation, quels qu'ils soient. MH se distingue d'une part par son traitement des points, qu'elle divise en deux classes suivant les groupes qu'ils séparent; et d'autre part par son traitement des virgules auxquelles elle attribue des notes plus basses (+2 ou +4) que ses homologues. Elle explique cela par la façon dont elle parle couramment : « je ne pense pas marquer systématiquement les virgules par de réelles pauses, mais plutôt par des variations dans l'intonation. »Enfin, il est à noter que BC précise n'utiliser que très rarement la note de 0 : « à mon sens, une pause est acceptable (positive) ou

ne l'est pas (négative), quel que soit par ailleurs le degré d'acceptabilité. »

3.2 Approche statistique

Afin d'objectiver ces tendances d'annotations, nous avons représenté dans la Figure 2 les histogrammes des annotateurs, sur les six textes confondus. Ces histogrammes permettent de mettre en lumière

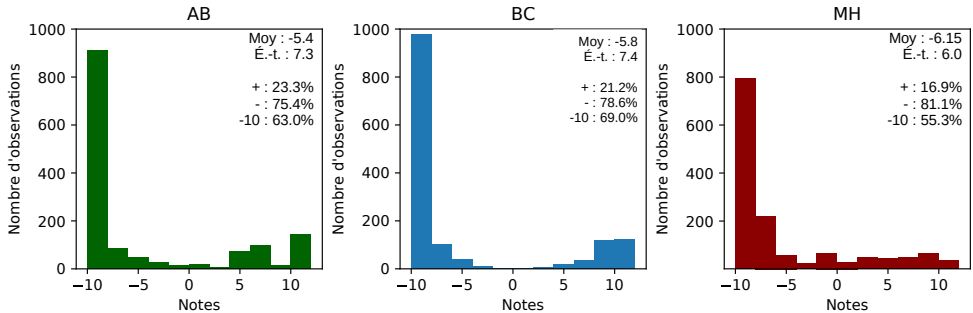


FIGURE 2 – Histogrammes des annotateurs, tous textes confondus. *Moy* : Moyenne, *É.-t.* : Écart-type, + : ratio de scores positifs, - : ratio de score négatif.

un usage principal de la note -10 : sur le total des 1436 pauses annotées, plus de la moitié sont aberrantes (score de -10) et plus des trois quarts sont jugées comme n'étant pas naturelles (score négatif) par les annotateurs. Ceci est cohérent avec le comportement de lecture qu'aurait un sujet sain, qui ne s'arrêterait que quelques dizaines de fois (à des pauses notées positivement) dans un texte contenant entre 200 et 250 mots. Ces représentations permettent également de mettre en lumière trois distributions différentes des annotations.

Chez BC, nous constatons une forte polarisation autour des notes -10 et +8/+10 conséquence de la polarité de sa conception des pauses de lecture : « j'ai attribué la note de 10 à chaque pause faisant suite aux différents points possibles, ne serait-ce que pour sanctionner la règle – très élémentaire – selon laquelle « une phrase commence par une majuscule et finit par un point ». ». On retrouve également une distribution partagée entre les deux pôles -10/+10 chez AB, mais qui est nuancée par une utilisation décroissante des niveaux intermédiaires au fur et à mesure que l'on s'approche de la notation neutre de 0.

Enfin, si l'on retrouve un pôle -10/-8 chez MH, les notes positives sont utilisées de manière quasiment uniforme. Cela est une volonté de sa part : « Afin d'aboutir à une notation utilisant autant que possible tous les degrés présents dans l'échelle de notation et prenant en compte tous les découpages de phrase possibles, j'ai lu plusieurs fois des passages du texte à haute voix, en faisant varier mon rythme de lecture et mon intonation. »

Ces différences observées entre les trois distributions sont statistiquement significatives (test de Kruskal-Wallis, $H = 14.7$, $df = 2$, $p = 6.5 \times 10^{-4}$). Un test post-hoc de Mann-Whitney sur chaque paire d'annotateurs permet de révéler une différence significative entre les annotations de AB et BC (MW, $p = 0.04$) et entre celles de MH et de BC (MW, $p = 3.9 \times 10^{-4}$), mais pas entre celles de MH et AB (MW, $p = 0.11$)

4 Lieux de désaccords

4.1 Évaluation subjective des annotateurs

Les six textes ont des caractéristiques différentes – longueur de phrases, complexité de la ponctuation, dialogues – qui peuvent influencer l’annotation des pauses de lecture. Nous avons ainsi identifié dans le rapport d’annotation quatre catégories de désaccords dont nous présentons des exemples prototypiques dans le tableau Tableau 1.

- **Dialogues (a)** : Dans son rapport, MH précise qu’elle a « rencontré plus de difficultés pour l’annotation des dialogues ». En effet, en raison de la double oralité de la lecture de ces passages à voix haute et de la mise en forme particulière, il peut être difficile d’estimer si une pause est naturelle ou non à la lecture.
- **Virgules (b)** : Au-delà du fait que MH annote les virgules différemment de ses confrères, ces dernières peuvent, suivant le contexte ou les caractéristiques de lecture de l’annotateur, être source de désaccord entre les annotateurs.
- **Ponctuation "exotique" (c)** : De même, les signes de ponctuation tels que les parenthèses ou les guillemets conduisent à des différences d’annotation.
- **Ponctuation implicite (d)** : Le rapport des annotateurs précise que « Les principaux points litigieux concernent des endroits où une pause non signalée par la ponctuation pourrait éventuellement être marquée ». Que la ponctuation puisse être marquée ou non relève dans ces cas-là de la pure évaluation personnelle de l’annotateur, rendant ces pauses plus difficilement objectivables. Dans une certaine mesure, les désaccords dus aux dialogues sont une forme spécifique de ponctuation implicite.

Ref	T	Localisation	AB	BC	MH	Moy.	É.-t.
(a)	4	« – Ah ! Ça P c’est drôle ... »	4	4	-10	-0.66	6.6
(b)	4	« je ne dessinerai pas mon avion, ; c’est un dessin beaucoup trop compliqué pour moi »	6	-10	4	0	7.12
(c)	4	« quand il aperçut pour la première fois mon avion (je ne dessinerai pas mon avion »	-10	8	4	0.66	7.72
(d)	0	« J’ai donc dû choisir un autre métier P et j’ai appris à piloter des avions. »	0	6	0	2	6.8

TABLE 1 – Exemples de désaccords d’annotations. *T* : Texte ; *Moy.* : Moyenne, *É.-t.* : Écart-type.

4.2 Approche statistique

D’un point de vue statistique, le désaccord entre annotateurs peut se traduire par l’écart-type interannotateurs. Afin d’avoir une vue générale de ces désaccords, nous avons représenté dans la Figure 3 un histogramme des écarts-types interannotateurs. Alors que presque 75% des annotations semblent conduire à un consensus solide ($\sigma < 1$), certaines pauses conduisent à des notes plus dispersées, dont l’écart-type d’annotation est supérieur à 6 (cf. Section précédente).

Dans un premier temps, nous avons donc cherché quel(s) texte(s) conduisent au meilleur accord interannotateurs. Nous avons donc calculé un test statistique de Kruskal-Wallis afin d’objectiver les variations de consensus d’annotation entre les textes, représentées Figure 3.

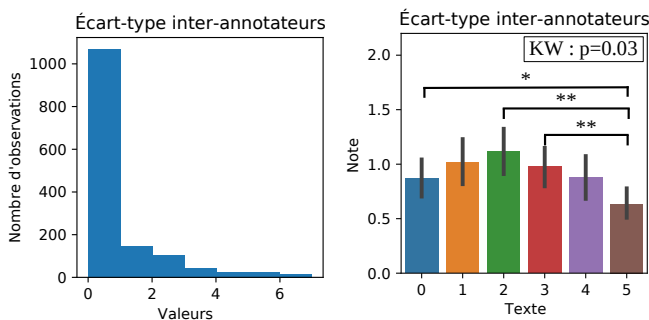


FIGURE 3 – **Gauche** : Histogramme des écarts-types interannotateurs. **Droite** : Distribution des écarts-types interannotateurs en fonction des textes (moy. \pm \acute{e} .-t.). *KW* : Test de Kruskal-Wallis ; Tests de Mann-Whitney : * : $p < 0.05$, ** : $p < 0.01$.

Les diff erences observ ees sont statistiquement significatives (*KW*, $H = 12.1$, $ddof = 5$, $p = 0.03$). Une analyse post-hoc (tests de Mann-Whitney appliqu es   chaque paire de textes) permet de mettre en lumi ere une diff erence significative entre le texte n 0 et le texte n 5 ($p = 0.02$), entre le texte n 2 et le texte n 5 ($p = 2.4 \times 10^{-3}$) et entre les textes n 3 et n 5 ($p = 4.5 \times 10^{-3}$). Le texte 5 est ainsi celui conduisant au meilleur consensus entre les trois annotateurs, avec des  carts-types interannotateurs plus faibles que sur les autres textes.

4.3 Confrontation des deux approches

En confrontant les deux approches des lieux de d esaccords– exp erience des annotateurs et approche statistique, un plus grand  cart-type interannotateurs moyen sur un texte devrait  tre la cons equence d’un plus grand nombre de lieu de d esaccords dans ce texte. Nous avons donc d enombr  et report  dans le Tableau 2 les lieux de d esaccord objectifiables selon les crit eres pr ec edemment  tablis.

Texte	n�0	n�1	n�2	n�3	n�4	n�5
Ponctuation "exotique"	2	0	2	0	2	4
Virgules	15	16	14	10	14	18
Dialogues	0	3	4	9	8	9
Total d�esaccord	17	19	20	19	24	31
Total pauses	217	253	248	229	237	252

TABLE 2 – Nombre de lieux de d esaccords potentiels dans chaque texte.

Ce tableau fait appara tre une contradiction entre la moyenne des  carts-types sur le texte n 5, qui est significativement plus faible que sur les autres textes (Figure 3) et le nombre de lieux potentiels de d esaccord, pour lequel il en poss ede le plus (Tableau 2). Une premi ere hypoth ese pour expliquer cette diff erence repose sur les pauses implicites, qui n’ont pas  t  incluses dans le pr ec edent tableau par manque de crit ere objectif de mesurabilit e. Le texte n 2 serait alors celui avec le plus de pauses implicites. Cette cat egorie de pauses, qui a  t  identifi ee par les annotateurs comme causant le plus d esaccords, est en dehors de la port ee de nos analyses statistiques et reste du domaine de l’appr eciation subjective de chaque annotateur. Une deuxi eme hypoth ese, pouvant expliquer la d ecroissance des d esaccords interannotateurs   partir du texte 2, pourrait  tre un effet d’entra nement. Au fur et   mesure

qu'ils gagnent en expérience et en aise, les trois annotateurs mesureraient de manière différente, mais plus précise un même score objectif.

5 Agrément interannotateurs et conclusion

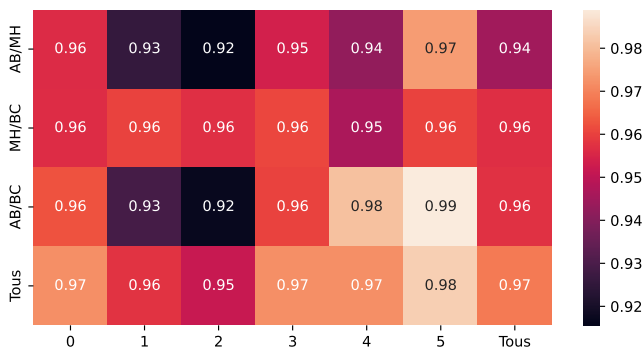


FIGURE 4 – ICC2k en fonction des textes et des combinaisons d’annotateurs

Afin de déterminer si les différences observées précédemment compromettent l’utilisation de la moyenne des trois annotateurs comme vérité terrain fiable, nous étudions dans cette partie l’agrément interannotateurs de l’annotation des pauses. Pour cela, nous calculons la corrélation intraclasse – ICC, une mesure d’agrément permettant de prendre en compte le nombre total d’observations dans chaque classe d’annotation. L’ICC se décline en 6 mesures différentes suivant l’hypothèse que l’on veut tester (Shrout & Fleiss, 1979). Nous désirons à la fois généraliser les résultats à toute population qui a les mêmes caractéristiques que nos annotateurs (c’est à dire aux orthophonistes de manière générale) et utiliser la moyenne des annotations comme vérité terrain : nous utilisons donc l’ICC de type 2 dans sa version estimant la fiabilité de la moyenne des annotations (notée ICC2k dans la suite). L’ICC2k calculée sur chaque texte et pour chaque combinaison d’annotateurs est représentée dans la Figure 4.

Toutes les valeurs de cette matrice sont supérieures à 0,9, avec une ICC tous annotateurs et tous textes confondus atteignant 0,97. Cela traduit un effet minoritaire des désaccords entre annotateurs dans les notes moyennes et témoigne d’une excellente fiabilité de l’annotation moyenne. De plus, l’ICC2 simple tous textes et annotateurs confondus (non représenté dans la figure) atteint la valeur de 0,91 : l’utilisation de l’annotation moyenne au lieu de celle d’un seul annotateur conduit à une vérité terrain plus fiable. Cette matrice est également cohérente avec les observations faites sur l’écart-type interannotateurs. En effet, les deux minimum sont observés sur le texte n°2, qui est celui ayant le plus grand écart-type interannotateurs moyen. Au contraire, les meilleurs consensus sont observés sur le texte 5, entre tous les annotateurs (ICC = 0,98) et plus particulièrement entre AB et BC (ICC = 0,99), ce qui avait déjà été observé en étudiant l’écart-type interannotateurs.

En conclusion, malgré la présence de lieux de désaccords identifiés, mais pas toujours mesurables de manière objective, les très bonnes valeurs d’accord interannotateurs permettent de valider l’utilisation de la moyenne interannotateurs comme vérité terrain pour la future extraction automatique de marqueurs.

Références

- ARAND D., BONNET M., HURWITZ T., MITLER M., ROSA R. & SANGAL R. B. (2005). The Clinical Use of the MSLT and MWT. *SLEEP*, **28**(1), 123–144.
- BROWN C. M. & HAGOORT P. (2000). *The Neurocognition of Language*. Oxford University Press.
- DENES P. B. & PINSON E. N. (1963). *The Speech Chain : The Physics and Biology of Spoken Language*. Bell telephone laboratories, edition.
- DUPUY L., MORIN C. M., SEVIN E., TAILLARD J., SALLES N., BIOULAC S., AURIACOMBE M., MICOULAUD-FRANCHI J. & PHILIP P. (2021). Smartphone-based virtual agents and insomnia management : A proof-of-concept study for new methods of autonomous screening and management of insomnia symptoms in the general population. *Journal of Sleep Research*.
- EYBEN F. & SCHULLER B. (2015). Opensmile. *ACM SIGMultimedia Records*, **6**, 4–13.
- FAGHERAZZI G., FISCHER A., ISMAEL M. & DESPOTOVIC V. (2021). Voice for Health : The Use of Vocal Biomarkers from Research to Clinical Practice. *Digital Biomarkers*, p. 78–88.
- FREITAG M., AMIRIPARIAN S., PUGACHEVSKIY S., CUMMINS N. & SCHULLER B. (2017). audeep : Unsupervised learning of representations from audio with deep recurrent neural networks. *The Journal of Machine Learning Research*, **18**(1), 6340–6344.
- GOLDMAN J.-P., HONNET P.-E., CLARK R., GARNER P. N., IVANOVA M., LAZARIDIS A., LIANG H., MACEDO T., PFISTER B., RIBEIRO M. S., WEHRLI E. & YAMAGISHI J. (2016). The SIWIS Database : A Multilingual Speech Database with Acted Emphasis. In *Interspeech 2016*, p. 1532–1535.
- LOW D. M., BENTLEY K. H. & GHOSH S. S. (2020). Automated assessment of psychiatric disorders using speech : A systematic review. *Laryngoscope Investigative Otolaryngology*, **5**(1), 96–116.
- MARTIN V. P., ARNAUD B., ROUAS J.-L. & PHILIP P. (2021a). Automatic Estimation of Pauses Duration and Location in Read Speech as Biomarkers of Sleepiness. [*Unpublished*].
- MARTIN V. P., ROUAS J.-L., MICOULAUD-FRANCHI J.-A., PHILIP P. & KRAJEWSKI J. (2021b). How to Design a Relevant Corpus for Sleepiness Detection Through Voice? *Frontiers in Digital Health*, **3**, 124.
- NASREEN S., ROHANIAN M., HOUGH J. & PURVER M. (2021). Alzheimer’s Dementia Recognition From Spontaneous Speech Using Disfluency and Interactional Features. *Frontiers in Computer Science*, **3**, 49.
- ROMANA A., BANDON J., PEREZ M., GUTIERREZ S., RICHTER R., ROBERTS A. & PROVOST E. M. (2021). Automatically Detecting Errors and Disfluencies in Read Speech to Predict Cognitive Impairment in People with Parkinson’s Disease. In *Interspeech 2021*, p. 1907–1911.
- SHAH Z., SAWALHA J., TASNIM M., QI S.-A., STROULIA E. & GREINER R. (2021). Learning Language and Acoustic Models for Identifying Alzheimer’s Dementia From Speech. *Frontiers in Computer Science*, **3**, 624–659.
- SHROUT P. E. & FLEISS J. L. (1979). Intraclass correlations : Uses in assessing rater reliability. *Psychological Bulletin*, **86**(2), 420–428.
- SORENSEN K. (2019). Millennials’ Acceptance of Voice Activated Shopping. *Textiles, Merchandising and Fashion Design : Dissertations, Theses, & Student Research*.