

Make text look like speech: disfluency generation using sequence-to-sequence neural networks

Henri Lasselin, Gwénolé Lecorvé

▶ **To cite this version:**

Henri Lasselin, Gwénolé Lecorvé. Make text look like speech: disfluency generation using sequence-to-sequence neural networks. [Rapport de recherche] Univ Rennes, CNRS, IRISA, France; IRISA, équipe EXPRESSION. 2018. hal-01738344

HAL Id: hal-01738344
https://inria.hal.science/hal-01738344

Submitted on 20 Mar 2018

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

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



Master research Internship

Bibliographic report

Make text look like speech: disfluency generation using sequence-to-sequence neural networks

Domain: Document and Text Processing - Machine Learning

Author:
Henri Lasselin

Supervisor:
Gwénolé Lecorvé
Expression - Irisa

ENSSAT
LANNION

Abstract: La synthèse de discours spontanés naturels est un défi à relever. Une manière de s'en approcher est de produire des discours disfluents. En effet, les disfluences sont très présentes en parole spontanée. Récemment, des travaux ont proposé une méthode permettant de générer des disfluences à l'aide de modèles de langage et de champs aléatoires conditionnels. Toutefois, les réseaux de neurones permettent de traiter de nombreux problèmes en traitement automatique des langues et il peut être judicieux de les utiliser pour produire des disfluences. Dans ce document, nous dressons l'état de l'art des disfluences ainsi que des modèles séquence-à-séquence afin de réaliser ce travail lors d'un stage de master et de suivre les pistes les plus appropriées.

Table des matières

1 Introduction **1**

2 Étude des disfluences **1**
2.1 Structure des disfluences . . . 1
2.2 Niveau prosodique . . . 3
2.2.1 Pauses silencieuses . . . 3
2.2.2 Allongements . . . 3
2.2.3 Troncatures . . . 3
2.3 Niveau syntaxique . . . 3
2.3.1 Pauses remplies . . . 4
2.3.2 Répétitions . . . 4
2.4 Niveau sémantique . . . 4
2.4.1 Révisions . . . 5
2.4.2 Faux départs . . . 5
2.5 Compositions de disfluences . . . 5
2.6 Production de disfluences . . . 6

3 Modèles séquence-à-séquence **7**
3.1 Champs aléatoires conditionnels . . . 8
3.2 Réseaux de neurones . . . 8
3.2.1 Réseaux de neurones récurrents . . . 9
3.2.2 Long Short-Term Memory . . . 10
3.2.3 Gated Recurrents Units . . . 10
3.2.4 Generative Adversarial Networks . . . 11
3.2.5 Mécanisme d'attention . . . 11
3.3 Applications des réseaux de neurones séquence-à-séquence . . . 12

4 Conclusion **13**

1 Introduction

Récemment, de nombreuses avancées ont été faites en synthèse de la parole, notamment grâce aux progrès en apprentissage automatique. Celles-ci permettent de produire des discours clairs et fluides comme lors de discours préparés ou de lectures de texte. Cependant, la production de parole spontanée reste un défi à relever car cette parole est naturellement expressive. En particulier, elle contient souvent des disfluences, sujet d'étude de ce rapport.

Les disfluences sont des phénomènes qui interrompent le flux de la parole, sans ajouter de contenu propositionnel [Tree, 1995]. Longtemps, elles ont été considérées comme étant des éléments perturbant le discours, c'est pourquoi des recherches ont été faites sur la suppression de celles-ci [Hassan et al., 2014]. Cependant, plusieurs études ont montré que les disfluences présentent plusieurs intérêts pour la communication. Elles permettent de prévenir les auditeurs de la complexité du discours à suivre, les aident à en comprendre la structure et rendent possible la correction d'erreurs faites précédemment [Tree, 2001, Rose, 1998] (cités par [Adell et al., 2012]). Les disfluences sont donc très importantes pour produire une parole spontanée et naturelle.

Récemment, des travaux ont été réalisés pour générer des textes disfluents [Qader et al., 2014, Qader, 2017]. Ces travaux proposent une implémentation fondée sur des modèles de langage et des champs aléatoires conditionnels. Notre objectif est de poursuivre cette tâche grâce à des réseaux de neurones et sous l'angle d'un modèle séquence-à-séquence. En effet, la génération de textes disfluents peut être vue comme la transformation d'un texte fluide (une séquence de mots) en un texte disfluent (une autre séquence de mots).

Ce document présente une étude bibliographique préalable à la réalisation de ce travail lors d'un stage de master. Afin de suivre les pistes les plus adéquates, nous présenterons tout d'abord les disfluences, la manière de les représenter ainsi que leurs rôles. Puis, nous nous intéresserons aux tâches séquence-à-séquence, aux modèles permettant de les traiter ainsi qu'à leurs applications. Enfin, notons que la synthèse audio des discours disfluents, bien qu'étant la motivation de fond, reste en dehors du cadre de la présente étude et du stage. Nous nous concentrerons uniquement sur la production de textes disfluents.

2 Étude des disfluences

Dans cette partie, nous allons étudier les disfluences. Nous commencerons par donner la façon dont elles sont structurées. Nous présenterons ensuite différentes catégories de disfluences avant de nous intéresser à leur composition. Enfin, nous verrons comment les travaux traitent actuellement ce phénomène.

2.1 Structure des disfluences

Une disfluence est une discontinuité dans un énoncé. Cette discontinuité peut introduire des modifications de l'énoncé autour de son point central. Des études ont révélé des régularités dans la structure des disfluences [Shriberg, 1994]. La structure adoptée dans ce dernier article est présentée dans la figure 1. Elle permet de représenter tous les types de disfluences comme une suite de sections, chacune ayant un rôle dans la disfluence. Ces sections s'articulent autour d'un point central, dit point d'interruption (PI), qui marque une coupure dans le discours : c'est ici que le locuteur se rend compte d'une erreur. Avant ce PI, le *reparandum* correspond à la partie erronée du discours. Selon les personnes, le *reparandum* est uniquement le mot contenant l'erreur ou bien la zone complète qui

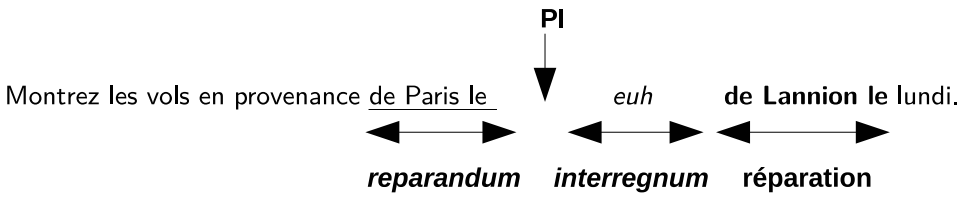


FIGURE 1 – Structure des disfluences proposée par [Shriberg, 1994]

contient ce mot (c'est cette dernière version qui est montrée dans la figure). Ensuite, juste après le PI, l'*interregnum* indique l'interruption de la parole par un intervalle entre l'erreur et la réparation de celle-ci. Enfin, la réparation contient la correction du discours. C'est cette partie qui rectifie l'erreur faite dans le *reparandum*. Par la suite, nous utiliserons le même format que sur la figure 1 : le *reparandum* sera souligné, l'*interregnum* sera en italique et enfin, la réparation sera en gras. De plus, le PI sera indiqué entre parenthèses.

Il existe plusieurs types de disfluences, tous conformes à ce modèle mais avec des particularités, et chacun ayant des origines, des rôles et des impacts différents. On peut relever les disfluences suivantes : les allongements, les répétitions, les révisions, les faux départs ainsi que les pauses. Parmi ces dernières, on peut distinguer les pauses silencieuses et les pauses remplies. On peut noter que la terminologie peut varier en fonction des articles.

Les PI étant les éléments centraux des disfluences, il est capital de prédire leur fréquence et leurs emplacements au sein du discours afin de produire ces dernières. On peut tout d'abord relever que la fréquence des disfluences dépend de beaucoup d'éléments [Shriberg, 1994]. Par exemple, elle varie en fonction du genre : les femmes feraient globalement moins de disfluences que les hommes. Les disfluences surviennent plus fréquemment lorsque le discours n'est pas préparé car le locuteur doit anticiper ce qu'il va dire tout en parlant. Il existe également des positions où insérer les PI. Par exemple, il y a davantage de disfluences entre deux mots qui n'ont pas l'habitude d'être juxtaposés. Les disfluences se produisant plus souvent lorsque le locuteur planifie la suite de sa phrase, près de la moitié de celles-ci survient avant des mots-outils [Maclay and Osgood, 1959, Blankenship and Kay, 1964] (cités par [Shriberg, 1994]). La position des disfluences est importante car [Dall et al., 2014] montre qu'il existe des endroits propices aux disfluences et qu'à l'inverse, la parole est perçue comme peu naturelle lorsqu'elles sont insérées à de mauvais endroits.

Cependant, ces articles s'intéressent aux caractéristiques des disfluences et non à leurs origines. Pourtant, les disfluences dépendent de l'intention du locuteur et de multiples autres facteurs liés à son état mental (énervé, stressé...) ou encore son expérience personnelle. Par exemple, le locuteur est en train de parler et le mot « voiture » vient dans la conversation. Il va alors peut-être se rappeler de sa propre voiture et peut produire une disfluence en conséquence.

Il existe plusieurs manières de catégoriser les disfluences : selon leur complexité ou selon leur rôle par exemple. Nous avons choisi d'étudier chaque type de disfluences en les regroupant selon leur influence sur différents niveaux d'abstraction du langage, à savoir la prosodie, la syntaxe et la sémantique.

2.2 Niveau prosodique

Certaines disfluences influent nettement sur la prosodie, c'est-à-dire qu'elles modifient le rythme et l'intonation du discours. On retrouve dans cette catégorie les pauses silencieuses, les allongements et les troncatures.

2.2.1 Pauses silencieuses

Les pauses sont des interruptions dans le discours. Elles ne possèdent ni *reparandum*, ni réparation. Dans le cas des pauses silencieuses, aucun son n'est prononcé dans l'*interregnum* : il n'y a qu'un silence (*cf.* l'exemple suivant).

Montrez-moi les vols en provenance de (PI) **silence** Lannion le lundi.

Il est tout d'abord nécessaire de mentionner que toutes les pauses silencieuses ne sont pas des disfluences et il peut être compliqué de distinguer les deux [Tree, 1995]. Par exemple, les pauses dues à la ponctuation sont inhérentes au langage.

Les pauses silencieuses sont les disfluences les plus fréquentes [Betz et al., 2015]. Elles surviennent aussi bien en anticipation d'une difficulté que lorsque celle-ci est déjà survenue. Cette étude montre que la durée de ces pauses est très variable : elle peut atteindre 800ms sans être perçue comme non naturelle. [Shriberg, 1994] note que les pauses silencieuses peuvent aider les interlocuteurs à discerner la partie réparation dans le discours.

2.2.2 Allongements

Les allongements sont des disfluences particulières. En effet, un allongement seul contient un *reparandum* sans réparation associée, il ne possède pas d'*interregnum* non plus : il consiste en le prolongement de la syllabe précédant le PI. Tout comme les pauses silencieuses, les allongements augmentent la durée du discours et permettent au locuteur de réfléchir à ce qu'il va dire après. Ils sont complexes à étudier car il est difficile de déterminer de manière absolue à partir de quelle durée une syllabe est considérée comme allongée, cette durée étant propre à chaque corpus. Cependant, [Betz et al., 2015] a montré que dans un contexte donné (même personne, même corpus), une syllabe allongée dure généralement deux fois plus longtemps qu'une syllabe normale.

2.2.3 Troncatures

Les troncatures sont des disfluences pour lesquelles le PI se situe au milieu d'un mot. Cela survient lorsque le locuteur perçoit son erreur alors qu'il prononce un mot.

Selon les corpus, le taux de troncatures dans la parole spontanée varie énormément (de 22 à 60%) [Shriberg, 1994]. Pour le corpus utilisé dans [Betz et al., 2015], ce taux atteint même 15%. Il n'y a pas d'explication claire à cette variation. Ce dernier article montre également que les troncatures sont peu appréciées dans les communications humaines mais qu'elles restent nécessaires car elles facilitent la correction d'erreur.

2.3 Niveau syntaxique

Les disfluences peuvent également modifier la syntaxe en introduisant de nouveaux syntagmes. Ces disfluences changent la structure de la phrase, sans en modifier le sens. Dans cette catégorie, on retrouve les pauses remplies et les répétitions.

2.3.1 Pauses remplies

Les pauses remplies sont similaires aux pauses silencieuses dans le sens où elles ne sont composées que d'un *interregnum*. Cependant, celui-ci ne contient pas un silence mais contient des éléments qui n'ajoutent aucune information. En anglais, ces pauses sont *« uh »* et *« um »* (resp. pauses courtes et longues), les éléments correspondants en français sont « euh » et « hmmm ».

Certaines études considèrerent que les marqueurs de discours tels que « enfin », « voilà » et « en fait » (en anglais *« I mean »*, *« well »* et *« you know »*) font également partie de cette catégorie. La phrase suivante est un exemple de pause remplie :

Montrez-moi les vols en provenance de (PI) *euh* Lannion le lundi.

[Shriberg, 1994] relève que contrairement aux autres disfluences, le taux de pauses remplies est stable, peu importe la complexité de la tâche. Ceci peut expliquer pourquoi la plupart des systèmes de génération de discours disfluents se concentraient sur cette famille de disfluences. Pour [Maclay and Osgood, 1959], celles-ci reflètent un besoin de pauses combiné à un besoin de continuer de parler, c'est-à-dire d'éviter les silences. De même, en s'intéressant à la gestuelle du locuteur, on remarque que les gestes et les pauses remplies ne coexistent pas. Ainsi, si on interdit au locuteur de faire des gestes en s'exprimant, il aura davantage tendance à faire des pauses remplies : les gestes et les pauses remplies semblent occuper le même rôle dans la parole, bien que celui-ci ne soit pas clairement établi [Christenfeld et al., 1991].

Une autre étude a également montré que la production de ces disfluences détériorait le naturel de la synthèse de la parole [Betz et al., 2015]. Cependant, ce manque de naturel était davantage dû à la mauvaise qualité de synthèse de ces disfluences qu'à leur présence.

2.3.2 Répétitions

Les répétitions consistent en la répétition, sans modification, d'une partie de la phrase. Elles n'ont pas d'*interregnum* et la partie réparation est l'exacte recopie du *reparandum*, comme dans l'exemple suivant :

Montrez-moi les vols en provenance de Lannion le (PI) **le** lundi.

[Tree, 1995] montre que, dans un discours spontané, les répétitions permettent d'augmenter l'attention de l'auditeur sur le mot suivant celle-ci. En effet, le temps de réaction pour détecter le mot suivant est plus court avec répétition, que sans. Cependant, il relève qu'une répétition générée artificiellement n'avait pas d'influence sur le temps de réaction, sûrement à cause de la faible qualité audio des répétitions ajoutées. Les répétitions n'ont donc, au pire, aucune influence sur la parole spontanée et, au mieux, aident à la compréhension de celle-ci : ce sont donc des disfluences intéressantes à produire. Elles peuvent permettre au locuteur de gagner du temps pour réfléchir à ses prochains mots. Elles peuvent aussi aider les interlocuteurs à se recentrer sur le discours après une longue pause [Shriberg, 1994].

2.4 Niveau sémantique

Les disfluences peuvent avoir un impact sur la sémantique en introduisant des variations de sens du discours. Cette catégorie comprend les révisions et les faux-départs.

2.4.1 Révisions

Les révisions ont la même structure que les répétitions, mais la partie réparation est différente du *reparandum*. L'exemple suivant montre le rôle principal des révisions : celui de corriger une erreur faite précédemment.

Montrez-moi les vols en provenance de Paris (PI) **de Lannion** le lundi.

Cet exemple montre également un phénomène relevé par [Shriberg, 1994] : la répétition d'une partie de la phrase précédant le mot corrigé (« de » dans l'exemple). Ce phénomène survient moins souvent lorsque l'erreur est phonétique comme dans la phrase suivante :

Montrez-moi les vols en provence (PI) **provenance** de Lannion le lundi.

Cela peut s'expliquer par le fait que l'interlocuteur comprend immédiatement la correction d'une erreur phonétique (il l'assimile à une répétition).

Dans certains cas, la partie réparation ne contredit pas le *reparandum* [Shriberg, 1994]. En effet, les révisions peuvent également servir à apporter des précisions, comme par exemple dans la phrase suivante :

Montrez-moi les vols (PI) **les dix premiers vols** en provenance de Lannion le lundi.

Ce dernier cas est sujet à discussion car toutes les précisions ne sont pas des révisions. Cela dépend de l'intention du locuteur : s'il avait planifié une précision, ce n'est pas une disfluence. À l'inverse, si de son point de vue il s'est trompé, on considère la précision comme une disfluence.

2.4.2 Faux départs

Les faux départs sont des disfluences pour lesquelles le locuteur abandonne la phrase qu'il énonçait et en commence une nouvelle. Ils peuvent donc être vus comme un cas particulier de révisions. L'exemple suivant montre un faux départ :

Je souhaite (PI) **Montrez-moi les vols en provenance de Lannion le lundi.**

[Tree, 1995] a étudié les faux départs. Cette étude montre que ces derniers perturbent les interlocuteurs. En effet, les interlocuteurs se forment une image mentale du discours qu'ils écoutent. Après un faux départ, ils doivent alors détruire cette image et en construire une nouvelle. Ces travaux montrent également que cette gêne est la même quelque soit la position du faux départ dans la phrase.

2.5 Compositions de disfluences

Au delà des archétypes présentés pour chaque type de disfluences, il est fréquent que plusieurs disfluences coexistent dans un énoncé. [Betz et al., 2015] propose deux scénarios. Le premier survient lorsque le locuteur se rend compte tardivement d'un changement dans le plan qu'il s'était fixé. Il interrompt immédiatement la parole (ce qui peut produire une troncature) puis, va chercher un nouveau plan de discours. Tant qu'il ne le trouve pas, il va produire d'autres disfluences en commençant par des pauses silencieuses, puis des pauses remplies. Le deuxième scénario est similaire

au premier à l'exception près que l'information d'un changement de plan arrive tôt. Le locuteur va alors commencer par allonger la syllabe qu'il prononce avant de produire les autres disfluences.

[Betz et al., 2015] a également remarqué qu'une troncature seule (qui génère un fragment de mot) semblait abrupte et peu naturelle. Il suggère que cela vient du fait que, dans le langage naturel, le locuteur change d'intonation et/ou allonge la syllabe avant de couper le mot. De plus, le locuteur coupe rarement un mot sans effectuer une révision ensuite. Par exemple, on peut supposer qu'il a mal prononcé un mot, s'en rend compte et se corrige immédiatement :

Montrez-moi les vols en provenance de Lon (PI) *enfin* **Lannion** le lundi.

Un marqueur de discours est également présent dans cet exemple. En effet, les pauses (et donc les marqueurs de discours) servent à ralentir le discours. Elles sont donc très utiles en combinaison des disfluences influant sur la sémantique. En outre, cela permet aux interlocuteurs de comprendre qu'une correction du discours va avoir lieu. Ce rôle peut aussi être rempli par les répétitions [Tree, 1995].

2.6 Production de disfluences

Les disfluences ont longtemps été perçues comme des éléments perturbant le discours. Cependant, dans l'objectif de générer une parole spontanée naturelle, plusieurs travaux ont récemment été réalisés afin de générer des discours difluents. Toutefois, ces travaux ne se concentrent que sur certains types de disfluences. Rappelons que notre perspective de travail se concentre sur la production de disfluences, et non leur synthèse.

Parmi les travaux de l'état de l'art, [Adell et al., 2007] n'étudie que la production de pauses remplies « uh » et « um ». La méthode consite à trouver tout d'abord le point d'interruption (PI), puis à produire la pause. La recherche de PI se fait grâce à un modèle de langage et à un arbre de décision. Ce dernier classifie chaque mot du texte en deux catégories : ceux qu'il faut faire suivre d'une disfluence et les autres. Afin de faire cette classification, l'arbre de décision s'appuie sur des probabilités du modèle de langage et des étiquettes morphosyntaxiques (nom, verbe, déterminant...).

Récemment, des travaux ont été faits sur la production de plusieurs types de disfluences : les répétitions et les pauses [Qader, 2017]. Le modèle théorique proposé permet néanmoins de produire des révisions et donc des faux-départs.

Le principe de cette méthode est de voir une disfluence comme le résultat d'une fonction qui transforme une phrase fluide. Ce processus de transformation est décomposé : il y a ainsi une fonction de transformation f_T par type de disfluences T. Ces fonctions prennent en paramètre une séquence de mots sous forme de vecteur. Elles donnent en résultat une autre séquence de mots ayant la disfluence associée. En se basant sur ce processus, plusieurs disfluences peuvent être générées en composant les fonctions comme le montre la figure 2. Chaque disfluence peut être générée 0, 1 ou plusieurs fois. L'ordre de composition est fixé : les révisions sont générées en premier car elles sont complexes. Les répétitions viennent ensuite. En effet, si elles étaient produites avant les révisions, ces dernières pourraient interrompre une répétition. Les pauses sont produites en dernier car elles peuvent être insérées au milieu des autres disfluences et elles sont plus faciles à produire.

[Qader, 2017] propose une implémentation de ce principe. Celle-ci peut produire des répétitions et des pauses. Pour chaque disfluence, la génération est décomposée en la prédiction de PI puis en l'insertion de la disfluence. La prédiction de PI utilise un CRF afin d'étiqueter les mots (l'étiquette indique si le mot est suivi d'une disfluence ou non). Afin de faire cette prédiction, le texte fluide

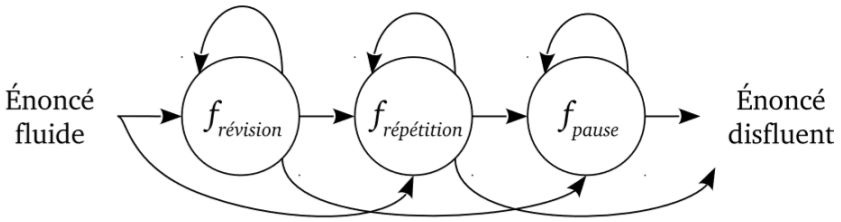


FIGURE 2 – Processus de production de disfluences (Source : [Qader et al., 2014])

.

est au préalable annoté avec des descripteurs (POS, fréquence du mot, nombre de syllabes, position du mot...). Pour l'insertion de disfluences, l'algorithme génère un ensemble de phrases candidates selon la disfluence. Ensuite, celles-ci sont évaluées à l'aide d'un modèle de langage et la phrase la plus probable est conservée. Un critère d'arrêt existe pour chaque disfluence. Il permet de choisir le degré de disfluence de la phrase.

Enfin, l'évaluation des discours disfluents produits n'est pas un sujet d'étude dans la littérature mais est un aspect primordial qu'il faut considérer avec attention. Une évaluation objective peut être réalisée en s'intéressant aux statistiques des systèmes comme le rappel ou la précision. Cependant, une telle méthode n'est pas suffisante pour évaluer la production de disfluences. En effet, plusieurs positions de disfluences sont acceptables pour chaque énoncé et, pour chacune de ces positions, plusieurs types de disfluences sont également possibles [Dall et al., 2014]. On peut alors se tourner vers une évaluation subjective : les énoncés disfluents sont présentés à des testeurs qui en évaluent la qualité. Ces énoncés peuvent être présentés sous forme écrite ou sonore. Dans le premier cas, les testeurs devront imaginer que le texte est prononcé. Dans le second cas, les énoncés peuvent être prononcés par des humains ou par un système de synthèse. D'un côté, faire prononcer les énoncés par des humains coûte cher et prend du temps. D'un autre côté, synthétiser les énoncés est plus abordable, mais les résultats de l'évaluation dépendent énormément de la qualité du système de synthèse qui est généralement faible pour des textes disfluents car les systèmes ne sont pas prévus pour ce type de texte. Une fois ce premier travail effectué, des questions sont posées aux testeurs, comme par exemple « Pensez-vous que les pauses remplies rendent le texte (plus/autant/moins) naturel ? ». Ces questions sont à choisir minutieusement car les résultats peuvent être influencés par celles-ci [Dall et al., 2014]. Malgré son coût et le temps requis par les testeurs, l'évaluation subjective reste cependant la méthode la plus pertinente pour évaluer la production de disfluences.

Nous avons donc étudié les différentes disfluences que l'on peut être amené à produire. La production de disfluences pouvant être vue comme le passage d'une séquence de mots fluide en une séquence de mots disfluente, nous allons maintenant nous intéresser aux modèles séquence-à-séquence.

3 Modèles séquence-à-séquence

Dans cette partie, nous allons étudier les modèles séquence-à-séquence. Ces modèles ont pour but de générer une séquence d'éléments à partir d'une autre séquence et peuvent donc nous permettre de produire un énoncé disfluent à partir d'un énoncé fluide. Nous allons commencer par présenter différents modèles d'apprentissage automatique, notamment des modèles neuronaux, pouvant réali-

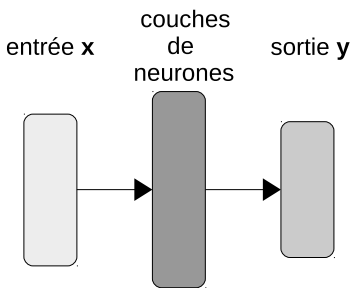


FIGURE 3 – Schéma d'un réseau de neurones.

ser des tâches séquence-à-séquence. Enfin, nous regarderons quelles applications ont ces réseaux de neurones séquence-à-séquence.

3.1 Champs aléatoires conditionnels

Les champs aléatoires conditionnels (CRF pour *Conditional Random Fields*) sont des modèles probabilistes. Ces modèles sont très utiles en traitement automatique des langues car ils permettent, entre autres choses, de faire de l'annotation séquentielle [Constant et al., 2011]. En effet, à partir d'une séquence d'entrées $\mathbf{x} = (x_1, \ldots, x_T)$, ces modèles calculent la probabilité d'une séquence d'étiquettes $\mathbf{y} = (y_1, \ldots, y_T)$ selon la formule suivante :

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_\theta(x)} \exp(\sum_{t=1}^{T} \sum_{k=1}^{K} \theta_k f_k(y_{t-1}, y_t, x_t)), \quad (1)$$

avec f_k des fonctions caractéristiques, θ_k les poids associés estimés sur les données d'apprentissage de façon à minimiser le taux d'erreur d'étiquetage et $Z_\theta(x)$ un facteur de normalisation. Les fonctions caractéristiques donnent généralement un résultat binaire (0 ou 1) indiquant la présence d'une caractéristique donnée parmi les paramètres fournis en entrée [Lafferty et al., 2001].

La séquence choisie par le modèle est celle ayant la probabilité $p(\mathbf{y}|\mathbf{x})$ la plus élevée. Les CRF sont particulièrement intéressants car ils permettent de prendre en compte les dépendances entre les étiquettes.

3.2 Réseaux de neurones

Les réseaux de neurones sont des modèles capables d'apprendre des relations très complexes à partir de données. Comme le montre la figure 3, ce sont des modèles qui prennent un vecteur d'entrée et le transforme successivement grâce à des couches de neurones jusqu'à produire une sortie, elle aussi sous forme vectorielle. Les neurones sont des cellules qui reçoivent des valeurs numériques en entrée, les transforment via une fonction f, et transmettent le résultat aux neurones de la couche suivante. La transformation réalisée par un neurone consiste tout d'abord à faire une somme du vecteur d'entrée pondérée par un vecteur de poids $\boldsymbol{\theta}$, puis à appliquer une fonction d'activation (par exemple la fonction sigmoïde).

Le but de la phase d'apprentissage est de trouver la pondération $\boldsymbol{\theta}$ qui minimise l'erreur entre la sortie du réseau et celle souhaitée. Pour cela, la technique la plus utilisée est celle de la rétropro-

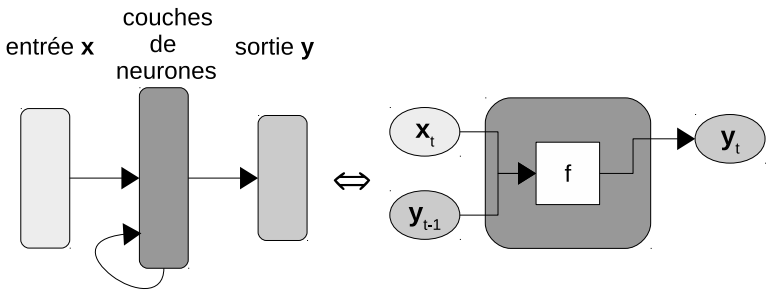


FIGURE 4 – Schéma d'un bloc RNN

pagation du gradient qui consiste à corriger les erreurs en fonction de l'importance des éléments qui y ont participé.

On peut représenter un réseau de neurones qui admet $\mathbf{x}$ en entrée par :

$$\mathbf{y} = f_{\boldsymbol{\theta}}(\mathbf{x}), \tag{2}$$

avec $\mathbf{y}$ la sortie et $\boldsymbol{\theta}$ les paramètres du réseau. Il existe de très nombreux réseaux de neurones qui diffèrent selon la manière dont les couches sont organisées, les connexions entre les neurones ou encore le fonctionnement interne des neurones.

3.2.1 Réseaux de neurones récurrents

Les réseaux de neurones récurrents (RNN pour *Recurrent Neural Network*) sont des réseaux de neurones particulièrement adaptés aux données séquentielles.

La structure des RNN est semblable à celle des réseaux de neurones classiques. Cependant, la sortie de la couche de neurones correspondant au $(t-1)$-ième élément de la séquence est réinjectée en tant qu'entrée de la couche pour l'élément t (*cf.* figure 4). Ainsi, la t-ième sortie $\mathbf{y}_t$ d'un bloc RNN se calcule comme :

$$\mathbf{y}_t = f_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{y}_{t-1}), \tag{3}$$

avec $\mathbf{y}_{t-1}$ la $(t-1)$-ième sortie, $\mathbf{x}_t$ la t-ième entrée et $\boldsymbol{\theta}$ les paramètres du réseau. Ces réseaux permettent donc de prendre en compte une dépendance entre les entrées d'une séquence.

Ils sont beaucoup utilisés dans les travaux de traitement du langage naturel pour plusieurs raisons [Young et al., 2017]. Premièrement, dans le langage naturel, le sens d'un mot dépend en partie des mots précédents. De plus, les phrases n'ont pas toutes le même nombre de mots ; or, les réseaux de neurones ont un vecteur d'entrées de taille fixe. Le jeu de chaînage de l'équation (3) permet d'apporter une solution à ces deux problèmes.

La technique utilisée pour l'apprentissage est une variante de celle utilisée pour les réseaux de neurones classiques : la rétropropagation du gradient à travers le temps. Cependant, les RNN connaissent le problème de disparition du gradient [Pascanu et al., 2013]. À cause de ce dernier, les systèmes ont du mal à apprendre de longues dépendances. D'autres réseaux de neurones, comme les *Long Short-Term Memory*, permettent de contourner ce phénomène.

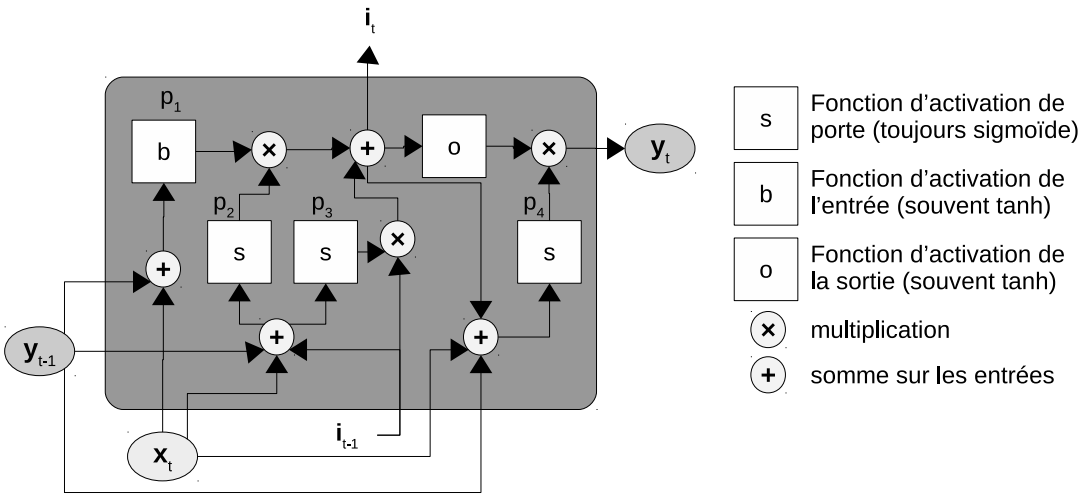


FIGURE 5 – Schéma d'un bloc LSTM. p_1 est appelé bloc d'entrée, p_2 porte d'entrée, p_3 porte d'oubli et p_4 porte de sortie. (D'après [Greff et al., 2017])

3.2.2 Long Short-Term Memory

Les *Long Short-Term Memory* (LSTM) sont un type particulier de RNN beaucoup plus complexes que ces derniers [Greff et al., 2017]. Comme le représente le système d'équations (4), la sortie $\mathbf{y}_t$ dépend toujours de la sortie $\mathbf{y}_{t-1}$ et de l'entrée $\mathbf{x}_t$. Toutefois, elle dépend également d'une nouvelle information récurrente $\mathbf{i}_{t-1}$, appelée mémoire, qui saisit les dépendances à long terme. Pour produire la valeur $\mathbf{i}_t$ à l'instant t, le bloc LSTM est composé de plusieurs « portes » qui ont pour objectif de conserver, supprimer ou modifier de l'information (voir la figure 5 pour plus de détails).

$$\begin{cases} \mathbf{y}_t = f_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{y}_{t-1}, \mathbf{i}_{t-1}) \\ \mathbf{i}_t = g_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{y}_{t-1}, \mathbf{i}_{t-1}) \end{cases}, \tag{4}$$

avec $\boldsymbol{\theta}$ les paramètres du réseau. Grâce à cette structure complexe, les LSTM sont efficaces pour capturer les dépendances à long terme, contrairement aux RNN classiques. De plus, ils possèdent leurs avantages (dépendances à court-terme, séquences de longueur variable). Cependant, l'apprentissage est plus coûteux à cause de ses nombreuses connexions et donc de ses paramètres à estimer.

3.2.3 Gated Recurrents Units

Les *Gated Recurrents Unit* (GRU) sont une autre variante de RNN, introduite récemment par [Cho et al., 2014]. Ces réseaux sont plus simples que les LSTM car ils possèdent une porte en moins.

Une étude a comparé les GRU, les LSTM et les RNN classiques [Chung et al., 2014]. Elle montre que, pour des tâches autres que le traitement du langage naturel (l'étude ne s'intéressant pas à ce domaine), les réseaux GRU et LSTM sont plus performants que les RNN classiques. Toutefois, cette étude ne parvient pas à déterminer lequel des deux premiers est le meilleur. Par conséquent, de par leur plus grande simplicité, les réseaux GRU seront privilégiés lorsque la puissance de calcul disponible est limitée [Greff et al., 2017].

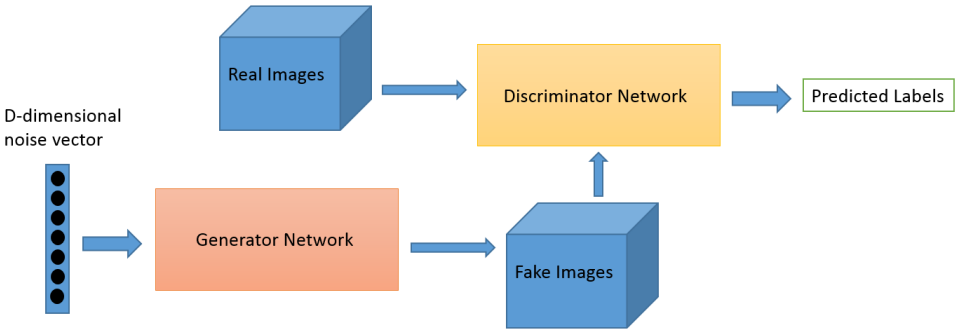


FIGURE 6 – Principe des GAN (Source : [Bruner and Deshpande, 2017])

.

3.2.4 Generative Adversarial Networks

Les Generative Adversarial Networks (GAN), proposés par [Goodfellow et al., 2014] font partie de la catégorie des modèles génératifs, c'est-à-dire ayant pour but de créer des données. Comme en témoigne la figure 6, le principe des GAN est de mettre en concurrence deux modèles. Le premier est un réseau générateur qui est chargé de produire des données à partir d'un vecteur de bruit. Le second est un réseau discriminatif qui apprend à différencier les données générées de données réelles fournies par l'utilisateur. Ce modèle peut être utilisé pour produire toutes sortes de données, aussi bien des images que des textes ou des signaux audio.

3.2.5 Mécanisme d'attention

Un réseau de neurones récurrents réalise des prédictions à partir de données séquentielles. Chaque élément apporte une part d'information que le réseau utilise pour produire une sortie. Toutefois, il est possible que certains éléments d'une séquence contiennent davantage d'informations utiles à la prédiction. Le mécanisme d'attention a pour objectif de repérer quels sont les éléments les plus utiles.

Par exemple, [Bahdanau et al., 2014] propose une architecture RNN de type encodeur-décodeur avec un mécanisme d'attention. Un encodeur-décodeur est un modèle en deux composantes qui permet de faire des correspondances entre une séquence d'entrée et une séquence de sortie. La partie encodeur est chargée de représenter la séquence d'entrée en un vecteur de taille fixée, tandis que la partie décodeur utilise cette représentation afin de générer une séquence en sortie. Cependant, tous les éléments de la séquence d'entrée n'ont pas la même importance pour obtenir la séquence de sortie. Le rôle du mécanisme d'attention est donc de donner de l'importance à la position des éléments au sein de la séquence d'entrée. Pour cela, on attribue un poids à chaque élément encodé en entrée. Ces poids sont ensuite mis à jour après chaque sortie produite. Le décodeur calcule alors un vecteur de contexte en faisant une somme, pondérée par ces poids, des éléments encodés. Ce vecteur de contexte permet au décodeur de prêter davantage attention à certains éléments pour produire sa sortie. À chaque sortie produite par le décodeur, le vecteur de contexte est recalculé.

3.3 Applications des réseaux de neurones séquence-à-séquence

Dans cette section, nous étudierons divers travaux réalisant des tâches séquence-à-séquence en traitement automatique du langage naturel à l'aide de réseaux de neurones. De cette façon, nous pourrons nous en inspirer en conservant leurs atouts.

Tout d'abord, les réseaux de neurones sont très utiles dans ce domaine car ils parviennent à capturer le sens des mots. [Mikolov et al., 2013] a mis en avant ce phénomène. Dans cet article, des phrases (séquence de mots) étaient mises en entrée d'un RNN classique. Les mots étaient encodés en *« one-hot »*, c'est-à-dire que le mot est représenté par un vecteur de dimension égale à la taille du dictionnaire. Ce vecteur contient des 0 partout et un 1 au niveau de la dimension correspondant au mot encodé. Cette représentation est souvent utilisée pour ce type de travaux. Le réseau de neurones essayait de prédire une distribution de probabilités des mots. En observant la sortie d'une couche intermédiaire, on obtient une représentation vectorielle de l'entrée. Les auteurs ont alors montré que cette représentation, appelée *« embedding »*, modélise les informations sémantiques et syntaxiques des mots. En effet, on remarque que le décalage (distance et orientation) entre le vecteur du mot *« man »* et celui de *« woman »* est le même qu'entre les mots *« king »* et *« queen »*. Les mêmes constatations sont faites pour les pluriels ou encore les conjugaisons.

Dans [Sutskever et al., 2014], les réseaux de neurones sont utilisés pour l'apprentissage séquence-à-séquence. Cet article propose en effet un modèle pour faire de la traduction automatique. Ce modèle a pour but d'estimer la probabilité qu'une séquence de mots soit la traduction correcte de la séquence de mots en entrée. Pour cela, il contient deux composantes. La première est un réseau LSTM avec 4 couches cachées qui prend en entrée la séquence de mots source et produit un vecteur représentant la séquence. La seconde est un LSTM simple qui prédit une traduction à partir de ce dernier vecteur. On retrouve donc la structure d'encodeur-décodeur. Le système commence par lire la séquence d'entrée mot par mot. Celle-ci se termine par un symbole qui en indique la fin. Une fois ce symbole lu, le système produit la séquence de sortie également mot par mot, en la terminant par un symbole de fin. Les auteurs de cet article ont remarqué que donner la séquence d'entrée à l'envers améliorait la qualité des traductions : au lieu de prédire que la séquence $\langle A,B,C \rangle$ se traduit par $\langle W,X,Y,Z \rangle$, le système était entraîné à prédire que $\langle C,B,A \rangle$ se traduit par $\langle W,X,Y,Z \rangle$.

Dans ce domaine, une grande partie des réseaux de neurones sont des LSTM. Toutefois, des variantes sont parfois utilisées. On peut évoquer les LSTM bidirectionnels (BLSTM) qui permettent de lire l'entrée en entier avant de produire une sortie. Pour cela, pendant qu'un LSTM lit les données (de gauche à droite), un autre LSTM les parcourt dans l'autre sens. La sortie produite peut être la concaténation des sorties de ces deux réseaux. Dans [Rao et al., 2015], un modèle complexe a été utilisé pour produire une conversion graphème (plus petite unité d'écriture) vers phonème (plus petite unité sonore permettant de distinguer deux mots dans une langue). Ce modèle utilise un LSTM et un BLSTM en parallèle, puis un second LSTM reçoit les sorties de ces deux réseaux pour prédire les phonèmes.

Dans [Hermann et al., 2015], plusieurs modèles ont été mis en compétition afin de réaliser une tâche de compréhension écrite. Chaque système recevait en entrée un texte (le contexte) ainsi qu'une question. Il devait prédire la réponse à la question grâce au contexte associé. Ces sytèmes renvoient la réponse qui maximise la probabilité que la réponse soit exacte sachant le contexte et la question. Les modèles les plus performants ont été « le lecteur attentif » et « le lecteur impatient ». Ces derniers utilisent un mécanisme d'attention pour repérer quelle partie du contexte est la plus à même de répondre à la question. Le « lecteur attentif » est un modèle comportant deux composants. Le premier composant est un LSTM bidirectionnel chargé d'encoder le contexte. Le second

est un autre LSTM bidirectionnel qui encode la question. Ensuite, une représentation de la paire contexte/question est produite à partir d'une somme pondérée des encodages et permet au système de prédire la réponse. Le « lecteur impatient » est semblable au précédent. Cependant, tandis que les deux encodages sont indépendants pour ce dernier, le « lecteur impatient » relit le contexte à la lecture de chaque élément de la question afin de produire un encodage propre à chacun de ceux-ci.

Comme le montre ces différents travaux, les réseaux de neurones sont utilisés avec succès pour traiter de nombreux problèmes de traitement automatique du langage naturel. Ces réseaux sont de différentes complexités et de différentes formes. On remarque toutefois que ce ne sont pas les seuls facteurs ayant un rôle dans la qualité des systèmes. On peut par exemple citer l'astuce de [Sutskever et al., 2014] qui consiste à inverser la séquence en entrée. Comme nous avons fait pour les disfluences, une bonne compréhension du phénomène que l'on veut modéliser peut en effet contribuer aux performances des systèmes.

4 Conclusion

Dans cette revue bibliographique, nous avons cherché à mettre au jour les éléments utiles à la production automatique de disfluences. Nous avons ainsi pu voir que les disfluences possèdent des caractéristiques importantes pour la synthèse de discours spontanés naturels et qu'il est donc nécessaire d'apprendre à produire des textes disfluents. La production de textes disfluents pouvant être assimilée au passage d'une séquence de mots fluide à une séquence de mots disfluente, nous avons également étudié différents modèles séquence-à-séquence. De plus, nous avons porté un intérêt particulier aux réseaux de neurones car de nombreux problèmes issus du domaine du traitement automatique du langage naturel sont traités avec succès par ces modèles. Cependant, il n'existe pas encore de réseaux de neurones chargés de produire des textes disfluents. Cette tâche est partiellement traitée (uniquement pour certaines disfluences) par des modèles de langages et des modèles probabilistes. La méthode récemment proposée au sein de l'IRISA [Qader, 2017] utilise par exemple des modèles de langage et des CRF.

Le sujet de mon stage intervient dans ce contexte. Les réseaux de neurones étant efficaces pour d'autres problèmes proches, il est intéressant de les utiliser pour produire des textes disfluents. Un premier travail à effectuer durant ce stage sera de choisir le type précis, ainsi que la topologie du modèle neuronal à utiliser. À cause des dépendances entre les mots, on pourra commencer par un réseau GRU qui est moins complexe qu'un LSTM. Le travail de production de textes disfluents peut être vu comme un problème de traduction « texte fluide » (langage source) vers « texte disfluent » (langage cible). C'est pourquoi les modèles dédiés à la traduction automatique seront une bonne source d'inspiration. Il faudra aussi s'intéresser au jeu de données dont nous disposerons. En effet, il y aura sûrement un travail de préparation des données à réaliser. Selon ces dernières, il faudra peut-être les nettoyer de toutes disfluences afin de pouvoir entraîner le modèle neuronal. Cette tâche peut s'avérer compliquée en cas d'imbrications de disfluences. D'autres questions, comme la stratégie de production des disfluences, seront à prendre en considération. Il faudra en effet se demander s'il faut proposer un modèle par disfluence ou un modèle global, ou encore s'il faut se restreindre à certaines disfluences.

Références

[Adell et al., 2007] Adell, J., Bonafonte, A., and Escudero, D. (2007). Filled pauses in speech synthesis : towards conversational speech. *Lecture Notes in Computer Science*, 4629.

[Adell et al., 2012] Adell, J., Bonafonte, A., and Escudero, D. (2012). Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. *Speech Communication*, 54 :459–476.

[Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

[Betz et al., 2015] Betz, S., Wagner, P., and Schlangen, D. (2015). Micro-structure of disfluencies : Basics for conversational speech synthesis. *Proceedings of Interspeech*.

[Blankenship and Kay, 1964] Blankenship, J. and Kay, C. (1964). Hesitation phenomena in spontaneous english speech : A study in distribution. *Word*, 20 :360–372.

[Bruner and Deshpande, 2017] Bruner, J. and Deshpande, A. (2017). `https://www.oreilly.com/learning/generative-adversarial-networks-for-beginners`.

[Cho et al., 2014] Cho, K., Merriënboer, B. V., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

[Christenfeld et al., 1991] Christenfeld, N., Schachter, S., and Bilous, F. (1991). Filled pauses and gestures : it's not coincidence. *Journal of Psycholinguistic Research*, 20.

[Chung et al., 2014] Chung, J., Cho, K., Gülçehre, Ç., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

[Constant et al., 2011] Constant, M., Tellier, I., Duchier, D., Dupont, Y., Sigogne, A., and Billot, S. (2011). Intégrer des connaissances linguistiques dans un crf : application à l'apprentissage d'un segmenteur-étiqueteur du français. *TALN*.

[Dall et al., 2014] Dall, R., Tomalin, M., Wester, M., Byrne, W., and King, S. (2014). Investigating automatic and human filled pause insertion for speech synthesis. *Proceedings of Interspeech*.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, pages 2672–2680.

[Greff et al., 2017] Greff, K., Srivastava, R., Koutnik, J., Steunebrink, B. R., and Schmidhuber, J. (2017). LSTM : A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28 :2222–2232.

[Hassan et al., 2014] Hassan, H., Schwartz, L., Hakkani-Tür, D., and Tur, G. (2014). Segmentation and disfluency removal for conversational speech translation. *Proceedings of Interspeech*.

[Hermann et al., 2015] Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. *Advances in Neural Information Processing Systems 28*.

[Lafferty et al., 2001] Lafferty, J., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289.

[Maclay and Osgood, 1959] Maclay, H. and Osgood, C. E. (1959). Hesitation phenomena in spontaneous english speech. *Word*, 15 :19–44.

[Mikolov et al., 2013] Mikolov, T., Yih, W., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of the North American Chapter of the Association for Computionnal Linguistics : Human Language Technologies*, pages 746–751.

[Pascanu et al., 2013] Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *Proceedings of the 30-th International Conference on Machine Learning*, 28.

[Qader, 2017] Qader, R. (2017). *Pronunciation and disfluency modelling for spontaneous speech synthesis*. PhD thesis, University of Rennes 1.

[Qader et al., 2014] Qader, R., Lecorvé, G., Lolive, D., and Sébillot, P. (2014). Ajout automatique de disfluences pour la synthèse de la parole spontanée : formalisation et preuve de concept. *Proceedings of TALN*.

[Rao et al., 2015] Rao, K., Peng, F., Sak, H., and Beaufays, F. (2015). Grapheme-to-phoneme conversion using long-short-term memory recurrent neural networks. *IEEE International Conference on Acoustics, Speech and Signal Processing*.

[Rose, 1998] Rose, R. L. (1998). *The communicative value of filled pauses in spontaneous speech*. PhD thesis, University of Birmingham.

[Shriberg, 1994] Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies*. PhD thesis, University of California.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associate, Inc.

[Tree, 1995] Tree, J. E. F. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, 34 :709–738.

[Tree, 2001] Tree, J. E. F. (2001). Listeners' uses of um and uh in speech comprehension. *Memory and cognition*, 29 :320–326.

[Young et al., 2017] Young, T., Hazarika, D., Poria, S., and Cambria, E. (2017). Recent trends in deep learning based natural language processing. *Computation and Language*.