Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

VoiceStick : un corpus de parole spontanée pour le guidage vocal de drones

Allan Henry[1, 2, 3, 4] Solange Rossato[1, 2] Christian Graff[1, 3]
Sylvain Huet[1, 4] Jose-Ernesto Gomez-Balderas[1, 4]
(1) Univ. Grenoble Alpes, 38000 Grenoble, France
(2) LIG, 38000 Grenoble, France
(3) LPNC, 38000 Grenoble, France
(4) GIPSA-lab, 38000 Grenoble, France
prenom.nom@univ-grenoble-alpes.fr

RÉSUMÉ

Cet article introduit VoiceStick, le premier corpus francophone de parole spontanée dédié au guidage vocal de drones, comblant ainsi un manque notable dans les ressources pour l'interaction humain-robot en langue française. Constitué auprès de 29 binômes dans un paradigme asymétrique guide-pilote en réalité mixte, le corpus capture la dynamique naturelle d'une interaction spontanée. Totalisant 4 219 énoncés pour 19 829 mots, VoiceStick se distingue par une richesse lexicale témoignant de la liberté d'expression accordée aux locuteurs. Le corpus intègre un étiquetage dual corrélant transcriptions textuelles et commandes motrices réelles, permettant de distinguer l'intention sémantique de l'action pragmatique effective. Une expérimentation de référence via une architecture en cascade atteint 94 % d'exactitude sur les commandes explicites, tandis que la performance de 70 % sur la parole spontanée illustre la complexité des ambiguïtés inhérentes à ce type d'interaction.

ABSTRACT

VoiceStick : A Spontaneous Speech Corpus for Drone Voice Guidance

This paper introduces VoiceStick, the first French-language spontaneous speech corpus dedicated to voice-guided drone control, addressing a significant gap in human-robot interaction resources beyond English. Collected from 29 dyads using an asymmetric guide-pilot paradigm in mixed reality, the corpus captures the natural dynamics of spontaneous interaction. Comprising 4,219 utterances totaling 19,829 words, VoiceStick is characterized by a rich lexical diversity reflecting the freedom of expression granted to participants. The corpus features a dual-labeling scheme correlating textual transcriptions with actual motor commands recorded via joystick, enabling a distinction between semantic intent and pragmatic action. A cascade baseline achieves 94 % accuracy on explicit commands, while the 70 % performance on spontaneous speech highlights the complexity of the ambiguities inherent to this type of interaction.

MOTS-CLÉS : Parole spontanée, Interaction Humain-Robot (IHR), Corpus, Commande vocale, Pilotage de drone.

KEYWORDS: Spontaneous speech, Human-Robot Interaction (HRI), Corpus, Voice commands, Drone control.

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 702–714.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

1 Introduction

La parole spontanée se distingue fondamentalement de la parole lue ou préparée par la présence de phénomènes linguistiques complexes : hésitations, répétitions, reformulations, troncations et variations prosodiques constituent autant de marqueurs d'une communication naturelle et non contrainte (Shriberg, 2001; Bazillon *et al.*, 2008). Ces caractéristiques, bien documentées en linguistique de corpus (Blanche-Benveniste, 1997; Crible *et al.*, 2017), représentent un défi majeur pour les systèmes automatiques de traitement de la parole, conçus majoritairement sur des données lues et contrôlées.

Dans le domaine de l'interaction humain-robot (IHR), et plus particulièrement de l'interaction humain-drone (IHD), la commande vocale s'impose comme une modalité d'interaction naturelle et accessible, notamment pour des opérateurs non experts (Tezza & Andujar, 2019; Wang *et al.*, 2024). Cependant, la quasi-totalité des systèmes de contrôle vocal pour drones reposent sur des vocabulaires prédéfinis et des grammaires rigides (Thomas *et al.*, 2019; Choutri *et al.*, 2022; Rajapaksha *et al.*, 2019), contraignant l'utilisateur à employer des formulations formatées telles que "décoller", "atterrir" ou "avancer". Cette approche ignore délibérément la richesse de la parole naturelle au profit d'une robustesse technique, créant un fossé entre la communication humaine authentique et les interfaces disponibles.

Des travaux récents explorent l'utilisation de la parole spontanée pour le guidage de drones (Atanov *et al.*, 2025), mais se heurtent à un verrou fondamental : l'absence de corpus adaptés. Des ressources anglophones telles que SLURP (Bastianelli *et al.*, 2020) existent, mais se limitent à des assistants domestiques en conditions contrôlées, sans la spontanéité d'une téléopération réelle. La définition de ce qui constitue une "commande vocale" varie considérablement selon les études (Choutri *et al.*, 2022; Fayjie *et al.*, 2017; Contreras *et al.*, 2020; Thomas *et al.*, 2019; Yapicioglu *et al.*, 2021), mais les ressources disponibles se limitent majoritairement à des mots isolés hors contexte (Warden, 2018; Poirier *et al.*, 2023), sans hésitations, sans reformulations, sans la prosodie d'urgence caractéristique d'une situation de téléopération réelle. Cette limite dépasse le domaine des drones : les corpus de commandes vocales pour robots mobiles (Contreras *et al.*, 2020; Poirier *et al.*, 2023) et assistants domestiques (Lugosch *et al.*, 2019) partagent le même biais vers des vocabulaires contrôlés au détriment de la parole spontanée. Cette absence de données écologiquement valides constitue un obstacle tant pour la recherche en TAL que pour le développement de systèmes robustes.

À ce verrou s'ajoute un déséquilibre linguistique marqué : la littérature s'appuie quasi exclusivement sur des données anglophones (Warden, 2018; Poirier *et al.*, 2023), laissant le français largement sous-représenté dans ce domaine applicatif malgré un besoin croissant (Poirier *et al.*, 2023). Or les spécificités phonétiques, syntaxiques et prosodiques du français spontané ne peuvent être capturées par des modèles entraînés sur d'autres langues.

Pour combler ce manque, nous présentons *VoiceStick*, un corpus de parole spontanée en français dédié à l'interaction vocale humain-drone. Collecté auprès de 29 binômes naïfs dans un paradigme de guidage guide-pilote en réalité mixte, il constitue à notre connaissance le premier corpus francophone de commandes vocales spontanées pour la téléopération de drone. Nos contributions sont : (1) le protocole de collecte et le paradigme expérimental, (2) une analyse linguistique de la richesse et variabilité de la parole spontanée, et (3) une annotation sémantique et pragmatique des commandes pour la robotique francophone. Dans une démarche de science ouverte, ce corpus est mis à disposition de la communauté sous licence Creative Commons Attribution Non Commercial 4.0 International (CC BY-NC 4.0) [1].

1. https://zenodo.org/records/19882638 — DOI : 10.5281/zenodo.19882638

2 Matériel et méthodes

2.1 Description de l'expérience

Le protocole repose sur un cadre défini par une asymétrie informationnelle entre un expert (le guide), détenteur des informations environnementales liées à la cible, et un opérateur de terrain (le pilote), chargé de la navigation. La mise en œuvre de ce paradigme s'appuie sur une infrastructure de capture de mouvement, assurant un suivi de la position et de l'orientation du drone via des marqueurs optiques. Ce dispositif ne vise pas à reproduire un système de commande opérationnel, mais à créer les conditions d'émergence d'une parole spontanée ancrée dans l'action, difficilement atteignable par d'autres méthodes telles que la simulation ou la lecture de scripts préétablis.

Au sein de ce dispositif, le pilote assure le contrôle d'un drone (DJI® Tello™) en vol réel au moyen d'une seule manette Xbox® One. L'interface permet l'exécution de sept types de manœuvres réparties sur trois axes de contrôle : le joystick gauche assure les translations longitudinales (avant/arrière) et latérales (gauche/droite), les gâchettes contrôlent les translations verticales (montée/descente), et les boutons de tranche (bumpers) pilotent les rotations en lacet (gauche/droite). L'accès à la vue réelle de la caméra embarquée sur moniteur constitue ici la source d'information environnementale principale du pilote. L'intégration de ce retour visuel s'est avérée indispensable suite à des essais préliminaires ayant révélé que les sujets néophytes, en l'absence de boucle de rétroaction visuelle, éprouvaient des difficultés majeures à valider l'exécution de leurs commandes et à stabiliser la trajectoire de l'appareil face aux directives vocales.

Le guide, situé dans une pièce séparée (cf. Salle B, Figure 1), observait la scène à travers un modèle 3D de l'environnement selon le point de vue du drone, comme s'il utilisait une caméra embarquée virtuelle, dans laquelle la position de la cible virtuelle était visible. La communication avec le pilote se faisait exclusivement par instructions vocales, à l'aide d'un microphone (Ultravoice XM8500), afin de le guider vers la cible qui n'était pas visible pour lui. L'isolation acoustique entre les deux salles garantissait l'absence de bruit parasite lié au drone dans les enregistrements, assurant ainsi une qualité audio optimale des données collectées.

Cette expérience est intentionnellement unidirectionnel : seul le guide a la possibilité de parler avec le pilote. En effet, le retour destiné au guide est de nature visuelle et motrice, le déplacement effectif du drone sur son écran constitue la confirmation que l'instruction a été comprise et exécutée. L'absence de mouvement signale au contraire une incompréhension ou une ambiguïté, rendant un retour verbal du pilote redondant dans ce paradigme.

La cible consistait en une sphère colorée d'un rayon de 15 cm, positionnée aléatoirement dans l'une des 27 cellules d'une grille de $3\times3\times3$. Les cellules étaient espacées de 135 cm le long des axes horizontaux et réparties sur trois plans horizontaux situés à des hauteurs de 70 cm, 110 cm et 150 cm. Le pilote disposait d'un maximum de 120 secondes pour atteindre chaque cible. Une fois la cible atteinte, ou le temps imparti écoulé, une nouvelle cible était générée sans repositionnement du drone, suite à une pression sur une touche par le guide. Le protocole complet consistait en l'atteinte successive de six cibles par bloc, pour un total de deux blocs.

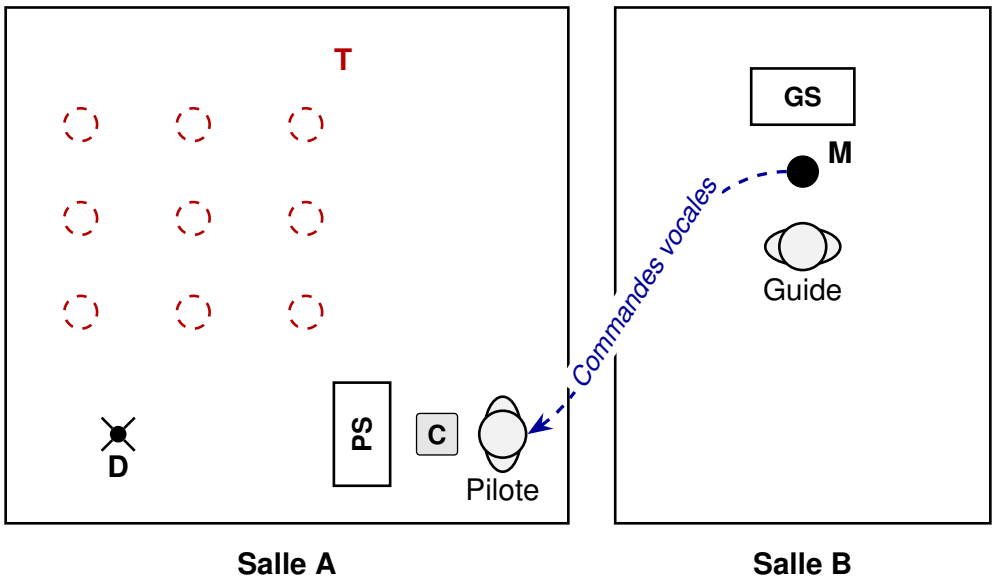

FIGURE 1 – Schéma du dispositif expérimental. Le drone (**D**) évolue dans la salle du pilote vers des cibles (**T**) invisibles pour celui-ci. Le pilote contrôle le drone au moyen d'une manette (**C**), guidé par les commandes vocales du guide transmises via un casque. Situé dans une pièce séparée, le guide observe la scène via une vue virtuelle (**GS**) et communique par microphone (**M**). Le pilote a uniquement accès à la vue réelle de la caméra embarquée du drone (**PS**). Dans la salle A, les neuf positions de cibles (**T**) sont disposées dans le plan horizontal (axes X/Y), tandis que leurs hauteurs varient selon l'axe vertical (Z).

2.2 Participants

L'étude a été menée auprès de 29 binômes, tous étudiants en psychologie, ayant déclaré une maîtrise complète de la langue française. Ce choix populationnel répond à des critères d'accessibilité et de disponibilité ; de plus, contrairement à des populations issues de formations techniques, ces participants présentaient une probabilité moindre d'avoir une expérience préalable du pilotage de drone, limitant ainsi les biais liés à une familiarité technique avec l'objet. Les participants (46 femmes, 12 hommes ; âge moyen : 20.1 ans, ET = 1.99) ont été recrutés individuellement et répartis en dyades selon les créneaux réservés, sans consigne de venir accompagné. Les rôles ont été assignés de manière aléatoire et maintenus tout au long de l'expérience. Dans la grande majorité des cas, les membres d'un binôme ne se connaissaient pas préalablement, ce qui limite les biais liés à une familiarité interpersonnelle susceptible de modifier les stratégies de communication spontanée. Aucune phase d'entraînement préalable n'a été proposée : ce choix délibéré visait à préserver la naturalité et la spontanéité des échanges, en évitant toute habituation à un vocabulaire ou à des formulations stéréotypées. Aucune contrainte lexicale ni liste de commandes préétablies n'a été imposée aux participants, favorisant ainsi l'émergence de phénomènes linguistiques propres à la parole spontanée, tels que les disfluences, les répétitions ou les autocorrections.

Le consentement éclairé de chaque participant a été recueilli individuellement avant le début de la session expérimentale, par le biais d'un formulaire de consentement dédié. Cette étude a été conduite en conformité avec le Règlement Général sur la Protection des Données (RGPD) et a fait l'objet d'un enregistrement auprès du délégué à la protection des données à l'Université Grenoble Alpes (numéro d'enregistrement : LPNC_RECH_5_RHS_HDS). À l'issue de l'expérience, les participants ont

complété un questionnaire visant à recueillir leurs impressions. Ce questionnaire portait notamment sur leur niveau de confort, leur perception de l'efficacité de la communication ou encore sur les difficultés rencontrées lors des différentes phases de la tâche.

En contrepartie de leur participation, les étudiants ont reçu des bons de crédits de cours, convertibles en points bonus pour leur examen. La participation à de telles études constitue par ailleurs une composante pédagogique reconnue de leur cursus.

3 Constitution du corpus *VoiceStick*

3.1 Prétraitement et annotation des données orales

Pour la conception de notre corpus, il est nécessaire de disposer d'un système capable de segmenter et de transcrire les enregistrements audio. Ces opérations permettent d'associer chaque segment de parole aux commandes effectuées sur la manette, et par conséquent aux mouvements réellement effectués par le drone. Pour la segmentation, nous avons d'abord utilisé WhisperX *(version large-v2)*, qui assure également la transcription des énoncés (Bain *et al.*, 2023). Nous avons comparé les performances *(Diarization Error Rate (DER))* de ce système avec un autre système de segmentation, PyAnnote *(version 2.1)* (Bredin *et al.*, 2019). La transcription est réalisée soit avec WhisperX, soit directement avec Whisper *(version large-v2)* pour chaque énoncé segmenté via PyAnnote. La seule différence réside dans la segmentation, étant donné que les modèles Whisper *large* sont utilisés dans les deux cas.

Afin d'évaluer ces outils, deux annotateurs ont réalisé indépendamment la correction des segmentations et des transcriptions automatiques sur 10 binômes. Les corrections manuelles ont été réalisées à l'aide du logiciel Praat (Boersma & van Heuven, 2001), retenu pour sa capacité à visualiser simultanément le signal acoustique, le spectrogramme et les annotations temporelles, ce qui facilite l'ajustement précis des frontières de segments et la vérification des transcriptions lexicales. Les transcriptions produites sont de nature lexicale et non phonétique : seul le contenu verbal est transcrit, sans notation des phénomènes sub-lexicaux tels que les allongements ou les qualités vocaliques. L'analyse des divergences a révélé un niveau d'accord élevé sur les transcriptions (WER inter-annotateurs : 6.12 %), tandis que les désaccords de segmentation se limitaient à des ajustements de frontières inférieurs au seuil de tolérance de 200 ms dans 76 % des cas (accord bidirectionnel). Les cas de divergences substantielles ont été résolus par révision du premier annotateur, qui disposait d'une expertise plus approfondie du protocole expérimental.

Conformément aux standards de la littérature, une marge de tolérance de 200 ms a été appliquée lors de l'évaluation des outils de segmentation (Bain *et al.*, 2023).

L'évaluation de la transcription indique un WER de 6,64 % pour WhisperX, contre 10,42 % pour PyAnnote + Whisper. Cette différence s'explique par les segments de parole fournis en entrée. En effet, WhisperX produit une segmentation qui diffère de celle de PyAnnote. En revanche, l'évaluation de la diarisation révèle une différence beaucoup plus marquée entre les deux systèmes. Le *DER* atteint 165,56 % pour WhisperX, contre seulement 29,63 % pour PyAnnote. Dans les deux cas, le taux d'omissions est faible (0,99 % pour WhisperX et 1,25 % pour PyAnnote), ce qui indique que la quasi-totalité des occurrences de parole présentes dans les données de référence sont correctement détectées. Cependant, WhisperX présente un taux de fausses alarmes (*FA*) extrêmement élevé (164,56 %), ce qui

signifie qu'il détecte de la parole dans de nombreuses zones de silence, entraînant une surestimation significative de la durée de parole. À l'inverse, PyAnnote maintient un taux de *FA* beaucoup plus limité (28,38 %).

En conclusion, bien que WhisperX obtienne un WER plus faible, ses performances en détection de segments de parole sont fortement dégradées par un excès de *FA*. La combinaison PyAnnote + Whisper offre une segmentation bien plus fiable. Pour la suite de nos analyses, nous retenons donc le couple PyAnnote + Whisper, qui constitue un meilleur compromis entre qualité de transcription et segmentation temporelle.

3.2 Commandes vocales du guide

À l'issue du traitement détaillé à la Section 3.1, nous avons obtenu 4 219 segments de commandes vocales, totalisant 19 829 mots. La distribution de la durée des segments présente une forte concentration entre 0 et 4 secondes. Les segments plus longs sont beaucoup plus rares et résultent souvent d'une segmentation imparfaite, au cours de laquelle des instructions consécutives n'ont pas été correctement séparées. Dans l'ensemble, la durée des segments s'étend de 0,12 s à 14,82 s (M = 2,04 s, ET = 1,85 s). Le nombre de mots par segment varie de 1 à 45 (M = 5,93, ET = 4,87). En termes de richesse lexicale, le corpus contient 669 mots uniques formant un total de 2 421 commandes vocales distinctes, ce qui souligne la diversité et la spontanéité du langage recueilli au cours de l'expérience.

Cette diversité lexicale témoigne de la liberté accordée aux locuteurs. À titre d'exemple, l'instruction de mouvement vers le haut ne se limite pas au verbe *« monte »* ; elle se décline en de nombreuses variations sémantiques et pragmatiques telles que *« prend un peu de hauteur »*, *« encore un peu plus haut »* ou des formulations plus imagées comme *« survole l'objectif »*. Cette variabilité constitue un défi majeur pour les systèmes de compréhension du langage naturel classiques et justifie l'intérêt d'un tel corpus pour la communauté.

3.3 De l'activité motrice à l'étiquetage pragmatique

Parallèlement à la capture des données orales, le protocole intègre l'enregistrement continu des entrées de la manette, reflétant l'activité motrice du pilote en réponse aux instructions. La Figure 2 illustre la complexité de ces signaux bruts pour l'énoncé *« Ok à ta gauche maintenant »*. On y observe une phase initiale de rotation à droite (attribuable à une erreur de manipulation ou à la poursuite d'une action antérieure) suivie d'une série d'impulsions vers la gauche.

Ce type de divergence met en lumière une caractéristique fondamentale de notre corpus : la nécessité de distinguer l'intention sémantique (ce qui est dit) de l'action pragmatique (ce qui est fait). Pour traiter cette dualité, nous avons instauré un protocole d'étiquetage dual. Chaque segment audio est ainsi associé à deux niveaux de vérité terrain : une étiquette sémantique, issue de la transcription manuelle, et une étiquette pragmatique, dérivée du comportement effectif enregistré sur la manette.

Afin d'extraire une action pragmatique unique par énoncé, nous appliquons une règle de sélection basée sur la direction dominante en termes de durée cumulée. Cette mesure est effectuée dans l'intervalle temporel s'étendant du début de l'énoncé (instant T) jusqu'à l'amorce de l'instruction suivante (instant $T + 1$). Dans l'exemple de la Figure 2, les trois premières impulsions sont retenues, tandis que la dernière, débutant après $T + 1$, est écartée. La rotation à gauche est alors validée

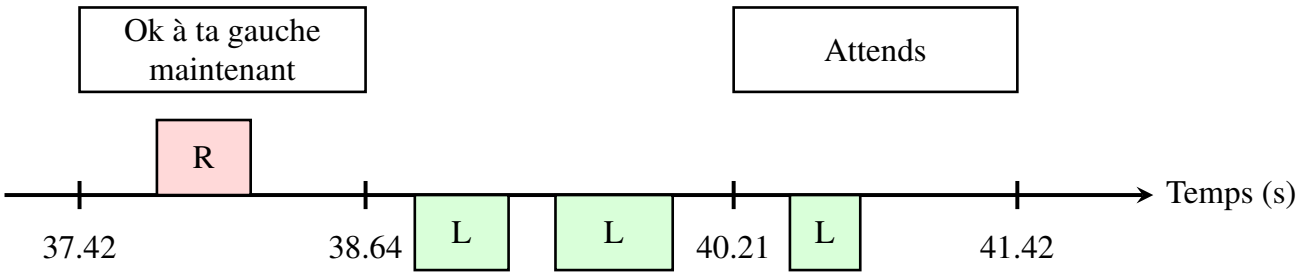

FIGURE 2 – Exemple d'une séquence de commandes de la manette enregistrées à la suite d'une instruction vocale. La commande en rouge représente une rotation à droite, et les commandes en vert représentent une rotation à gauche.

comme l'action principale associée à l'unité de parole. Ce critère a été retenu car la durée reflète l'engagement moteur effectif du pilote, contrairement à un comptage d'impulsions qui pondèrerait identiquement corrections brèves et maintiens prolongés, ou à une règle de première action, trop sensible aux artefacts de transition. Dans les cas où aucune activité motrice n'est détectée durant cet intervalle, une situation attendue, par exemple, suite à une consigne d'arrêt tel que *« Stop »*, l'action est alors étiquetée *« None »*, signalant l'absence de mouvement.

Ce processus de filtrage permet de réduire les 11 189 impulsions brutes à 4 219 unités d'action cohérentes. Cette simplification repose sur le constat que, dans un contexte de guidage spontané, les locuteurs privilégient majoritairement des instructions directionnelles uniques par segment de parole (Henry *et al.*, 2025).

3.4 Répartition et composition des jeux de données

Le corpus est structuré en trois sous-ensembles selon une répartition cible de 75 % pour l'entraînement (*train*), 15 % pour la validation (*validation*) et 10 % pour le test (*test*). Le jeu de test est exclusivement constitué de fichiers ayant fait l'objet d'une annotation manuelle, garantissant ainsi une évaluation finale sur une vérité terrain de haute précision.

La répartition détaillée des données par classe est présentée dans le Tableau 1. On y observe une forte prédominance de la classe *forward*, traduisant la fréquence naturelle de cette instruction dans les scénarios de guidage. À l'inverse, les commandes *backward*, *left* et *right* sont nettement moins représentées, ce qui reflète la tendance des participants à privilégier les mouvements vers l'avant, les pivots et les déplacements verticaux. Enfin, la standardisation du protocole de collecte permet d'envisager un accroissement futur du corpus tout en maintenant une structure de données cohérente. Compte tenu du déséquilibre distributionnel entre les translations latérales (*left*/*right*) et les rotations en lacet (*yaw left*/*yaw right*), et de leur proximité lexicale dans les instructions vocales spontanées, ces quatre classes sont fusionnées en deux classes unifiées (*left* et *right*) pour l'ensemble des expérimentations présentées dans ce travail, réduisant ainsi l'espace de sortie à sept classes.

	Forw.	Back.	Up	Down	Left	Right	Y. left	Y. right	None	Total
Train	868	181	245	260	138	106	435	386	587	3 206
Valid.	171	40	39	53	19	10	67	88	114	601
Test	98	11	37	35	23	13	64	37	94	412

TABLE 1 – Nombre d'énoncés par classe pour les jeux d'entraînement, de validation et de test.

3.5 Distribution des énoncés de guidage explicite selon les commandes

Afin de constituer une référence sémantique, nous avons extrait du corpus les énoncés contenant des marqueurs lexicaux explicites, appariés aux catégories de mouvement correspondantes selon la nomenclature suivante :

gauche	→ left	*avant, avance, devant, tout droit*	→ forward
droite	→ right	*recule, derrière, arrière*	→ backward
haut, monte	→ up	*stop, arrête*	→ none
bas, descend	→ down		

Cette procédure nous permet de sélectionner 1 117 énoncés distinct issus des jeux d'entraînement et de validation, ce qui représente environ un quart du corpus total. À partir de ces données, nous avons comparé la catégorisation obtenue par mots-clés avec celle dérivée des mouvements réellement effectués par le drone (cf. Figure 3). Avec une précision globale (*accuracy*) de 79 %, notre méthode d'association témoigne d'une cohérence satisfaisante entre le langage et l'action. Néanmoins, certaines erreurs subsistent, principalement dues à des imprécisions résiduelles de la segmentation temporelle : les commandes étant extraites entre le début de l'énoncé T et le début de l'énoncé $T+1$, des actions initiées à l'instant $T-1$ peuvent encore être enregistrées.

Le taux de correspondance de 79 % entre les mots-clés explicites et les actions réelles souligne la complexité du langage spontané. Les 21 % de divergence ne proviennent pas uniquement de biais de segmentation, mais illustrent également la nature même de la tâche : un pilote peut anticiper une commande, en corriger une autre ou interpréter une instruction relative à son propre référentiel spatial. Ce phénomène rejoint les observations de (Raux & Nakano, 2010) sur les corrections d'action en interaction située : les participants ajustent fréquemment leurs actions en cours d'exécution face à des instructions ambiguës ou tardives, ce qui se traduit ici par des divergences entre l'intention verbale du guide et l'action réalisée par le pilote. Ce corpus offre ainsi une base de données unique pour analyser ces phénomènes de désynchronisation entre l'intention verbale et la réalisation physique.

4 Expérimentation : Validation de l'utilisabilité du corpus

L'objectif de cette section n'est pas de proposer une architecture novatrice, mais de fournir un point de référence à l'aide d'outils standards afin de démontrer que le corpus *VoiceStick* est exploitable pour des tâches de compréhension automatique.

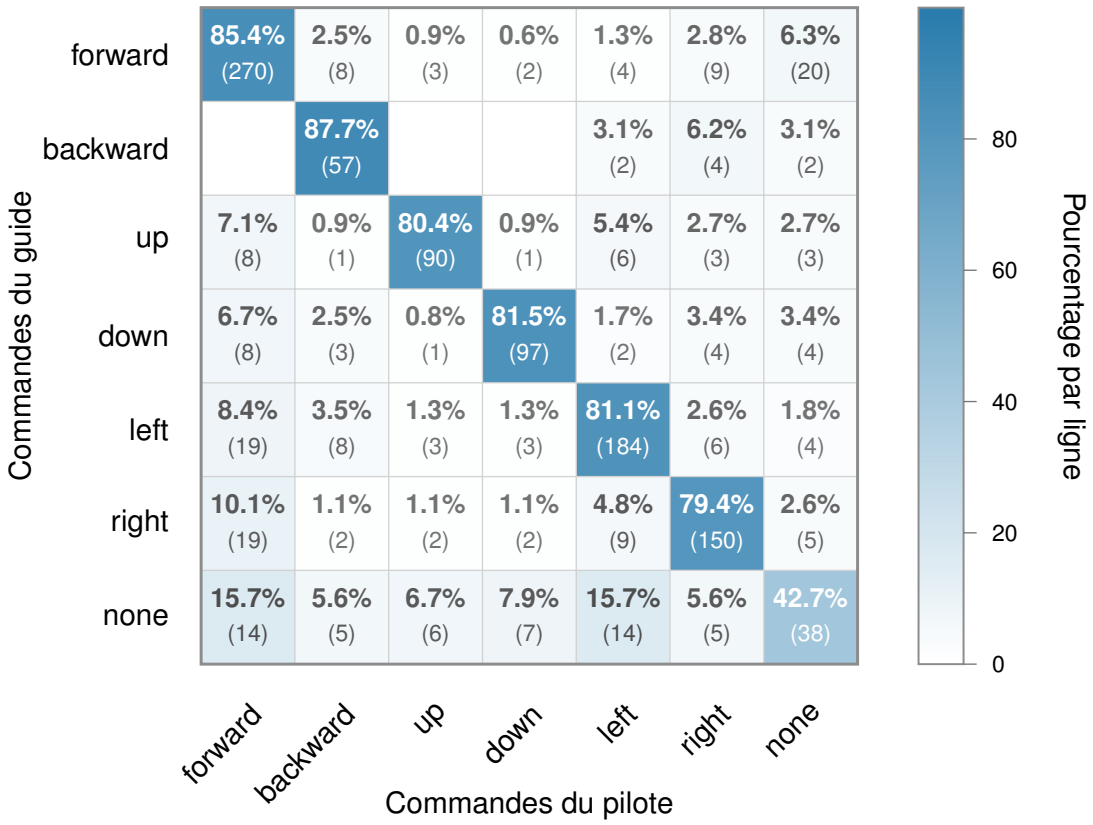

FIGURE 3 – Matrice de confusion illustrant la correspondance entre les commandes vocales explicites du guide (lignes) et les commandes effectivement exécutées par le pilote (colonnes). Les pourcentages indiquent le taux de correspondance pour chaque direction. La classe *none* regroupe les énoncés contenant des marqueurs d'arrêt explicites (*stop*, *arrête*), pour lesquels aucun mouvement n'est attendu. Le faible taux de correspondance observé (42,7 %) s'explique par le fait que le pilote peut poursuivre un mouvement amorcé avant de l'interrompre, introduisant une activité motrice résiduelle dans l'intervalle de mesure.

4.1 Architecture du système de référence

Nous utilisons un système de compréhension du langage parlé en cascade. La première étape assure la transcription de la commande vocale via le modèle Whisper. Le texte transcrit est ensuite injecté dans un classifieur d'intentions dont l'architecture est détaillée en Figure 4.

FIGURE 4 – Architecture générale du modèle de référence utilisé pour la validation.

Après une étape d'encodage en *Sentence Embeddings*, les représentations sont traitées par un réseau LSTM, suivi d'un mécanisme d'attention et d'une couche linéaire. Cette structure mémorielle est cruciale pour le traitement de la parole, car elle permet au modèle de prendre en compte le contexte des commandes précédentes pour interpréter des énoncés anaphoriques ou itératifs tels que *« encore »* ou *« continue un peu »*.

4.2 Configurations et prétraitement

Conformément à la simplification introduite en Section 3.4, nous retenons une configuration à sept classes comme système de référence.

Par ailleurs, une stratégie d'augmentation de données via le LLM Mistral AI (Jiang *et al.*, 2023) a été testée pour accroître la diversité lexicale de l'entraînement. Bien que ce processus ait généré plus de 9 000 paraphrases uniques, les gains de performance observés sont restés marginaux par rapport à un entraînement sur données réelles. En conséquence, nous retenons la configuration entraînée sur les données réelles comme référence pour l'évaluation de la ressource.

4.3 Résultats et capacité d'apprentissage

Les performances du système de référence ont été évaluées afin de quantifier la complexité linguistique du corpus *VoiceStick*. L'analyse met en évidence une forte capacité d'apprentissage du modèle sur le lexique de base, atteignant une exactitude de 94 % sur le sous-ensemble des commandes explicites. Ce résultat démontre que la ressource est exempte de bruit majeur pour les instructions univoques. En revanche, la performance globale s'établit à 70 % sur l'intégralité du jeu de test, reflétant le défi scientifique posé par la parole spontanée.

Une analyse qualitative des erreurs révèle que la majorité des divergences ne relève pas d'une défaillance structurelle du système, mais de l'ambiguïté intrinsèque des interactions humaines en situation réelle. Des énoncés tels que *« tourne »*, dépourvus de spécification directionnelle, imposent au pilote un choix arbitraire. Dans ces conditions, un désaccord entre la prédiction du modèle et l'action réelle du pilote ne traduit pas nécessairement une erreur de compréhension, mais plutôt une impossibilité de lever l'ambiguïté sémantique sans recours à un contexte visuel ou pragmatique plus large.

5 Discussion

Un choix méthodologique central de ce travail mérite d'être explicité : le refus délibéré du paradigme du Magicien d'Oz (*Wizard of Oz*, WoZ), dans lequel un opérateur humain caché simule les réponses d'un système automatique à l'insu du participant (Fraser & Gilbert, 1991). Notre protocole repose au contraire sur une configuration humain-humain transparente, dont le discours porte la trace : on observe ainsi des tournures de politesse telles que *« est-ce que tu peux te retourner ? »*, caractéristiques d'une interaction sociale ordinaire. Ce choix constitue une valeur ajoutée scientifique pour les travaux futurs : disposer de cette référence humain-humain permettra de mesurer en quoi le comportement langagier évolue lorsque le locuteur sait qu'il s'adresse directement à un système autonome (Riek, 2012).

La richesse lexicale du corpus témoigne de la liberté accordée aux locuteurs : avec 669 mots uniques et 2 421 commandes vocales distinctes, *VoiceStick* reflète la variabilité naturelle du langage spontané en situation d'action. Une même intention directionnelle peut ainsi se manifester sous des formes très diverses, allant de l'impératif direct *« monte »* à des formulations imagées comme *« survole l'objectif »*, en passant par des constructions modalisées telles que *« essaie d'aller un peu plus haut »*. Cette variabilité, documentée en linguistique de corpus (Blanche-Benveniste, 1997; Stoean, 2001), constitue

précisément le défi que les systèmes à vocabulaire contraint ne peuvent pas relever, et justifie l'intérêt d'une telle ressource pour la communauté TAL et robotique.

L'écart de performance entre les commandes explicites (94 %) et la parole spontanée (70 %) ne doit pas être interprété comme une limitation du système, mais comme un résultat scientifique à part entière : il quantifie le coût réel de la spontanéité pour les systèmes de compréhension automatique. Il illustre précisément pourquoi les approches reposant sur des vocabulaires prédéfinis (Choutri *et al.*, 2022; Rajapaksha *et al.*, 2019) échouent dès que le locuteur s'en écarte. Contrairement à ces ressources, *VoiceStick* capture la parole telle qu'elle émerge dans une téléopération réelle, et son étiquetage dual, corrélant transcription et commande motrice effective, permet de rendre ces divergences analysables plutôt que de les masquer derrière une vérité terrain unique.

6 Conclusion

Cet article présente *VoiceStick*, le premier corpus francophone de parole spontanée dédié au guidage vocal de drones. Par son protocole de collecte en conditions écologiques et son étiquetage dual corrélant intention sémantique et action pragmatique effective, il offre une ressource inédite pour la communauté TAL et robotique francophone.

L'expérimentation de référence valide l'exploitabilité du corpus pour des tâches de compréhension automatique, et l'écart entre commandes explicites (94 %) et parole spontanée (70 %) quantifie objectivement le défi scientifique qu'il pose. Plusieurs limites méritent toutefois d'être soulignées : la population, homogène et non experte, peut ne pas refléter la diversité des usages réels, et la taille actuelle du corpus reste modeste. Ces limites constituent autant de pistes pour des travaux futurs : l'extension à de nouveaux locuteurs et contextes, l'intégration d'indices prosodiques et paralinguistiques issus du signal acoustique brut, et le développement d'architectures de bout en bout exploitant directement ces caractéristiques sans recours à la transcription intermédiaire.

Dans une démarche de science ouverte, *VoiceStick* est disponible sous licence CC BY-NC 4.0, accompagné de ses transcriptions, annotations sémantiques et pragmatiques, fichiers audio segmentés et métadonnées.

Remerciements

Ce travail est soutenu par l'Agence nationale de la recherche dans le cadre du programme « Investissements d'avenir » (ANR-15-IDEX-02) et a été partiellement soutenu par ROBOTEX 2.0 (subventions ROBOTEX ANR-10-EQPX-44-01 et TIRREX ANR-21-ESRE-0015) financé par le programme français Investissements d'avenir. Ce travail a bénéficié de la collaboration des membres du projet SAMGuide (ANR-21-VE33-0011-01). Les auteurs tiennent à remercier tous les participants qui ont pris part à l'expérience de collecte de données.

Références

ATANOV S., MOLDAMURAT K., BAKYT M., ZINAGABDENOVA D., MOLDAMURAT A., ZHUMAZHANOV B. & MAIDANOV A. (2025). Intelligent voice control system for UAV with mobile robot. *Indonesian Journal of Electrical Engineering and Computer Science*, **38**(2), 1061. DOI : 10.11591/ijeecs.v38.i2.pp1061-1072.

BAIN M., HUH J., HAN T. & ZISSERMAN A. (2023). WhisperX : Time-Accurate Speech Transcription of Long-Form Audio. arXiv :2303.00747 [cs], DOI : 10.48550/arXiv.2303.00747.

BASTIANELLI E., VANZO A., SWIETOJANSKI P. & RIESER V. (2020). SLURP : A Spoken Language Understanding Resource Package. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 7252–7262, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.588.

BAZILLON T., JOUSSE V., BÉCHET F., LINARÈS G. & LUZZATI D. (2008). La parole spontanée : transcription et traitement. *ATALA (Association pour le Traitement Automatique des Langues)*.

BLANCHE-BENVENISTE C. (1997). *Approches de la langue parlee en francais*. Ophrys.

BOERSMA P. & VAN HEUVEN V. (2001). Speak and unSpeak with Praat. *Glot International*, **5**(9/10), 341–347.

BREDIN H., YIN R., CORIA J. M., GELLY G., KORSHUNOV P., LAVECHIN M., FUSTES D., TITEUX H., BOUAZIZ W. & GILL M.-P. (2019). pyannote.audio : neural building blocks for speaker diarization. arXiv :1911.01255 [eess], DOI : 10.48550/arXiv.1911.01255.

CHOUTRI K., LAGHA M., MESHOUL S., BATOUCHE M., KACEL Y. & MEBARKIA N. (2022). A Multi-Lingual Speech Recognition-Based Framework to Human-Drone Interaction. *Electronics*, **11**(12), 1829. DOI : 10.3390/electronics11121829.

CONTRERAS R., AYALA A. & CRUZ F. (2020). Unmanned Aerial Vehicle Control through Domain-Based Automatic Speech Recognition. *Computers*, **9**(3), 75. DOI : 10.3390/computers9030075.

CRIBLE L., DEGAND L. & GILQUIN G. (2017). The clustering of discourse markers and filled pauses : A corpus-based French-English study of (dis)fluency. *Languages in Contrast*, **17**(1), 69–95. DOI : 10.1075/lic.17.1.04cri.

FAYJIE A. R., RAMEZANI A., OUALID D. & LEE D. J. (2017). Voice enabled smart drone control. In *2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN)*, p. 119–121, Milan : IEEE. DOI : 10.1109/ICUFN.2017.7993759.

FRASER N. M. & GILBERT G. N. (1991). Simulating speech systems. *Computer Speech & Language*, **5**(1), 81–99. DOI : https ://doi.org/10.1016/0885-2308(91)90019-M.

HENRY A., GRAFF C., ROSSATO S., GOMEZ-BALDERAS J.-E. & HUET S. (2025). Voice Commands for Guidance to a 3D Position : To Collect Spontaneous Data. In *Proceedings of the 18th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, p. 410–411, Corfu Island Greece : ACM. DOI : 10.1145/3733155.3734918.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., CASAS D. D. L., BRESSAND F., LENGYEL G., LAMPLE G., SAULNIER L., LAVAUD L. R., LACHAUX M.-A., STOCK P., SCAO T. L., LAVRIL T., WANG T., LACROIX T. & SAYED W. E. (2023). Mistral 7B. arXiv :2310.06825 [cs], DOI : 10.48550/arXiv.2310.06825.

LUGOSCH L., RAVANELLI M., IGNOTO P., TOMAR V. S. & BENGIO Y. (2019). Speech Model Pre-Training for End-to-End Spoken Language Understanding. In *Interspeech 2019*, p. 814–818 : ISCA. DOI : 10.21437/Interspeech.2019-2396.

POIRIER S., CÔTÉ-ALLARD U., ROUTHIER F. & CAMPEAU-LECOURS A. (2023). Efficient Self-Attention Model for Speech Recognition-Based Assistive Robots Control. *Sensors*, **23**(13), 6056. DOI : 10.3390/s23136056.

RAJAPAKSHA S., ILLANKOON V., HALLOLUWA N. D., SATHARANA M. & UMAYANGANIE D. (2019). Responsive Drone Autopilot System for Uncertain Natural Language Commands. In *2019 International Conference on Advancements in Computing (ICAC)*, p. 232–237, Malabe, Sri Lanka : IEEE. DOI : 10.1109/ICAC49085.2019.9103346.

RAUX A. & NAKANO M. (2010). The Dynamics of Action Corrections in Situated Interaction. *11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 165–174.

RIEK L. (2012). Wizard of Oz Studies in HRI : A Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction*, p. 119–136. DOI : 10.5898/JHRI.1.1.Riek.

SHRIBERG E. (2001). To 'errrr' is human : ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, **31**(1), 153–169. DOI : 10.1017/S0025100301001128.

STOEAN C. S. (2001). LES ACTES DE LANGAGE DANS LE DISCOURS. THÉORIE ET FONCTIONNEMENT. *Armand Colin*.

TEZZA D. & ANDUJAR M. (2019). The State-of-the-Art of Human–Drone Interaction : A Survey. *IEEE Access*, **7**, 167438–167454. DOI : 10.1109/ACCESS.2019.2953900.

THOMAS C., JOSEPH THOMAS J., BHARADWAJ R., MONDAL A. K., DEVALLA V. & OMKAR S. N. (2019). Design and Development of an Android Application for Voice Control of Micro Unmanned Aerial Vehicles. In *AIAA Aviation 2019 Forum*, Dallas, Texas : American Institute of Aeronautics and Astronautics. DOI : 10.2514/6.2019-3363.

WANG T., ZHENG P., LI S. & WANG L. (2024). Multimodal Human–Robot Interaction for Human-Centric Smart Manufacturing : A Survey. *Advanced Intelligent Systems*, **6**(3), 2300359. DOI : 10.1002/aisy.202300359.

WARDEN P. (2018). Speech Commands : A Dataset for Limited-Vocabulary Speech Recognition. arXiv :1804.03209 [cs], DOI : 10.48550/arXiv.1804.03209.

YAPICIOGLU C., DOKUR Z. & OLMEZ T. (2021). Voice Command Recognition for Drone Control by Deep Neural Networks on Embedded System. In *2021 8th International Conference on Electrical and Electronics Engineering (ICEEE)*, p. 65–72, Antalya, Turkey : IEEE. DOI : 10.1109/ICEEE52452.2021.9415964.