# PréGDR TAL
# Axe de réflexion : Productions langagières

D. Lolive

juin 2018

## 1 Participants


— Aurélien Bossard, LIASD, Université Paris 8, aurelien.bossard@iut.univ-paris8.fr
— Florian Boudin, LS2N, Université de Nantes, florian.boudin@univ-nantes.fr
— Michael Filhol, LIMSI, CNRS, michael.filhol@limsi.fr
— Gwénolé Lecorvé, IRISA, Université de Rennes 1, gwenole.lecorve@irisa.fr
— Damien Lolive, IRISA, Université de Rennes 1, damien.lolive@irisa.fr
— François Portet, LIG, francois.portet@imag.fr


## 2 Thèmes et périmètre d'étude

Le texte, le geste et la parole permettent de produire des énoncés structurés par un langage dans le but de communiquer un contenu sémantique. Ce dernier peut être réalisé de différentes manières selon le média utilisé. Ces productions langagières peuvent être réalisées en combinant ces différentes modalités ainsi que possiblement d'autres supports (e.g. images, sons...). On peut noter qu'il existe des interactions fortes entre ces modalités. Par exemple, les émotions ne s'expriment pas seulement par la parole ou le geste mais résultent d'un processus cognitif complexe se traduisant par l'usage de différentes modalités. En conséquence, dans cet axe, nous tentons de gommer les frontières disciplinaires existantes pour faire ressortir les points communs qui existent sur le thème des productions langagières au sens large.

L'axe « productions langagières », décrit dans ce document, est en interaction avec d'autres axes du GDR. Notamment, des liens peuvent être faits tout au long du document avec les axes « multimodalité », « multilinguisme et multiplicité des langues », « ressources linguistiques » et « apprentissage ».

Parmi les applications principales des travaux concernés par l'axe « productions langagières », on peut citer :

— l'accessibilité : La loi de 2005 sur l'égalité des chances vise à favoriser l'accès aux lieux publics, au numérique pour les personnes sourdes, malentendantes ou muettes. La génération de la langue des signes, de la parole ou de texte peut servir dans ce sens. La synthèse de parole peut, par exemple, être utilisée pour assister des personnes lors d'une interaction ou bien en décrivant une scène. L'adaptation d'un texte ou sa simplification peut être appliquée dans ce contexte pour favoriser l'accessibilité à l'information, à certaines oeuvres, lieux, etc. ;
— les systèmes de dialogue interactifs, les assistants personnels, les chatbots : ce champs applicatif s'est beaucoup développé ces dernières années et la demande est croissante, autant du côté des industriels que des utilisateurs ;
— le résumé automatique : cela inclut les cas mono/multimédia, mono/multi-documents, multilingues/cross-langues ainsi que le résumé de documents de spécialité ou de jeux de données. Le résumé de documents de spécialités suppose un contexte et des connaissances bien définies ;
— la génération créative et/ou multimédia : L'industrie créative développe de plus en plus l'assistance technologique pour la génération de romans, d'histoires, de pièces musicales ou de scénarios de films.

Dans ces différentes applications, la personnalisation de la génération pour prendre en compte l'utilisateur, la langue, la modalité, la situation ou le flux du discours est nécessaire.

## 3 État des lieux

### 3.1 Des données vers le texte

La génération automatique de texte (GAT) ou Natural Language Generation (NLG) est le domaine d'application du traitement automatique du langage naturel (TALN) qui s'intéresse à la tâche de production d'un texte naturel qui

communique une information à une audience à partir de données fournies au système, principalement sous forme non linguistique. La tâche a initialement portée sur le cas idéal où l'entrée est une donnée sémantique formelle, puis s'est progressivement étendue à tout type d'entrées, d'où l'émergence du terme data-to-text. La complexité de cette tâche est due, d'une part, à la très grande variété des domaines d'application, des types d'entrées (signaux, statistiques, base de données), des buts communicationnels, ainsi que des informations pertinentes à communiquer (quoi dire ?), et, d'autre part, au manque de connaissances concernant la "meilleure" façon de réaliser un énoncé à partir d'une information donnée (comment dire ?). Pour un état de l'art de l'évolution de la GAT, le lecteur est invité à consulter [33, 26, 4, 9].

Si la première question (quoi dire ?) est surtout traitée en utilisant de la connaissance sur le domaine applicatif, la deuxième question (comment dire ?) fait souvent appel, outre la linguistique, à des connaissances sur le domaine. La conséquence est que dans les systèmes de GAT, il existe une tension entre les raisonnements purement linguistiques et ceux liés au domaine d'application. Par ailleurs, la GAT a connu un développement industriel important en regard de la communauté qui la constitue et de la diversité des applications visées (rapport, dialogue, article de journaux, génération créative et divertissante, génération multimédia, en-têtes, etc.). Enfin, il n'existe pas de théorie linguistique ou de modèle cognitif computationnel couvrant l'ensemble des sous-tâches de la GAT permettant de résoudre ces choix (quoi/comment dire ?). Il en a résulté un grand nombre d'approches et de systèmes dont peu sont transférables d'un domaine d'application à un autre. Il n'en reste pas moins que depuis les années 90 un pipeline « consensuel » des sous-tâches de la GAT a émergé [26] qui est instancié, parfois avec une grande liberté, dans la plupart des systèmes. Ce pipeline, surtout utilisé dans les approches symboliques (grammaires) est depuis début 2010 concurrencé par les approches dites de bout-en-bout (end-to-end).

La communauté GAT (NLG) est principalement structurée autour des conférences INLG (qui remplace ENLG depuis 2017, elle-même ayant remplacé EWNLG en 2005) organisées par le SIG-GEN de l'ACL. Des sessions NLG sont également régulièrement organisées dans les conférences de l'ACL ainsi qu'à COLING. Enfin, SigDial est également un lieu de croisement important des communautés dialogue et GAT.

La GAT est très largement dominée par les équipes de recherche anglo-saxonnes ciblant l'anglais. Quelques équipes francophones sont visibles à l'international. En France, on peut citer, sans prétention à l'exhaustivité, les travaux de l'équipe SYNALP (C. Gardent) du Loria, de l'ancienne équipe/UMR-INRIA Alpage (L. Danlos) de Rocquencourt, du LIS (anciennement LIF) à Marseille (M. Zock, mais qui restent anciens) et plus récemment du LIG Grenoble (F. Portet). D'autres acteurs francophones importants peuvent être trouvés hors métropole tels qu'au RALI (G. Lapalme) ou au DLT (F. Lareau) à Montréal. Les industriels en France s'intéressent également à la GAT, tels que Navers Labs (ex Xerox, M. Dymetman), Syllabs (`https://www.syllabs.com/`), Synapse (`http://www.synapse-developpement.fr/`) ou encore Yseop (`https://yseop.com/`). Il existe quelques formations dans lesquelles les étudiants sont initiés à la GAT telles que le master science du langage IDL de l'université Grenoble Alpes.

La tâche de GAT est présente dans de nombreuses autres applications du TALN (traduction automatique, résumé automatique, dialogue, questions-réponses) mais les approches sont différentes (avec souvent un usage réduit de la sémantique) de part le fait que les types des entrées-sorties sont figées ou que la sémantique est restreinte sur la tâche (dialogue). En particulier, la génération texte-à-texte, qui consiste à transformer un texte source en un texte cible, est une des problématiques historiques du TALN. Elle concerne un grand nombre d'applications comme la génération de paraphrases, la production de résumés, les chatbots ou la simplification de texte. L'arrivée récente des modèles neuronaux, et plus particulièrement des modèles séquence-à-séquence (entre autres [27]), couplée avec l'accès à de grandes quantités de données pour les entraîner, a permis de réaliser un saut qualitatif sans précédent dans le domaine.

Les travaux sur la génération texte-à-texte, dominés par les équipes de recherche américaines, portent essentiellement sur la langue anglaise qui est dotée de corpus de taille conséquente, dépassant le milliard de mots [12]. En France, plusieurs équipes de recherche travaillent sur ces tâches mais leur visibilité à l'international est plutôt restreinte. De manière non exhaustive, on retrouve la génération texte-à-texte dans les travaux des équipes ILES (LIMSI, Paris), TALEP (LIF, Marseille), LIASD (Paris), TALNE (LIA, Avignon) et TALN (LS2N, Nantes). On peut également mentionner l'existence de projet de recherche collaboratives relevant de la génération texte-à-texte comme RPM2 (2007-2010), ALECTOR (2006), ASADERA (2016), TREMoLo (2017-2021).

## 3.2 Langue des signes (LS)

En synthèse de langue des signes à partir d'entrées formalisées (nous excluons ici le rejeu de mouvements enregistrés sans abstraction linguistique), les premières avancées se sont produites dans les années 2000. Dans le projet collaboratif européen ViSiCAST en particulier, le consortium conçoit un signeur virtuel capable d'animer les descriptions en HamNoSys [16], une représentation formelle de signes lexicaux issue des travaux de Prillwitz et al. [24] à Hambourg, focalisée à l'origine sur les articulations manuelles. Un logiciel est alors disponible en Java et permet à tous de synthétiser des séquences de signes. Ceci assoit une première génération d'avatars signants, aux mouvements robotiques des bras mais couvrant l'ensemble des possibilités d'HamNoSys, et qui restera l'état de l'art pendant quelques années.

Depuis, les principaux verrous pour une synthèse acceptable restent les suivants :

— **multi-linéarité :** la LS met en jeu plusieurs articulateurs simultanément et dont la synchronisation peut être complexe, alors que l'approche majoritaire réduit la production globale à une séquence d'unités lexicales, supposant toujours une analogie avec la suite de mots formant le discours écrit. A priori, les signeurs virtuels ne présentent technologiquement aucune limite dans la synchronisation de leurs articulateurs (toutefois il peut en manquer, par ex. les muscles utiles à l'animation du visage sont nombreux). La limite se trouve dans les modèles. L'effort le plus avancé aujourd'hui pour permettre une représentation multi-linéaire des structures profondes des énoncés signés est AZee [5].

— **iconicité :** en LS, de nombreuses productions dites iconiques échappent aux lexiques de formes lemmatisées. On peut par exemple projeter des relations topologiques et des distances dans l'espace continu, ou prendre le rôle d'actants du discours sans aucune marque comparable aux unités lexicales. Sans les unifier toutes, des travaux existent s'attaquant aux structures iconiques de taille, de forme ou de positions/mouvements relatifs d'entités localisées dans l'espace [17]. Les structures hautement iconiques de prise de rôle (parfois appelés « transferts personnels ») sont encore absentes des modèles.

— **rendu de mouvements naturels :** la synthèse pure à partir d'unités minimales de descriptions produit généralement des mouvements à l'aspect robotique. Pour progresser, le domaine a besoin de modèles mathématiques sur les mouvements humains mis en jeu en LS (étude du système biomécanique et dynamique du mouvement). À ce jour, les mouvements synthétiques les plus naturels sont ceux de DePaul University [19]. Ces résultats sont principalement le fruit de processus appliqués à des formes de référence animés par des animateurs experts, ce qui ne permet toujours qu'une approche partielle des productions très iconiques.

Générer des signes à partir de textes suppose une traduction automatique, et les méthodes actuelles d'apprentissage fonctionnant pour la traduction texte-à-texte supposent toujours une séquence d'unités lexicales discrètes en sortie. Si certains groupes [7] s'y sont essayés en alignant les séquences de mots écrits avec des séquences classiques de gloses (étiquettes de signes) côté LS pour apprendre des modèles de traduction, les résultats dans ce domaine souffrent toujours de la limite indiquée plus haut des représentations en séquences.

La recherche en France se trouve aujourd'hui au centre des évolutions du traitement automatique de la LS puisqu'y sont hébergés les travaux sur AZee, modèle vers lequel se tournent les équipes d'animation graphique de signeurs virtuels : DePaul University d'abord [6], et le DFKI en Allemagne ensuite [22]. Des efforts sont également en cours sur les modèles biomécaniques en collaboration avec les sciences du mouvement [1]. On note aussi des travaux en animation au niveau phonétique et de la coarticulation (transition entre gestes motivés linguistiquement) à partir de capture de mouvement [11].

### 3.3 Du texte depuis et vers la parole

La reconnaissance automatique de la parole (RAP) vise à transcrire un discours oral vers une forme écrite. Il s'agit d'une tâche historique car elle ouvre la voie aux interactions humain-machine. La RAP implique principalement des travaux en acoustique afin de faire le lien entre signal de parole et phonologie (description des trames et modélisation temporelle), ainsi qu'en TAL afin de produire des modèles permettant de désambiguïser linguistiquement des hypothèses de transcriptions homophones. Historiquement, ces domaines ont été dominés par des approches statistiques [15] : modèles de Markov cachés, mélanges de gaussiennes, modèle n-grammes… Comme beaucoup de domaines, la RAP a subi une inflexion méthodologique du fait des réseaux de neurones. Ainsi, l'état de l'art en modélisations acoustique et linguistique est aujourd'hui dominé par des méthodes fondées sur des réseaux de neurones convolutionnels et/ou récurrents [2, 21, 13, 23]. La poursuite de ces travaux amène la communauté à étudier des solutions de bout-en-bout, transformant le signal brut en une séquence de mots via un unique modèle neuronal et sans réel moteur pour le décodage [18]. Les principaux challenges en RAP portent sur les points suivants : la production de modèles toujours plus robustes à des changements de contextes par rapport à leur apprentissage (bruits, locuteurs, thèmes, nouveaux mots…) ; le traitement des langues peu dotées (où les réseaux de neurones ne peuvent directement s'appliquer) ; et l'intégration de traitement linguistiques ou para-linguistiques en post-traitement voire au sein de la RAP (analyse syntaxique, nettoyage des disfluences, émotions, étiquetage en rôles sémantiques…

La synthèse de parole a pour rôle de produire un signal de parole à partir d'un texte fourni en entrée, éventuellement accompagné de consignes précisant la manière de réaliser le texte. Des travaux existent également sur la synthèse dite audio-visuelle qui se distingue par l'adjonction d'un avatar animé en sus du signal de parole. Une première étape consiste à extraire des informations linguistiques et prosodiques. De nombreux travaux se situent dans ce cadre et s'intéressent notamment à la prédiction des phonèmes et de l'intonation. Pour cela, de multiples autres analyses et traitements sont nécessaires, par exemple la standardisation du texte (réécriture d'abréviations, uniformisation de la convention orthographique…), l'analyse du sens de la phrase (détection de l'opinion, identification des mots porteurs d'expressivité…) ou encore de sa structure (positions où respirer, incises…). Il existe de nombreux travaux sur ces différentes tâches. Ces informations permettent d'alimenter la deuxième étape du processus de synthèse qui consiste en la construction du signal. Pour cela, les principales solutions se fondent sur la sélection d'unités [14], l'usage de modèles

paramétriques, ou la combinaison des deux [20, 28]. Le modèle utilisé pour la synthèse paramétrique peut prendre la forme d'un modèle de Markov caché [3] ou d'un réseau de neurones profond [30]. De manière récente des approches end-to-end reposant sur des réseaux de neurones, permettant d'éviter l'usage de connaissances expertes, sont apparues [29]. Cependant, les relations complexes qui permettent de passer du texte à la parole font que ces approches n'offrent pas encore une qualité suffisante. Les systèmes de synthèse fondés sur la sélection d'unités permettent d'obtenir une parole synthétisée de qualité mais dont l'expressivité est liée au jeu de données. Au contraire, les systèmes paramétriques offre une qualité moindre, limitée par la qualité du vocodeur utilisé, mais dont l'expressivité peut être contrôlée plus facilement.

Au niveau national, les activités ont diminué en RAP sur les 10 dernières années. Les laboratoires encore actifs sont principalement le LIMSI, le LIUM et le LORIA. En synthèse, de nombreuses équipes travaillent sur les étages amont comme la phonétique, la phonologie et la prosodie (e.g. ATILF, LLING, LLF, LPL, LPP). Pour la génération de parole proprement dite, la majorité des travaux sont conduits à l'étranger. Néanmoins, on peut citer l'IRCAM, l'IRISA, le GIPSA et le LORIA qui possèdent des activités en synthèse de parole, parfois audio-visuelle. Le paysage industriel national est relativement riche pour la RAP à travers des entreprises qui, outre la production de transcriptions, analysent le discours (par exemple, Vocapia, VoxPass, Allo-Media…) et plus limité en synthèse (principalement Voxygen). Au niveau international, les sociétés travaillant sur le sujet sont principalement les GAFAM mais il existe de multiples entreprises spécialisées (par exemple, Innoetics en synthèse de la parole).

L'activité en termes de projets de recherche collaborative est importante, notamment en France avec des projets financés par l'ANR comme GVLex (2008), VISAC (2009, synthèse audio-visuelle), Phorevox (2012), VERA (2012), ChaNTeR (2014), SynPaFlex (2015), ArtSpeech (2015), PASTEL (2016). Au niveau international, on peut également citer des projets comme Simple4All (FP7-ICT, 2011), MALORCA (H2020, 2016-2018), SUMMA (H2020, 2016-2019). Enfin, le domaine de la RAP est particulièrement marqué par de nombreuses campagnes d'évaluation : par exemple, ESTER 1 et 2, ETAPE, les éditions de MediaEval, ainsi que de multiples autres initiatives, notamment lancée par la DARPA et la IARPA aux États-Unis. En synthèse, seule la campagne internationale Blizzard Challenge existe.

# 4 Grands enjeux

## 4.1 Génération multimodale, située et interactive

Un discours n'est généralement pas désincarné car il apparaît sur un support, implique des acteurs (au moins rédacteur et lecteur(s), virtuels ou non), dans une situation donnée et pour un but communicatif fixé. Cet enjeu de génération multimodale, située et interactive va de paire avec l'analyse du contexte d'interaction et reste encore largement non résolu, à moins de développer des solutions ad hoc. Par exemple, le contrôle de l'expressivité, que ce soit pour le rendu de personnages, la distinction entre narration et style direct, ou encore la restitution des émotions, restent problématiques. En particulier, l'affect peut s'exprimer de différentes manières selon la modalité utilisée. Néanmoins, lors d'une interaction naturelle entre deux humains, l'expression de l'affect est essentielle pour transmettre de l'information permettant d'interpréter le message transmis. De nombreuses applications nécessitent sa prise en compte explicite comme la génération d'un message en langue des signes, ou encore la création et la lecture d'histoires pour enfants. Des travaux s'intéressent à l'analyse de l'affect dans les différentes modalités et d'autres à sa prise en compte lors de la génération. Néanmoins, des progrès sont encore attendus dans ce domaine. Cet enjeu d'adaptation au contexte, à plus long terme, est fondamental pour atteindre un niveau de qualité, d'expressivité et de naturel à la hauteur des interactions naturelles. Il est d'autant plus prégnant avec l'émergence des assistants intelligents des GAFA qui cherchent à s'immiscer dans l'intimité de l'habitat. Cette émergence de l'Internet des objets et des espaces intelligents est une très belle opportunité d'étudier la génération en contexte et de faire émerger des relations entre les situations et les choix de la GAT. Bien qu'une génération interactive multimodale ne soit pas obligatoirement par interaction langagière, des ponts évidents avec la communauté dialogue apparaissent ainsi que des problématiques éthiques importantes à prendre en considération.

## 4.2 Accessibilité

La génération de texte, de parole et de geste permet d'apporter une réponse à la problématique de l'accessibilité à des personnes en situation de handicap. Dans ce cadre, le développement de systèmes personnalisables, adaptables, permettant de prendre en compte la personnalité de l'utilisateur est nécessaires. La reconstruction de la voix d'une personne doit à la fois être de bonne qualité mais être suffisamment flexible. De même, pour la langue des signes, la génération d'un message ne se limite pas à une succession de mouvements stéréotypés mais doit inclure des mouvements contrôlés et précis pour être acceptée. L'accessibilité peut également recouvrir des groupes de la population, comme les enfants via la simplification de texte ou encore les personnes âgées à travers des réalisations peu ambiguës pour les sens (articulation, volume sonore, clarté gestuelle…). Le développement de ces technologies représente donc un enjeu pour l'accessibilité des espaces publics, des outils numériques, des contenus et de l'information.

## 4.3 Apprentissage automatique, apprentissage profond

L'arrivée de l'apprentissage profond et des réseaux de neurones (RN) profonds permet d'espérer un changement de paradigme bénéfique à multiples égards dans la manière de concevoir les systèmes dans la plupart des domaines du TALN. En effet, ces approches peuvent apporter des réponses au problème de transfert de modèles d'une application à une autre pour aider le prototypage rapide de nouveaux systèmes. Ils peuvent également apporter une réponse aussi bien dans la tâche de sélection de l'information que dans le cas de données hétérogènes. Cependant, leur véritable apport concernant l'acquisition de connaissances sur la tâche de génération reste encore largement inconnu. Par ailleurs, les RN et notamment les modèles séquence-à-séquence (seq2seq) ne garantissent pas une sortie correcte (phénomène d'hallucination, de répétition, etc.) ce qui n'est pas sans poser problème pour certains industriels (e.g., génération de notice de médicament, banque, commerce, etc.). Comme pour de nombreux domaines, le passage aux modèles neuronaux a laissé son lot de laissés-pour-compte. Les ressources nécessaires pour entraîner ces modèles (larges quantités de données, cartes de calcul GPU, temps d'entraînement importants) constituent évidemment un verrou à lever. En conséquence, l'objectif dans cet enjeu sera d'orienter l'emballement pour les méthodes profondes tout en maintenant l'intérêt sur les thématiques non abordées par ces méthodes ainsi que sur les approches alternatives. Enfin, les méthodes profondes sont une excellente opportunité d'étudier différentes tâches, de favoriser les ponts entre les domaines du TALN (traduction automatique, résumé automatique, génération automatique de légendes d'image), et d'attirer ainsi plus de chercheurs sur ce domaine. Elles sont également un fer de lance de l'intelligence artificielle pour le grand public et constituent donc un levier de recrutement et de financement.

## 4.4 Corpus, outils, langues minoritaires et sous-dotées

Les corpus et les outils de TAL constituent un enjeu majeur transversal à toutes les applications de génération de texte, de geste et de parole. Pour bon nombre d'applications, le manque de données est un frein à leur développement. Non seulement le volume mais également la qualité et l'hétérogénéité sont en cause ainsi que le déficit d'outils adaptés pour leur captation ou traitement. C'est encore plus vrai dans le cas de langues minoritaires ou sous-dotées, par exemple pour la langue des signes française ou le breton dans le cas des langues régionales. Ce manque de données limite notamment l'usage de méthodes statistiques même si dans certains cas, il est possible de s'appuyer sur des langues proches. On peut noter que le développement de l'outillage numérique est encouragé par la délégation générale à la langue française et aux langues de France (DGLFLF). Concernant le français, très peu de travaux portent sur cette langue et il n'existe aucun ensemble de données standardisé. De manière générale, il convient d'étudier les moyens de pallier le manque régulier de données pour chaque nouvelle application.

## 4.5 Évaluation

L'évaluation est une problématique difficile transverse à toutes les applications de génération de contenus. En effet, les contenus produits sont destinés à des utilisateurs finaux et leur évaluation nécessite la plupart du temps l'introduction d'un testeur humain. La difficulté est donc de trouver des métriques objectives corrélées à la perception permettant de réduire l'intervention humaine. Depuis une quinzaine d'années, la communauté GAT a mis une grande énergie dans la définition de méthodes d'évaluation [25] et dans des campagnes d'évaluations spécifiques pour les sous-tâches de GAT telles que la détermination du contenu, le micro-planning et la réalisation [10, 8]. Les métriques d'évaluation par corpus utilisées restent non satisfaisantes et inadaptées pour valider certains choix linguistiques de haut niveau et ne prédisent pas le succès des systèmes de GAT dans l'application visée (rapports descriptifs, aide à la décision, texte distrayant, texte affectif, etc.). Si les modèles neuronaux ont permis un bond en avant dans les domaines de la traduction, de la paraphrase ou de la simplification de texte, où la tâche consiste à transformer une phrase vers une autre relativement proche, ils échouent encore sur d'autres domaines, comme les chatbots ou le résumé automatique, en raison de la dimension et de la variabilité des données d'entrée et de sortie [32]. Pour la génération de parole et de geste, l'évaluation perceptive est également incontournable pour évaluer la qualité de sortie. En parole, de nombreuses méthodes existent, e.g. tests AB, ABX, MOS, MUSHRA. Néanmoins, le développement de méthodologies se focalisant sur l'évaluation de la parole dans le contexte applicatif visé est nécessaire afin de prendre en compte la capacité d'acceptation de l'utilisateur en fonction de la tâche. Pour les métriques objectives, le constat est le même que pour le texte, à l'heure actuelle, elles sont insuffisantes et ne permettent pas de conclure sur la qualité d'un système. Les efforts initiés dans la communauté doivent être poursuivis avec les autres communautés du TALN, notamment pour soutenir le développement des méthodes d'apprentissage automatique qui dépendent de fonctions de coût en lien avec l'objectif de génération.

# 5 Positionnement

## 5.1 Éthique

La génération de textes, de voix et de gestes nécessite l'utilisation de données naturelles. Dans les trois cas, ces données portent des informations personnelles sur l'identité ou les orientations de la personne qui les a produites. La conservation et l'utilisation de ces données sont régies par la législation (droit d'auteur, droit à l'image, propriété intellectuelle, RGPD). La génération en elle-même pose d'autres difficultés. À titre d'exemple, on ne peut synthétiser un message vocal sans l'accord de la personne prêtant sa voix. Par ailleurs, le contrôle de la génération et l'effet que la production peut avoir sur l'audience sont des problèmes (e.g. cas du robot microsoft Tay qui dérape sur twitter). Dans ce cas, la question de la responsabilité, notamment légale, des contenus produits se pose. Par la reproduction fidèle du comportement humain, l'utilisation de ces technologies amène aussi à se questionner sur leur acceptabilité et l'éthique de leur emploi. Sur le plan des relations humains-machine, un contenu artificiel ou un outil de production peuvent en effet être rejetés pour de multiples raisons : par exemple, la crainte d'être espionné, la déconsidération liés au fait de ne pas avoir une interaction avec un humain, l'agacement face à un manque d'ergonomie ressenti... Dans certains domaines, une réponse peut être d'inclure la population cible en amont dans la recherche ou du cycle de conception pour faire émerger des technologies adaptées, c'est le cas par exemple de la synthèse de la LSF pour la population sourde. Il peut être également utile d'avertir l'utilisateur lorsqu'un contenu ou une interaction est artificielle. Sur un second volet se pose la question de l'incidence sociétale à travers de difficiles problèmes tels que la suppression/mutation de l'emploi, les situation tolérables pour l'utilisation d'une machine à la place de l'humain (par exemple, en cas de problème grave dans des hôpitaux). Ces questions éthiques, partagés par tous les domaines de l'intelligence artificielle, appellent probablement avant tout une réponse politique mais ne prive pas la communauté d'une réflexion amont.

# 6 Programmatique

La section programmatique est commune avec l'axe 4, Intermodalité et multimodalité.

# Références


[1] Benchiheub, M. ; Berret, B. ; Braffort, A. (2016), Collecting and Analysing a Motion-Capture Corpus of French Sign Language, Workshop on the Representation and Processing of Sign Languages.

[2] Bengio et al. (2006). Neural probabilistic language models. Studies in Fuzziness and Soft Computing Volume 194, Springer, chapter 6.

[3] Black, A. W., Zen, H., and Tokuda, K. (2007, April). Statistical parametric speech synthesis. In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on (Vol. 4, pp. IV-1229). IEEE.

[4] Danlos, L. et Roussarie, L. (2000). Ingénierie des Langues, chapitre La génération automatique de textes, page 354 p. Traité IC2 (Information, communication et commande). Hermès, Paris.

[5] M. Filhol, M. N. Hadjadj, A. Choisier (2014), Non-manual features : the right to indifference, Representation and Processing of Sign Languages : Beyond the manual channel, Language resource and evaluation conference (LREC), Iceland.

[6] M. Filhol, J. McDonald, R. Wolfe (2017), Synthesizing Sign Language by connecting linguistically structured descriptions to a multi-track animation system, in Universal Access in Human-Computer Interaction (UAHCI), Springer LNCS 10278.

[7] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. Piater, and H. Ney (2012). RWTH-PHOENIX-Weather : A Large Vocabulary Sign Language Recognition and Translation Corpus. In Language Resources and Evaluation (LREC), pages 3785-3789, Istanbul, Turkey.

[8] Claire Gardent, Anastasia Shimorina, Shashi Narayan and Laura Perez-Beltrachini. Creating Training Corpora for NLG Micro-Planners. ACL 2017

[9] Albert Gatt, Emiel Krahmer. Survey of the State of the Art in Natural Language Generation : Core tasks, applications and evaluation. Journal of Artificial Intelligence Research, 61, 65-170.

[10] Gatt A., Belz A. (2010) Introducing Shared Tasks to NLG : The TUNA Shared Task Evaluation Challenges. In : Krahmer E., Theune M. (eds) Empirical Methods in Natural Language Generation. EACL 2009, ENLG 2009. Lecture Notes in Computer Science, vol 5790. Springer, Berlin, Heidelberg

[11] Sylvie Gibet, Nicolas Courty, Kyle Duarte, Thibaut Le Naour (2011), The SignCom system for data-driven animation of interactive virtual signers : Methodology and Evaluation. TiiS 1(1) : 6 :1-6 :23

[12] K. M. Hermann, T. Kočiský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom (2016). Teaching machines to read and comprehend. In Advances in Neural Information Processing Systems, vol. 28, p. 14

[13] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition : The shared views of four research groups. IEEE Signal Processing Magazine, 29(6), 82-97.

[14] Hunt, A. J., and Black, A. W. (1996, May). Unit selection in a concatenative speech synthesis system using a large speech database. In Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on (Vol. 1, pp. 373-376). IEEE.

[15] Jelinek, F. (1998). Statistical Methods for Speech Recognition, MIT Press.

[16] Kennaway, R. (2001), Synthetic Animation of Deaf Signing Gestures, International Gesture Workshop, London.

[17] Fernando López-Colino, José Colás (2011), The Synthesis of LSE Classifiers : From Representation to Evaluation, In Journal of Universal Computer Science, vol. 17, no. 3 (2011), 399-425

[18] Lu et al. (2015). A study of the recurrent neural network encoder-decoder for large vocabulary speech recognition. In Interspeech ; pp. 3249-3253.

[19] McDonald, J., Wolfe, R., Schnepp, J., Hochgesang, J., Jamrozik, D., Stumbo, M., Berke, L., Bialek, M., Thomas, F. (2015), An automated technique for real-time production of lifelike animations of American Sign Language, pp. 1–16

[20] Merritt, T., Clark, R. A., Wu, Z., Yamagishi, J., and King, S. (2016, March). Deep neural network-guided unit selection synthesis. In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on (pp. 5145-5149). IEEE.

[21] Mikolov, T., Karafiát, M., Burget, L., Černocký, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In Interspeech.

[22] F. Nunnari, M. Filhol, A. Heloir (2018), Animating AZee descriptions using off-the-shelf IK solvers, workshop on the representation and processing of Sign Languages, Miyazaki, Japan.

[23] Palaz, D. (2016). Towards End-to-End Speech Recognition. EPFL.

[24] Prillwitz S. ; Leven, R. ; Zienert, H. ; Hanke, T. and Henning, J (1989). HamNoSys version 2.0, Hamburg notation system for sign languages, an introductory guide International studies on Sign Language communication of the Deaf, Signum press, Hamburg.

[25] Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. Comput. Linguist. 35, 4 (December 2009), 529-558.

[26] Ehud Reiter and Robert Dale. 2000. Building Natural Language Generation Systems. Cambridge University Press, New York, NY, USA.

[27] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112).

[28] Wan, V., Agiomyrgiannakis, Y., Silen, H., and Vit, J. (2017). Google's next-generation real-time unit-selection synthesizer using sequence-to-sequence lstm-based autoencoders. In Interspeech (pp. 1143-1147).

[29] Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... and Le, Q. (2017). Tacotron : Towards End-to-End Speech Synthesis. In Interspeech.

[30] Wu, Z., Valentini-Botinhao, C., Watts, O., and King, S. (2015, April). Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on (pp. 4460-4464). IEEE.

[31] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... and Bengio, Y. (2015, June). Show, attend and tell : Neural image caption generation with visual attention. In International Conference on Machine Learning (pp. 2048-2057).

[32] J. Yao, X. Wan and J. Xiao (2017). Recent advances in document summarization. In Knowledge Information System, vol. 53 n. 2, pp. 297-336.

[33] M. Zock, G. Sabah. La génération automatique de textes : trente ans déjà, ou presque. Langages, 26(106), 1992, pp. 8-35.