Évaluation perceptive de l'anticipation de la prise de parole lors d'interactions dialogiques en français

Rémi Uro, Albert Rilliard, David Doukhan, Marie Tahon, Antoine Laurent

Évaluation perceptive de l'anticipation de la prise de parole lors d'interactions dialogiques en français

Rémi Uro[1,2], Albert Rilliard[2], David Doukhan[1], Marie Tahon[3], Antoine Laurent[3]
(1) Institut National de l'Audiovisuel, Paris, France.
(2) Université Paris Saclay, CNRS, LISN, France.
(3) LIUM, Le Mans Université, France.
{ruro,ddoukhan}@ina.fr albert.rilliard@lisn.fr,
{marie.tahon,antoine.laurent}@univ-lemans.fr

RÉSUMÉ
Cette étude présente un test perceptif évaluant les indices permettant la planification de la prise de parole lors d'interactions orales spontanées. Des Unités Inter-Pauses (IPU) ont été extraites de dialogues du corpus REPERE et annotées en terminalité. Afin de déterminer quels paramètres affectent les jugements de la possibilité de prendre la parole, les stimulus ont été présentés sous forme audio ou textuelle. Les participant·es devaient indiquer la possibilité de prendre la parole « Maintenant », « Bientôt » ou « Pas encore », à la fin des IPU tronqués de 0 à 3 mots prosodiques. Les participant·es sont moins susceptibles de prendre la parole pour les frontières non terminales en modalité audio que textuelle. La modalité audio permet également d'anticiper une fin de tour de parole au moins trois mots avant sa fin, tandis que la modalité textuelle permet moins d'anticipation. Ces résultats soutiennent l'importance des indices contenus dans la parole pour la planification des interactions dialogiques.

ABSTRACT
A perceptual evaluation of the anticipation of turn-taking in French dialogic interactions

This study presents a perceptual test evaluating the cues allowing for turn-taking planning in French spontaneous interactions. Inter-Pausal Units (IPUs) were extracted from dialogues from the REPERE corpus and annotated with regard to terminality. In order to determine which parameters affect the possibility of turn-taking, stimuli were presented with audio-only or text-only modality. Participants had to indicate whether they could take the floor "Now", "Soon" or "Wait" after an IPU with 0 to 3 prosodic words removed. Participants were less likely to take the floor for non-terminal boundaries with the audio modality than with the text one. The audio modality also allows for the anticipation of the end of a turn up to three words before its end, while the text modality allows for less anticipation. These results support the importance of speech cues for the planning of dialogic interactions.

MOTS-CLÉS : Tour de parole, analyse de conversation, évaluation perceptive, TRP.

KEYWORDS: Turn-taking, conversation analysis, perceptual evaluation, interruption, TRP.

1 Introduction

Lors des interactions parlées, la gestion des tours de parole est critique pour intervenir au moment adéquat : sans couper la parole de son interlocuteur et sans laisser trop de silence (Bosch *et al.*, 2005;

Stivers *et al.*, 2009; Levinson & Torreira, 2015). Il est fondamental pour cela d'anticiper ces moments pertinents de prise de parole (communément appelés TRP - Transition Relevance Places (Sacks *et al.*, 1974)) pour planifier son énoncé (Levinson, 2016). L'étude de ces phénomènes répond à un grand nombre d'enjeux théoriques et applicatifs : analyse conversationnelle, description de comportements des locuteur·ices et des différences culturelles, conception de systèmes automatiques de gestion du dialogue pour les interactions Humain-Machine (Skantze, 2021).

Grosjean montre que la fin d'une phrase lue est prédictible grâce à des indices prosodiques (Grosjean, 1996). Les travaux de Magyari & de Ruiter (2012) sur des conversations téléphoniques en néerlandais démontrent la capacité des auditeur·ices à prédire si le tour de parole courant va se terminer ou continuer ; cette capacité augmente plus on s'approche de la fin du tour. Les pauses sont également un aspect important de l'analyse conversationnelle. Avec la syntaxe et la prosodie, les pauses fournissent des indices robustes pour permettre la détermination automatique de la fin d'un tour de parole (Christodoulides, 2018). Gotoh & Renals (2000) montre que l'analyse de la durée des pauses permet une meilleure définition des frontières de phrases (*"sentence boundaries"*) que des approches basées sur des modèles de langue, pour des contenus de médias audiovisuels anglophones. D'autres travaux, basé sur des conversations jouées ou des enregistrements téléphoniques, mettent en avant l'importance des indices visuels (Bi & Swerts, 2017) ou lexicaux (Hjalmarsson, 2011; Oliveira, 2008) dans la gestion des tours de parole. Les travaux de Gambi *et al.* (2015); De Ruiter *et al.* (2006) concluent à l'absence d'impact de l'intonation pour l'anticipation des fins de tour de parole. Ces différentes études travaillent sur des matériaux et des styles de parole divers (parole lue ou conversations téléphoniques, énoncés élicités), et dans différentes langues (anglais, néerlandais, français, etc.) dont certaines montrent des performances divergentes sur ces aspects (Grosjean, 1996). Il n'est donc pas clair quels sont les indices (prosodie, syntaxe, lexique, mouvements, etc.) les plus pertinents pour permettre des transitions fluides entre locuteur·ices en français.

Dans quelle mesure est-on capable d'anticiper le moment propice de prise de parole, et quels indices jouent dans cette décision ? Cet article présente une expérience visant à mieux comprendre entre les indices lexicaux et les indices transmis par le signal de parole, lesquels participent le plus à la capacité des auditeur·ices à anticiper les TRP, sur des énoncés de parole spontanée. Cette évaluation est fondée sur des segments de parole extraits semi-automatiquement de contenus de médias télévisuels présentant des interactions spontanées.

La Section 2 explique les processus de sélection et d'évaluation des données ainsi que le paradigme expérimental. Les résultats de perception sont en suite détaillés dans la Section 3 et discutés dans le contexte de la littérature en Section 4.

2 Méthode

2.1 Données

La tâche de perception envisagée consistait à présenter aux participant·es des unités de parole éventuellement tronquées de quelques mots à la fin, et de leur demander si, à la fin du stimulus, il leur semble possible de prendre la parole sans couper celle de leur interlocuteur·ice.

2.1.1 Sélection des unités

Les pauses étant des indices importants pour la segmentation de la parole, nous avons choisi d'utiliser des *Inter Pausal Units* (IPU, segment de parole entre deux pauses) –utilisées pour une variété de tâches d'analyse et de traitement de la parole (Levitan & Hirschberg, 2011; Prakash & Murthy, 2019; Bigi & Priego-Valverde, 2019)– comme unité de base pour cette étude.

Les IPU ont été extraites de REPERE (Giraudel *et al.*, 2012), corpus composé d'émissions de TV diffusées en France entre 2011 et 2012. Nous avons sélectionné uniquement les programmes des émissions *BFMStory*, *EntreLesLignes* et *CaVousRegarde*, qui présentent des interactions conversationnelles. Une détection automatique des pauses a été préférée à une annotation manuelle afin de limiter les biais humains lors de la sélection, et se rapprocher de conditions d'une tâche finale entièrement automatique. Pour cela, une segmentation en locuteur a été réalisée avec `LIUMSpkDiarization` (Meignier & Merlin, 2010) et les segments obtenus (en supprimant tous ceux contenant de la parole superposée) ont été transcrits avec le système du LIUM basé sur Kaldi (Povey *et al.*, 2011).

Les IPU ainsi obtenues sont les segments maximaux d'un·e même locuteur·ice entre deux pauses silencieuses, telles que prédites par `LIUMSpkDiarization`.

2.1.2 Annotation

Afin de proposer aux participant·es un ensemble varié d'IPU et d'éviter de leur faire évaluer le même extrait dans des conditions différentes, nous avons choisi de proposer aux participant·es deux IPU présentant les mêmes caractéristiques selon les facteurs contrôlés suivants : *Genre* du ou de la locuteur·ice (2 possibilités), *TRP* ou non à la fin de l'IPU (2 possibilités), *Modalité* de présentation du stimulus (2 possibilités, audio ou textuelle), et *Coupure* de 0 à 3 mots prosodiques (Nespor & Vogel, 2007; Wheeldon & Lahiri, 2002) à la fin de l'IPU (4 possibilités). Ainsi, un total de 64 (2 × Genre × TRP × Modalité × Coupure) IPU sont nécessaires.

Parmi les IPU obtenus automatiquement comme décrit ci-dessus, nous avons sélectionné un sous-ensemble d'IPU d'une durée de 6 à 12 secondes afin de limiter la complexité de l'étude. En raison du nombre restreint de femmes présentes dans le corpus REPERE, la durée maximale a été augmentée à 19 s pour les locutrices. Un certain nombre d'IPU a été retiré car présentant des questions ou des exclamations –exemples intéressants de TRP mais induisant un biais– ainsi que celles contenant des backchannels.

Les IPU sélectionnées ont été annotées par trois co-auteur·ices de cette étude, qui devaient pour chacune indiquer si elle faisait partie d'une conversation ou d'un monologue (e.g., présentation de nouvelles, discours, ...). Iels ont aussi annoté le type de frontière à la fin de chaque IPU pour différencier les fins terminales (présence de TRP) ou non-terminales (absence de TRP). Un total de 172 segments ont été annotés, résultant en un accord inter-annotateur de $0,70$ pour le dialogue et $0,73$ pour la terminalité (en utilisant le Kappa de Fleiss (Fleiss, 1971)) ce qui montre un accord substantiel pour ces deux tâches. Seuls les segments pour lesquels les trois annotateur·ices étaient d'accord ont été gardés. Les 64 IPU sélectionnées ont ensuite été traités manuellement afin de corriger la transcription automatique et déterminer les frontières des trois derniers mots prosodiques.

Des 32 IPU annotés comme terminales 22 apparaissaient en fin de tour dans l'émission originale et 10 apparaissaient à l'intérieur d'un tour de parole (i.e., la personne conserve la parole après). À l'inverse, sur les 32 IPU non terminales, 12 apparaissaient en fin de tour et 20 au sein d'un tour de

```
mais vous savez étant donné que les convocations sont attendues
étant donné que euh on a l' habitude de mettre la pression par
voie de presse à mon avis d' ici là on va | se | voir | souvent
```

FIGURE 1 – Exemple de transcription d'un énoncé utilisé dans l'expérience, les frontières de mots prosodiques sont représentées par le symbole « | »

parole. Du fait du faible nombre de femmes représentées dans le corpus REPERE, moins de locutrices différentes (20) que de locuteurs (27) sont présentes dans les stimulus.

La Figure 1 présente un exemple d'IPU transcrite avec les frontières de mots prosodiques de la fin représentées par le symbole « | ». Pour chaque IPU sélectionné, 8 versions différentes sont générées en découpant aux quatre positions différentes (0, 1, 2, ou 3 mots découpés) et en présentant dans les deux modalités (texte seul ou audio seul). Au final, un ensemble de 256 stimulus dont les caractéristiques sont présentées en Table 1, est utilisé durant le test.

TABLE 1 – Durée et nombre de mots des IPU sélectionnés

	Durée (s)	Nombre de mots
min. (s)	5.0	19
max. (s)	18.6	55
moy. (s)	8.9	34.7

Le nombre de syllabes de chaque énoncé (Table 2) a été calculé suivant la méthode décrite par Adda-Decker *et al.* (2005). Alors que Grosjean (1996) utilise des coupes de 3 syllabes dans une expérience similaire en français, nous observons une moyenne de 2 syllabes pour les coupes au niveau du mot prosodique effectuées sur les stimulus de notre étude.

2.2 Test de perception

2.2.1 Paradigme expérimental

Ce test utilise l'interface web PsyToolkit (Stoet, 2010, 2016), permettant la réalisation d'expériences et questionnaires en navigateur. Kochari (2019); Sasaki & Yamada (2019); Strickland & Stoops (2018), entre autres, montrent que des tests psychologiques classiques, effectués en ligne obtiennent des résultats comparables en condition de laboratoire, avec une même puissance de test statistique. Ainsi, l'utilisation d'une interface web a été préférée pour simplifier la tâche de recrutement de participant·es aux profils plus variés que ceux recrutés au laboratoire (Woods *et al.*, 2015).

TABLE 2 – Nombre de syllabes coupées

	Nb moyen de syllabe
IPU complet	53.4
1er mot prosodique	2.1
2ème mot prosodique	3.8
3ème mot prosodique	5.7

Après un court texte expliquant le cadre de l'étude et la durée estimée du test (20 min), il était demandé aux participant·es d'indiquer leur âge, genre et langue maternelle. L'expérience commençait par la modalité audio, supposée plus simple et motivante.

Pour chaque stimulus, les participant·es devaient soit écouter soit lire l'extrait. Les boutons de réponses apparaissaient à la fin de l'écoute pour la modalité audio et après la moitié de la durée de l'extrait audio pour la modalité textuelle. Les participant·es avaient ensuite 30 secondes pour indiquer si une prise de parole sans interrompre le tour courant était possible *Maintenant*, *Bientôt* ou *Pas encore*. Il leur était demandé de répondre le plus rapidement possible en suivant leur intuition. Le stimulus suivant démarrait après une seconde de pause, une fois la réponse soumise ou si les 30 secondes étaient écoulées.

Un lien vers l'interface de test en ligne a été envoyé à différentes listes de diffusion, incluant des communautés de recherche en informatique et en sociologie, et a été partagé sur des réseaux sociaux. Un total de 53 personnes francophones (29 s'identifiant comme femme, 21 comme homme et 3 comme autre), sans déficit visuel ou auditif non corrigé ont volontairement participé à l'expérience. Leurs âges varient de 20 à 63 ans, avec une moyenne de 35 ans. Cinq personnes ont indiqué avoir une langue maternelle autre que le français.

Soixante-quatre stimulus ont été présentés à chaque participant·e afin que chacun des 64 énoncés originaux soient évalués une et une seule fois, mais dans des conditions de présentations variées selon un design en carré latin mélangeant les facteurs contrôlés suivants : (i) 4 coupures de mot prosodiques (entre 0 aucun mot coupé et 3 derniers mots coupés), (ii) 2 genres, (iii) 2 modalités de présentation et (iv) IPU terminal/non-terminal. Les niveaux de ces quatre facteurs sont répartis selon quatre groupes de participant·es ; les participant·es sont attribués aléatoirement à l'un des quatre groupes.

2.2.2 Traitement statistique

Les variations de la proportion de chacune des trois réponses possibles (*Maintenant*, *Bientôt*, *Pas encore*) ont été modélisées en fonction des facteurs suivants : présence ou non de TRP, nombre de mots coupés (0 à 3), modalité de présentation (audio/texte) et genre du ou de la locuteur·ice (H/F) – regroupant ainsi ensemble les stimulus présentant les mêmes caractéristiques.

Les réponses sont analysées grâce à une régression polynomiale (Gries, 2021) à l'aide de la bibliothèque `R nnet`(Venables & Ripley, 2003), prenant la proportion de chaque catégorie de réponse (*Maintenant*, *Bientôt*, *Pas encore*) comme variable dépendante, et les variables *TRP*, *Coupure*, *Modalité* et *Genre* comme facteurs indépendants. Ces quatre facteurs indépendants et leurs interactions forment un modèle maximal qui est ensuite soumis à une procédure de simplification (Crawley, 2013) en supprimant itérativement les interactions d'ordre supérieur, tant que cela ne dégrade pas significativement le modèle. L'interaction quadruple et les interactions triples prenant en compte le genre, ainsi que les interactions doubles *(TRP × Genre)* et *(Modalité × Genre)* ont ainsi été supprimées. Le modèle minimal adéquat est ainsi basé sur les quatre facteurs principaux, quatre interactions doubles (*(TRP × Coupure) ; (TRP × Modalité) ; (Coupure × Modalité) ; (Coupure × Genre)*) et l'interaction triple *(TRP × Coupure × Modalité)*.

TABLE 3 – Table présentant la sortie du modèle minimal adéquat (Type III tests) présenté dans le texte : test du rapport de vraisemblance (LR χ^2), degrés de liberté (df), degrés de significativité (0.001 : '***' ; 0.01 : '**' ; 0.05 : '*').

Facteur	**LR χ^2**	**df**	**p**
TRP	214.23	2	***
Coupure	366.14	6	***
Modalité	62.71	2	***
Genre	18.29	2	***
(TRP × Coupure)	69.03	6	***
(TRP × Modalité)	74.46	2	***
(Coupure × Modalité)	14.61	6	*
(Coupure × Genre)	12.63	6	*
(TRP × Coupure × Modalité)	14.52	6	*

3 Résultats

Moins de $0,5\%$ des réponses n'ont pas été traitées parce que les participant·es avaient atteint la limite de 30 s de temps de réponse. Sur les réponses restantes, le modèle de régression montre un rôle significatif de tous les différents facteurs contrôlés lors de cette expérience. La table ANOVA correspondante est présentée dans la Table 3.

La réponse « Pas encore » est la plus fréquente (46%), suivie par « Bientôt » (32%) et enfin « Maintenant » (21%). Cela est lié au fait que les stimulus pour lesquels un changement de tour peut effectivement se produire sont en minorité, du fait des coupures de mots.

L'interaction entre TRP, Modalité et Coupure est décrite à la Figure 2. Cette interaction triple est présentée dans quatre graphiques de façon à montrer l'influence relative de chaque catégorie de réponse, sur les différentes coupures (en abscisse).

La réponse « Pas encore » est en effet prédominante, sauf pour quelques combinaisons de facteurs. Sa probabilité décroît à la coupure 0, pour toutes les combinaisons de modalité et de présence de TRP. Elle atteint ses plus bas niveaux pour les stimulus présentant un TRP, étant déjà plus faible que pour les stimulus sans TRP quel que soit le nombre de mots coupés. La probabilité de la réponse « Pas encore » dépend également de la modalité, montrant des variations plus importantes avec la modalité audio qu'avec le texte : elle est la plus forte, quelle que soit la position dans la phrase, pour les stimulus audio sans TRP, au-dessus de 60%, tandis qu'elle est la plus faible pour les stimulus audio avec TRP. Les présentations audio d'énoncés terminant par un TRP sont les seuls cas où la réponse « Pas encore » n'est jamais la plus probable : les participant·es ont répondu à plus de 50% « Bientôt » pour les coupures > 0, et « Maintenant » à 77% pour la coupure 0.

La réponse « Bientôt » est plutôt stable pour les stimulus sans TRP avec une probabilité autour de 25%, alors que la capacité d'anticipation change en fonction des mots coupés pour les stimulus avec TRP. Pour la modalité textuelle, sa probabilité est la plus forte pour un et deux mots coupés (autour de 40%, comparable aux réponses « Pas encore »), tandis que pour la modalité audio, il s'agit de la réponse la plus probable, autour de 50%, jusqu'aux énoncés complets (coupure=0 ; pour lesquels la réponse « maintenant » est choisie).

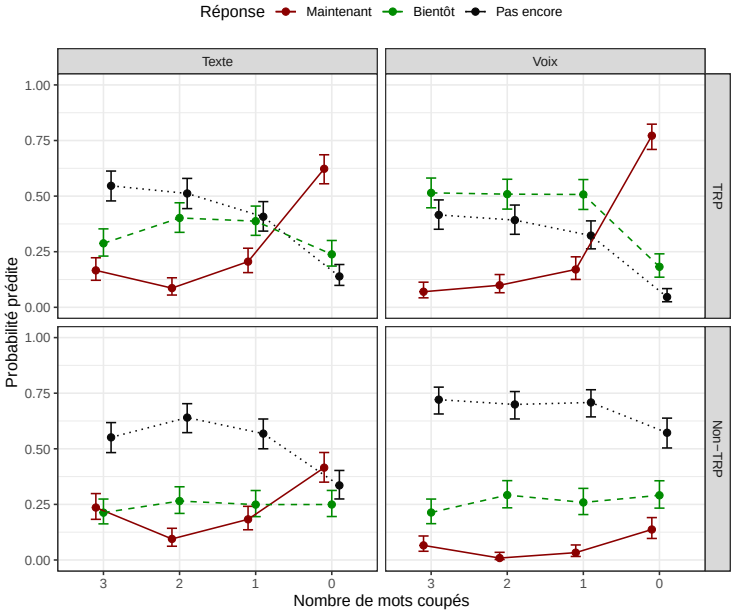

FIGURE 2 – Effet de la présence ou non de TRP (lignes), de la modalité de présentation (en colonne) et du nombre de mots coupés à la fin des IPU (abscisse) sur la probabilité des trois réponses possibles (« Maintenant » en rouge, « Bientôt » en vert, « Pas encore » en noir) estimée par le modèle.

La réponse « Maintenant » a un comportement opposé à la réponse « Pas encore », étant faible pour les coupures > 0 et augmentant pour les énoncés complets. Elle a aussi une probabilité plus élevée pour les énoncés complets avec TRP que sans, et surtout pour les énoncés en modalité audio : les stimulus complets audio sans TRP ont reçu 14% de réponse « Maintenant » contre 77% pour les ceux avec TRP. Pour la modalité textuelle, le taux de réponse passe de 42% pour les stimulus complets sans TRP à 62% pour ceux avec TRP.

Les inférences effectuées par les participants à propos de la terminalité des IPU sont plus tranchées pour les présentations orales que pour le texte. Avec la modalité textuelle, si la probabilité de la réponse « Maintenant » est plus faible pour les énoncés complets sans TRP qu'avec, il s'agit tout de même de la réponse la plus probable. À l'inverse, pour la modalité audio, les stimulus sans TRP ne sont presque jamais considérés comme pertinents pour une prise de parole, tandis que pour les stimulus avec TRP les participant·es anticipent la possibilité d'une prise de parole quel que soit le nombre de mots coupés (dans les limites de ce test).

L'interaction entre le nombre de mots coupés et le genre montre des différences à la coupure 0 (IPU complets), avec plus de réponses « Pas encore » et moins de « Maintenant » pour les énoncés produits par des hommes (18% « Pas encore » pour les femmes vs 28% pour les hommes, 53% « Maintenant » pour les femmes vs 45% pour les hommes). La probabilité de la réponse « Bientôt » augmente pour les femmes lorsque le nombre de mots coupés diminue. Cependant, ces observations sont les mêmes pour l'audio et le texte : il est possible que ce soit dû aux caractéristiques linguistiques et sémantique de ces énoncés plutôt qu'au genre, cette information n'étant pas disponible pour les stimulus textuels.

4 Discussion

L'effet principal sur la catégorie de réponse est lié à la triple interaction entre présence de TRP, nombre de mots coupés et modalité de présentation. S'il semble possible de décider qu'une IPU est terminée sur la base d'informations textuelles (hausse systématique des proportions « Maintenant » pour la coupure 0), distinguer entre IPU terminales ou non est bien mieux effectué si les participant·es ont accès aux informations de parole (proportion haute de « Bientôt » avant la fin des énoncés terminaux avant la coupure 0, puis réponse « Maintenant » très claire ; proportion élevée de « pas encore » pour les énoncés non terminaux, quelle que soit la coupure). Cela donne des arguments en faveur de l'importance des marques prosodiques pour la gestion du dialogue. Des indices de non-terminalité existent dans les deux modalités de présentation –il y a moins de réponses « Maintenant » pour les stimulus sans TRP dans les deux modalités– mais la présentation audio permet une meilleure distinction en fonction du nombre de mots coupés, sans confusion pour les énoncés complets : alors que les réponses « Maintenant » et « Pas encore » sont comparables pour les énoncés complets sans TRP présentés sous forme de texte, la réponse « Pas encore » n'est dominante que pour ceux présentés sous forme audio. Ainsi, les indices audio réduisent la probabilité de prendre la parole (réponses « Maintenant ») et augmentent la probabilité d'attendre (« Pas encore ») pour les séquences sans TRP.

À l'inverse, la prise de parole à la fin de séquences terminant par des TRP est plus probable pour les stimulus audio que textuels (62% de « Maintenant » pour le texte contre 77% pour l'audio). Les indices audio permettent aussi une meilleure anticipation : « Bientôt » n'est une réponse dominante que pour les stimulus audio, est l'est avant la fin de l'IPU. Les auditeur·ices peuvent prévoir au moins trois mots prosodiques (6 syllabes en moyenne) à l'avance s'il sera possible de prendre la parole. Ce résultat est cohérent avec les dynamiques de prise de parole présentée dans Levinson & Torreira (2015). Cette capacité d'anticipation n'est observée que pour la modalité audio.

L'augmentation de la réponse « Maintenant » est clairement liée aux énoncés avec 0 mots coupés, indiquant une forte capacité à déterminer qu'un énoncé est complet. Ceci est également observé pour la modalité textuelle, le gain permis par les indices audio est modeste (62% vs 77%). Les informations prosodiques auraient donc un rôle secondaire dans cette détermination. Elles sont par contre fondamentales pour l'anticipation –capacité nécessaire pour une prise de parole fluide–, résultat qui n'est pas reflété dans les réponses de ce test sur la seule base des contenus sémantiques. Nous montrons donc ici l'importance des indices audio pour le timing et l'efficacité de la gestion des tours de parole, sur des données dialogiques spontanées.

Si nous observons un effet du facteur *Genre* de la personne qui parle, le fait que cet effet soit visible aussi bien avec la modalité textuelle qu'audio pointe vers l'utilisation d'indices sémantiques ou syntaxiques plus que vers le genre perçu. Cette étude n'inclut pas un ensemble de locuteur·ices et de participant·es suffisamment large pour étudier l'impact des facteurs sociologiques tels que le rôle, le capital culturel ou la position sociale. Une étude multimodale incluant des indices visuels pourrait également être intéressante à l'avenir, le corpus d'origine étant issu de contenu télévisuel.

Remerciements

Ce travail a été partiellement financé par les projets ANR « Gender Equality Monitor » (ANR-19-CE38-0012) et ANR-DFG « La documentation automatique des langues à l'horizon 2025 » (CLD 2025, ANR-19-CE38-0015-04). Nous tenons à remercier les participant·es de cette étude.

Références

ADDA-DECKER M., BOULA DE MAREÜIL P., ADDA G. & LAMEL L. (2005). Investigating syllabic structures and their variation in spontaneous french. *Speech Communication*, **46**(2), 119–139. DOI : 10.1016/j.specom.2005.03.006.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

BI R. & SWERTS M. (2017). A perceptual study of how rapidly and accurately audiovisual cues to utterance-final boundaries can be interpreted in chinese and english. *Speech Communication*, **95**, 68–77. DOI : 10.1016/j.specom.2017.07.002.

BIGI B. & PRIEGO-VALVERDE B. (2019). Search for Inter-Pausal Units : application to Cheese ! corpus. In *9th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics*, p. 289–293, Poznań, Poland.

BOSCH L. T., OOSTDIJK N. & BOVES L. (2005). On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, **47**(1–2), 80–86. DOI : 10.1016/j.specom.2005.05.009.

CHRISTODOULIDES G. (2018). Acoustic correlates of prosodic boundaries in french a review of corpus data / correlatos acústicos de fronteiras prosódicas em francês : uma revisão de dados de corpora. *REVISTA DE ESTUDOS DA LINGUAGEM*, **26**(44), 1531–1549. DOI : 10.17851/2237-2083.26.4.1531-1549.

CRAWLEY M. J. (2013). *The R Book*. John Wiley & Sons, 2 édition.

DE RUITER J., MITTERER H. & ENFIELD N. (2006). Projecting the end of a speaker's turn : A cognitive cornerstone of conversation. *Language*, **82**, 515–535. DOI : 10.1353/lan.2006.0130.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

FLEISS J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **76**(5), 378–382. DOI : 10.1037/h0031619.

GAMBI C., JACHMANN T. & STAUDTE M. (2015). The role of prosody and gaze in turn-end anticipation. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

GIRAUDEL A., CARRÉ M., MAPELLI V., KAHN J., GALIBERT O. & QUINTARD L. (2012). The REPERE corpus : a multimodal corpus for person recognition. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 1102–1107, Istanbul, Turkey : European Language Resources Association (ELRA).

GOTOH Y. & RENALS S. (2000). Sentence boundary detection in broadcast speech transcripts. In *in Proc. of ISCA Workshop : Automatic Speech Recognition : Challenges for the new Millennium ASR-2000*, p. 228–235.

GRIES S. T. (2021). *Statistics for linguistics with R*. Mouton Textbook. Berlin, Germany : De Gruyter Mouton, 3 édition.

GROSJEAN F. (1996). Using prosody to predict the end of sentences in english and french : Normal and brain-damaged subjects. *Language and Cognitive Processes*, **11**(1–2), 107–134. DOI : 10.1080/016909696387231.

HJALMARSSON A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, **53**(1), 23–35. DOI : 10.1016/j.specom.2010.08.003.

KOCHARI A. R. (2019). Conducting web-based experiments for numerical cognition research. *Journal of Cognition*, **2**(1), 39. DOI : 10.5334/joc.85.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LEVINSON S. & TORREIRA F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, **6**. DOI : 10.3389/fpsyg.2015.00731.

LEVINSON S. C. (2016). Turn-taking in human communication – origins and implications for language processing. *Trends in Cognitive Sciences*, **20**(1), 6–14. DOI : 10.1016/j.tics.2015.10.010.

LEVITAN R. & HIRSCHBERG J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *INTERSPEECH*. DOI : 10.7916/D8V12D8F.

MAGYARI L. & DE RUITER J. (2012). Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in Psychology*, **3**.

MEIGNIER S. & MERLIN T. (2010). Lium spkdiarization : an open source toolkit for diarization. In *CMU SPUD Workshop*.

NESPOR M. & VOGEL I. (2007). *Prosodic Phonology*. DE GRUYTER. DOI : 10.1515/9783110977790.

OLIVEIRA M. (2008). Intonation as a cue to turn management in telephone and face-to-face interactions. *Speech Prosody*, p.4̃.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLIC˘EK P., QIAN Y., SCHWARZ P., SILOVSKY J., STEMMER G. & VESELY K. (2011). The kaldi speech recognition toolkit. p.4̃.

PRAKASH J. J. & MURTHY H. A. (2019). Analysis of inter-pausal units in indian languages and its application to text-to-speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **27**(10), 1616–1628. DOI : 10.1109/taslp.2019.2924534.

SACKS H., SCHEGLOFF E. A. & JEFFERSON G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, **50**(4), 696. DOI : 10.2307/412243.

SASAKI K. & YAMADA Y. (2019). Crowdsourcing visual perception experiments : a case of contrast threshold. *PeerJ*, **7**, e8339. DOI : 10.7717/peerj.8339.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SKANTZE G. (2021). Turn-taking in conversational systems and human-robot interaction : A review. *Computer Speech & Language*, **67**, 101178. DOI : 10.1016/j.csl.2020.101178.

STIVERS T., ENFIELD N. J., BROWN P., ENGLERT C., HAYASHI M., HEINEMANN T., HOYMANN G., ROSSANO F., DE RUITER J. P., YOON K.-E. & LEVINSON S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, **106**(26), 10587–10592. DOI : 10.1073/pnas.0903616106.

STOET G. (2010). PsyToolkit : A software package for programming psychological experiments using linux. *Behavior Research Methods*, **42**(4), 1096–1104. DOI : 10.3758/brm.42.4.1096.

STOET G. (2016). PsyToolkit. *Teaching of Psychology*, **44**(1), 24–31. DOI : 10.1177/0098628316677643.

STRICKLAND J. C. & STOOPS W. W. (2018). Feasibility, acceptability, and validity of crowdsourcing for collecting longitudinal alcohol use data. *Journal of the Experimental Analysis of Behavior*, **110**(1), 136–153. DOI : 10.1002/jeab.445.

VENABLES W. N. & RIPLEY B. D. (2003). *Modern applied statistics with S*. Statistics and Computing. New York, NY : Springer, 4 édition.

WHEELDON L. R. & LAHIRI A. (2002). The minimal unit of phonological encoding : prosodic or lexical word. *Cognition*, **85**(2), B31–B41. DOI : 10.1016/S0010-0277(02)00103-8.

WOODS A. T., VELASCO C., LEVITAN C. A., WAN X. & SPENCE C. (2015). Conducting perception research over the internet : a tutorial review. *PeerJ*, **3**, e1058. DOI : 10.7717/peerj.1058.