Traitement automatique du style dans le langage naturel : quelques contributions et perspectives

Gwénolé Lecorvé

UNIVERSITÉ DE RENNES 1

UMR IRISA

HABILITATION À DIRIGER DES RECHERCHES

Université de Rennes 1

Spécialité informatique

Traitement automatique du style dans le langage naturel : quelques contributions et perspectives

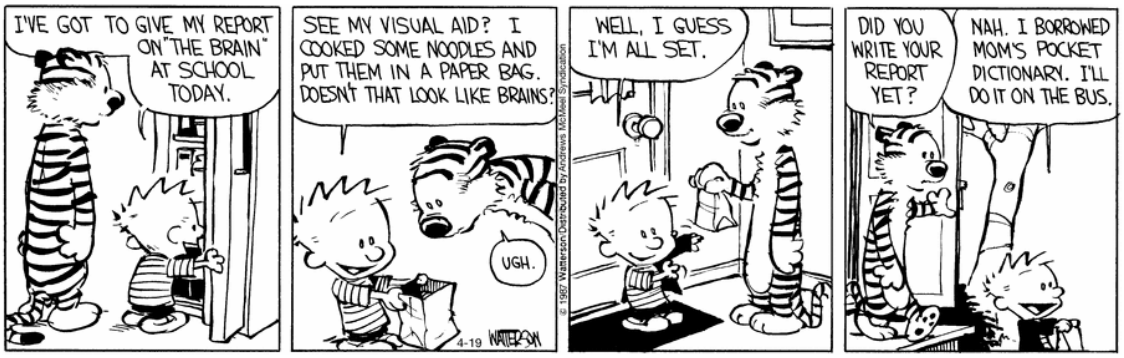
Natural Language Style Processing : Some Contributions and Perspectives

Gwénolé Lecorvé

Juillet 2020

Soutenue le 17 novembre 2020,
devant le jury composé de :

Élisa FROMONT Professeure des universités, Université de Rennes 1	Présidente
Laurent BESACIER Professeur des universités, Université Grenoble-Alpes	Rapporteur
Pierrette BOUILLON Professeure ordinaire, Université de Genève, Suisse	Rapportrice
Mathew MAGIMAI-DOSS *Senior researcher*, *Idiap Research Institute*, Suisse	Rapporteur
Iris ESHKOL-TARAVELLA Professeure des universités, Université Paris Nanterre	Examinatrice
Emmanuel MORIN Professeur des universités, Université de Nantes	Examinateur



(*Calvin and Hobbes*, par Bill Watterson, 22 avril 1987)

Remerciements

Comme on a tendance à le dire, le doctorat couronne un travail individuel alors que l'Habilitation à Diriger des Recherches (HDR) met en lumière un travail d'équipe, ou même, pour être exact à mon sens, des travaux de plusieurs équipes. Le reste du document l'illustre clairement mais je tenais à souligner explicitement ma gratitute à travers cette courte section.

Avant de faire le tour de ces différentes équipes, je tiens à remercier les membres de mon jury dont la présence et les échanges, suite aux retours de pré-soutenance des rapporteurs ou de chacun à la soutenance, m'honorent et contribueront certainement à nourrir mon futur scientifique.

Ensuite, j'adresse mes sincères remerciements aux différents et nombreux collègues avec qui j'ai collaboré depuis ces dix dernières années. Merci ainsi aux collègues « suisses » (Petr, John D., Phil ainsi que toute la clique des non permanents de l'époque) qui m'ont ouvert à d'autres pratiques et problématiques. À l'IRISA, et à Lannion en particulier, j'ai eu le plaisir et la chance d'être engagé dans la dynamique de création de l'équipe Expression, entre Lannion, Vannes et Lorient, et à la croisée des modalités orales, textuelles et gestuelles. Ce document et cette habilitation sont le fruit des réflexions, discussions et collaborations autour de la notion-clé d'expressivité dans le langage humain. Ainsi, je tiens à remercier tous les collègues d'Expression, avec une emphase plus particulière pour les doctorants et chercheurs post-doctorants avec qui j'ai travaillé (les travaux présentés ici sont les leurs avant d'être les miens), ainsi que Damien et Nicolas avec qui nous avons, dès le début, monté de multiples projets ou co-encadrements. À l'extérieur, je me dois également de remercier Delphine avec toute mon amitié et ma reconnaissance pour l'aventure scientifique et humaine que nous avons co-construite. Bien sûr, je tiens également à remercier tous ceux qui contribuent à un environnement propice à la recherche, au sein de l'IRISA ou de l'ENSSAT.

Pour terminer, un grand merci à mes proches qui ont participé, certes indictement mais indiscutablement, à l'aboutissement de cette étape-clé qu'est la HDR. Notamment, un immense et tendre merci à Marie pour son soutien et sa compréhension sans faille (notamment au moment de la rédaction en plein confinement !), ainsi qu'à Saskia, en dépit de la réduction drastique de la durée utile d'une journée normale de chercheur. Vous êtes les rayons de soleil de ma photosynthèse quotidienne.

Table des matières

Part I

English summary

CHAPTER 1
Introduction

The study of natural language is present in multiple scientific disciplines. In linguistics, this study is purpose itself and comes in many specific areas. In neuro-sciences also, language is at the heart of some works to understand how this capacity develops or deteriorates in humans. As a communication and bonding tool, anthropological disciplines study it to analyze society, relationships between humans, history or even art. Finally, natural language processing (NLP) [1] is a historic area of computer science. As the current boom in many so-called "smart" commercial products shows, this is a relatively mature field. However, in my opinion, two observations point out shortcomings. Firstly, while NLP has long focused on the meaning of the statements, the processing of style remains an aspect which is still rather poorly mastered. Yet, style reflects multiple factors useful in the analysis of a sentence/utterance and its context. Secondly, the dialogue between NLP and human sciences is relatively weak and even tends to decrease if we refer to the practices spreading in the deep learning techniques.

On the first point, the preponderance of studying the meaning rather than style undoubtedly comes from a logical prioritization to advance the field. In the same way as someone learns a new language, it is indeed more important to understand or be understood than to master the different ways that say something. Then, style processing contributes more to the natural language generation tasks than to analysis and recognition. Since the popularity of these production tasks is relatively recent (speech synthesis, chatbots, paraphrase generation, etc.), the rise of style processing is rather recent as well.

Regarding interdisciplinarity, the weakness of the dialogue is probably explained, in addition to the classic difficulty in agreeing on a shared vocabulary, by the predominance of statistical approaches in NLP, starting from the 1990s. In this framework, tasks are resolved based on the estimation of average phenomena from large amounts of data. This differs from the other disciplines which often aim to exhibit multiple explanatory dimensions from an initial average phenomenon. Furthermore, the transition to deep learning for almost all tasks in NLP accentuates the lack of dialogue. Indeed, for a given task, this paradigm tends to skip feature design phase and, thus, makes it useless to talk with with linguists, psychologists, sociologists, etc. From now on, the models often work directly on raw data and incorporate preliminary transformations in place of expert descriptors.

This habilitation manuscript is a synthesis of my research activities since obtaining my PhD in 2010. These activities, which are part of supervisions and collaborations, reflect my career development through the fields of automatic speech recognition (2007-2012), speech synthesis (since 2012) and written language processing (since 2017). Without claiming to inflect the observations just drawn up, my work has in common to contribute to the question of the *automatic processing of style* in natural language, whether oral or written. Moreover, although machine learning plays a primordial role in my work, the integration of considerations or problems arising from linguistics

1. In French : *Traitement Automatique du Langage Naturel, TALN*.

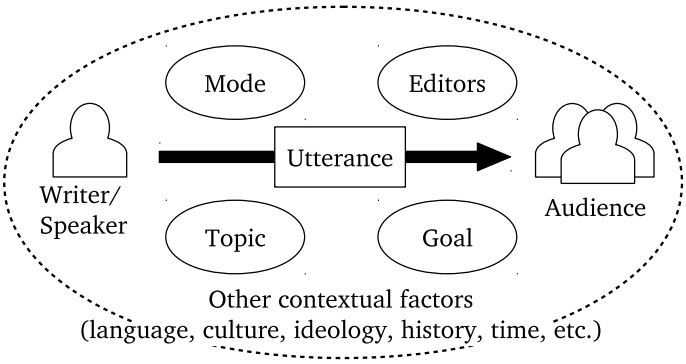


FIGURE 1.1 – Influential factors in verbal communication (adapted from (ARGAMON 2019)).

is another aspect which, I believe, gives coherence to my work. Before detailing the organization of this manuscript in Section 1.4, this introduction sets the general framework. Thus, Section 1.1 discusses the mechanisms which drive to natural language production. Section 1.2 presents some examples that illustrate the impact of these mechanisms on the linguistic observations of produced utterances. From there, we analyze in Section 1.3 elements defining the notion of style.

1.1 Building an utterance

In the stratified vision of language, linguistics defines multiple levels of abstraction (phonology, lexicon, morphology, morphosyntax, syntax, etc.). These levels provide tools to describe an utterance from its physical form, i.e. its signal or characters, to its meaning (semantic level) and how it fits into a more general context (pragmatic level). When comparing utterances, these levels are used to characterize differences. For example, the use of more or less simple syntactic structures, the use of more or less rich vocabulary, or the fluidity of speech are observable characteristics.

Nevertheless, these levels of abstraction do not explain the causes of these observations, except maybe through the general existence of the pragmatic level, although these causes are useful to get a fine understanding of a content. As summarized in the communication schema of Figure 1.1, the final form of an utterance is conditioned by multiple factors. First of all, it depends on the writer/speaker (his/her state of mind, opinion, age, social background, level of expertise), the *audience* (*idem*) and their relationship (social relation, emotional relation, shared knowledge, state of discussion, unit of time and place, etc.). The *topic* is also important. It obviously influences the semantic level, but it can also call for reactions from the person producing or receiving the content, and thus impact the style of communication. The *goal/intention* (e.g., inform, dictate, entertain), the mode (speech, novel, blog, email, TV show), and eventual text editing mechanisms usually impose codes to the form of the message. For example, news speech will tend to have a neutral tone and a common level of language, characterized by a syntax and lexicon that is not faulty and generally affordable to everyone. Finally, many other more general factors influence the utterance : language ; societal factors such as culture, ideologies, etc. ; or the temporal or spatial characteristics of a speech.

1.2 Variation examples

The causal links between contextual factors and linguistic observations are diverse. To situate the work developed in this manuscript, Table 1.1 lists pairs of utterances A and B that vary at different linguistic levels. For each variation, possible contextual factors are given. First of all, examples (1) and (2) show situations where the meaning differs. The other linguistic levels also

	Utterance A	Utterance B	Observable variations	Potential causes
(1)	Flowers bloom in the spring.	Parliament voted on the text.	Lexicon, syntax, semantics, pragmatics	Different topics
(2)	I love that story.	I hate this script.	Lexicon, syntax, prosody, semantics...	Sentiment, emotion
(3)	Before he left, he greeted me.	He greeted me and left.	Syntax, Lexicon	Target audience, author's command of the language
(4)	She works hard.	She works selflessly.	Syntax, Lexicon	Social origin, level of education, target audience
(5)	I have a nasopharyngitis.	I have a cold.	Lexicon	Target audience, expertise of the author
(6)	She's diligent.	She's zealous.	Lexicon, semantics	Sentiment
(7)	I want it.	That's what I want.	Syntax, prosody	Level of willingness, State of dialogue
(8)	"Like, uh... As you wish."	"As you wish."	Syntax, lexicon, prosody	Emotion, intention
(9)	"How many?"	"How- how- how many?"	Morphology, phonology, prosody	Emotion, pathology
(10)	"I didn't know that."	"I didn't know that."	Phonology, phonetics, prosody	Speaker/audience relationship, level of education, emotion
(11)	"Hello, Mommy."	"Good.day ma.man" (over-articulated)	Prosodie	Purpose (dictation?), sound environment, pathology
(12)	"How are you doing?"	"I'm fine."	Prosody, semantics	State of dialogue
(13)	*<Low-pitched Speech signal>*	*<High-pitched Speech signal>*	Acoustics	Age, gender of speaker, emotion
(14)	cu tmr! ;-)	See you tomorrow! Ha ha!	Lexicon, morphology, syntax	Media

TABLE 1.1 – Examples of variations between pairs of utterances.

vary because they are partly the instrument of the semantic level. In the other examples, the meaning varies little. Utterance pairs 3 to 7 are syntactic and lexical reformulations. Some vary in terms of complexity (3-5), perhaps due to varying degrees of language proficiency of the writer/speaker or target audience. Others (6-7) may reflect the speaker's state of mind (a feeling, a desire, etc.). Examples 8 to 12 bring together speech-specific variations, e.g. disfluencies at the lexical (8) or infra-lexical (9) level, variations in pronunciation (10), articulation (11) or intonation (12). Here again, the causes can be varied, although easily imaginable. Finally, other variations are specific to the mode of communication (speech signal, handwritten text, typed on a telephone). For example, the fundamental frequency of a speech signal may reflect information about the speaker (13). In yet another form, example 14 shows that the graphic codes of the language can be modified (SMS, Twitter, forums).

In this manuscript, the scope of study ranges from examples (3) to (10), i.e. we are not interested in semantic variations or those related to the medium. Nor is the aim to construct a general causal model. More modestly, we study particular cases, already highlighted by the linguistic literature, and propose computational models. The ambition of a unified model of these different cases is one of the perspectives that we will discuss at the end of the manuscript.

1.3 What is style ?

Although the aim of this document is not to propose a definition of style, nor to defend a particular approach, it is interesting to observe the interpretations that are made of it in human sciences and NLP in order to position the various works that we will present later.

In the social sciences and humanities, the notion of style is not uniformly established. It is therefore preferable to define style as the object of stylistic analysis. From a literary point of view, this analysis focuses on the artistic efforts and strategies of an author (SEBEOK 1960). In psychology, it studies the perception of a message (RIFFATERRE 1961). In particular, rhetoric is interested in the effectiveness of this message. The style is then often linked to the oral modality and to a logic of taking a position (JOHNSTONE 2009). In sociological fields, it can be associated with the coherent use of language within a linguistic community—a concept that we would rather qualify as a register in this manuscript (*cf.* chapter 4), as noted by the way ECKERT et RICKFORD (2001). Thus, quoting (CRESSOT et GALLO 1969), "the stylistic fact is therefore linguistic, psychological and social at the same time : we must be understood." In this, a consensual linguistic vision consists in approaching style as the set of choices made by the writer/speaker to use certain tools offered by the language (1) in order to facilitate the understanding of his/her utterance and, therefore, (2) according to the context of enunciation.

From the perspective of automatic processing, style can be more pragmatically defined through its association to meaning, the former modulating the latter with information that is relevant to the context of the enunciation. Among the questions surrounding this association, a major one is then whether style and meaning can be considered independent or intertwined. The practical consequence of a hypothesis of independence is that a given style can be decomposed into a set of explicit linguistic elements such as specific words, markers or syntactic structures (LEEFTINK et SPANAKIS 2019 ; LI et al. 2018). In the other, style is considered as a holistic concept, i.e. an implicit and integral component of a text/speech. To be able to distinguish style and meaning, models must then move from the linguistic space, where the two concepts cannot be separated, to latent spaces (TIKHONOV et YAMSHCHIKOV 2018). This concept has the advantage of being easily generalized to many practical cases. Indeed, given two corpora supposedly of different styles, style can be seen there as the dimension that invariably discriminates between the texts of these two corpora.

1.4 Organization of the manuscript

This document reflects my exploration of style-related issues in NLP over the past ten years. It is organized in short chapters, which are presented in the form of successive summaries grouped around common questions. Precisely, the chapters 2 to 5 present works on well-identified stylistic variations :

- Chapter 2 deals with modeling and adapting phoneme sequences to incorporate variability and accommodate different pronunciation styles.
- Chapter 3 describes work done on the insertion of disfluences in a text in order to give it a more natural, more human dimension.
- Chapter 4 deals with the question of the casual, neutral and formal registers.
- Chapter 5 provides an overview of work on language modeling for children.

Then, Chapters 6 and 7 deal with more transversal subjects :

- Chapter 6 discusses work related to the encoding of linguistic information by neural networks.
- Chapter 7 presents my activities of software development for research and technology transfer.

Finally, Chapter 8 brings together the research perspectives resulting from these various studies over a new ten-year horizon.

CHAPTER 2
Summary of the main contributions

The chapter summarises the main contributions that are described with more details in the French part of the manuscript. For each of them, the reference to the corresponding French chapter is given.

2.1 Pronunciation variants

The pronunciation associated with the words of a language is the key element that links the written and spoken worlds. In NLP, phonetisation is the task of predicting the pronunciation of words or statements by associating a sequence of phonemes with a sequence of graphemes. This step can be difficult because pronunciation is subject to many variations, e.g. depending on the speaker's habits or pathology, the degree of familiarity with the audience, or a specific accent (regional or foreign). These variations result in insertions, deletions and substitutions of phonemes with respect to the standard (also called *canonic*) pronunciation.

Most phonetisation tools mainly rely on manually constructed pronunciation lexicons for the common words of the language and on automatic grapheme-to-phoneme conversion for the others. For most languages, the phonetisation of an utterance is then limited to the concatenation of the individual pronunciations of its words. However, this approach is not viable for some languages (e.g. French) where transitions between words cause phonetic changes.

The production of pronunciation variants can serve several purposes. It may be useful to model all possible variants in a given language in a compact and exhaustive way or to preserve as much information as possible for post-processing. Alternatively, the objective may be to introduce new pronunciations from the canonical ones to imitate or adapt to a particular context. The task can then be seen as phoneme-to-phoneme conversion.

In Chapter 2, I present different contributions to these problems in the context of speech synthesis. The overall aim is to be able to control the phonetisation stage in order to imitate certain expressive features. The first work concerns the construction of phoneme lattices which represent possibilities of allowed variants in the language. Adaptation to different styles can then be achieved by reweighting arcs, without the risk of producing incorrect variants. A second approach is the production of style-specific variants thanks to an adaptation model. In this case, the allowed variants are more free. We have shown that this process is efficient and generic as we have successfully applied it to three different styles.

This work has been integrated into the IRISA speech synthesis system. The main limitation observed is the difficulty of the system to produce the signal of pronunciations different from those used for its construction. Despite advances in neural network approaches, this remains a problem and therefore a major perspective to be developed.

2.2 Automatic insertion of disfluencies

Disfluencies are a phenomenon that interrupts speech without adding any propositional content (TREE 1995). They occur mainly when speech proceeds faster than the thought process, which is particularly frequent when the speaker has not prepared her/his speech. Disfluencies play an important role in speech. They improve comprehension by listeners, signal the complexity of future utterances and, in the case of a dialogue, facilitate synchronisation between speakers (CLARK 2002). Many studies agree—although the terminology may differ—on a categorisation of disfluencies into three main families : *pauses*, *repetitions* and *revisions*(BOULA DE MAREÜIL et al. 2005 ; E. E. SHRIBERG 1999 ; TSENG 1999).

In NLP, most of the work on disfluencies is in the field of automatic speech recognition (HASSAN et al. 2014 ; KAUSHIK et al. 2010 ; LIU et al. 2006 ; STOLCKE et E. SHRIBERG 1996 ; STOLCKE, E. SHRIBERG et al. 1998). The main objective of these works is to integrate this phenomenon into the language model of the systems or to produce automatic transcriptions that are cleaned of any possible disfluencies. Thus, they have been more interested in the surface form of disfluencies than in the process that leads to their production. Studies in speech synthesis have been more scarce (ADELL, BONAFONTE et ESCUDERO 2007 ; ADELL, BONAFONTE et MANCEBO 2008 ; ADELL, ESCUDERO et al. 2012 ; DALL et al. 2014 ; SUNDARAM et NARAYANAN 2003). ADELL, BONAFONTE et MANCEBO (2008) explain this in part by the absence of disfluencies in the speech corpora on which the synthesis systems rely and the increased difficulty for linguistic pre-processing to work on disfluent texts. On the whole, the work on adding disfluencies focuses on the insertion of pauses, possibly categorized as types parencitebetz2015micro, but it does not propose anything for other disfluencies. Thus, unless the user makes disfluencies explicit in the text to be synthesised, the signals produced are generally perceived as unspontaneous.

Chapter 3 focuses on my work on the automatic insertion of disfluencies from the perspective of synthesising speech with a more spontaneous style. We present two techniques for inserting disfluencies, designed with a view to bringing greater expressive power to speech synthesis systems. The first approach is based on a linguistic framework derived from the literature where a elementary disfluencies are composed. The second, less accomplished, relies on sequence-to-sequence neural networks. Finally, we point out that a common aspect of this work is the difficulty of evaluating this task, as the most reliable solution of recording an actor to play back utterances is also the most costly.

In a similar way to the conclusions on pronunciation modeling, the main obstacle to the continuation of this work is the current difficulty in having disfluencies rendered by speech synthesis systems. Nevertheless, it should be noted that the automatic insertion of disfluencies is already perfectly feasible in a purely textual context, for example to humanise *chatbots*. In this case, it is likely that the families considered will have to be adapted to the specificities of digital writing (other linguistic markers, typos, emoticons, etc.).

2.3 Casual, neutral and formal registers

The notion of register refers to the way in which linguistic productions are evaluated and categorised within the same language community. It is addressed in numerous works in linguistics as well as in sociolinguistics. Thus, FERGUSON (1982) defines registers as a variation "*in which the linguistic structure varies according to the occasions of use*". URE (1982) associates this variation with human activities : "each language community has its own system of registers corresponding to the range of activities in which its members normally engage". Depending on the angle of study, there are various ways of partitioning the language space into different registers in the linguistic literature. For example, it may be a question of distinguishing between colloquial, popular and vulgar registers in satirical newspapers (ILMOLA 2012), the influence of different communication media (CHARAUDEAU 1997), degrees of specialisation (BORZEIX et FRAENKEL 2005 ; MOIRAND 2007)

or the opposition between functional and relational communication (BORZEIX et FRAENKEL 2005). This diversity also gives rise to a terminological difficulty, since the terms “level”, This diversity also reveals a terminological difficulty, since the terms “level”, “style” or “gender” co-exist with that of “register” and are the subject of debates (BELL 1984 ; BIBER 1991 ; BIBER et FINEGAN 1994 ; GADET 1996a,b). Since our work does not aim to make contributions on these definitional aspects, we arbitrarily adopt the term register, which comes from the British tradition (SANDERS 1993 ; URE 1982).

Chapter 4 transcribes work carried out on the case of casual, neutral and formal registers. This division is primarily motivated by pragmatism, as it is relatively consensual and unambiguous for the manual labelling of an initial dataset, while not precluding possible refinements for the future. The work summarised in this chapter is part of the exploratory ANR TREMoLo project, which I coordinate. This project aims at studying language registers and developing automatic methods for transforming texts from one register to another. This work is based on the extraction of linguistic patterns specific to given registers and on their consideration in a paraphrase generation process.

I review different contributions to the modelling of the notion of language register in NLP : the corpus validation of descriptors from the linguistic and sociolinguistic literature, the automatic construction of a large annotated textual corpus, and the automatic extraction of discriminating patterns. These different works are placed in a coherent aim of replicability to other textual styles. One of their originality is their positioning at the intersection of linguistics and data mining, each of which brings its own challenges. On the one hand, it is difficult to circumscribe the registers of colloquial, common and sustained language as an object of study in comparison with other notions of linguistics (genre, level, style, mode). On the other hand, the complexity of the classical mining algorithms makes their application difficult.

Finally, let one note that the paraphrase generation side of the project has recently started and is discussed in the conclusion and perspectives.

2.4 Texts for children

How an individual understands a text depends on the characteristics of the text and the abilities of the individual. This is particularly true according to age since, during childhood, cognitive, linguistic and cultural capacities change a lot. Echoing this, it is preferable to adapt your statements when addressing children. On this principle, multiple media offer children better access to current events and information from the web while respecting their cognitive abilities. These can be dedicated portals (Qwant Junior in France, Yahoo ! Kids in Japan), textual content (encyclopedias like Wikimini, Vikidia or Simple Wikipedia encyclopedias ; newspaper like Le P’tit Libé, Albert or Le Petit Quotidien) or multimedia content (“1 day 1 question” video pastilles, “1 day 1 news”, kids versions of Youtube and Netflix, etc.). Unfortunately, these initiatives are based on the work of experts and therefore suffer from limited power of impact. Artificial intelligence techniques are therefore an important lever for developing these initiatives.

In psycholinguistics, children’s understanding of language is a widely studied area. The work highlights factors such as the role of short-term phonological memory (GATHERCOLE 1999), the acquisition of temporal notions (HICKMANN 2012 ; TARTAS 2010) and the linguistic constructions that accompany them (VION et COLAS 1999), or the coherence and complexity of the emotions present in a story (BLANC 2010 ; DAVIDSON 2006 ; MOUW et al. 2019). For their part, studies in learning to read provide elements of analysis as to the ease that children may have in deciphering a text. FRITH (1985) argues that reading is acquired in three main stages related to the recognition of word-forms, then graphemes and, finally, morphemes. Subsequently, understanding the language is accompanied by the acquisition of knowledge about the language (syntactic construction, vocabulary, etc.). On the oral level, other works note that the intonation during the reading of a text—induced by the lexicon, the punctuation, the syntax, etc.—influences the perception of a text and that the understanding of this intonation evolves with age (AGUERT et al. 2009).

Finally, computational approaches have existed for a long time to link the readability of a text to a level of study. Historically, these are based on lexical and syntactic complexities, like the Flesch-Kincaid (FLESCH 1948) index, or the Dale-Chall formula which also considers the notion of "difficult" words (DALE et CHALL 1948). More recently and more generally in NLP, works on the simplification of the text for children (DE BELDER et MOENS 2010 ; GALA et al. 2018) or on the readability of texts (FRANÇOIS 2015 ; FRANÇOIS et FAIRON 2012) have been proposed.

Chapter 5 present my work on texts addressed to children. This is part of the collaboration with the University of Paris-Nanterre, which recently gave rise to the launch of the ANR TextToKids project. This line of research combines NLP and psycholinguistics issues. We have approached a first work on the prediction of age recommendation. This is a singular task in the literature for which these first results are very encouraging. Other work has focused on the linguistic peculiarities of temporality and emotions in journalistic texts for children. This work complements the first one in the sense that they open up the possibility in the future of explaining in natural language to authors inadequate elements of their text being written.

Beyond that, much work is still to be done to refine linguistic analyzes and generalize them to other genres and dimensions. In addition, the prediction models presented in this manuscript are still rudimentary and we are currently experimenting with more sophisticated ones.

2.5 Neural approaches to encode linguistic information

In a decade of work in NLP, neural networks have gone from being a tool for implementation to being the focus of proposed advances. This change has significantly altered the approach to problems. The current trend, often supported by the results, is to be agnostic about the nature of the data and the properties of the phenomena to be treated. It is thus relatively similar to build a speech synthesis system or a text simplification model. Linguistic expertise is replaced by deep learning expertise, which is necessary to understand and train the sometimes complex architectures of the state of the art.

Deep learning has also given rise to a new use of neural networks : that of encoding information thanks to the concept of embedding. This encoding has the advantage of converting heterogeneous sets of information (symbolic or numerical, of multiple dimensions, sequential or non-sequential) into homogeneous latent representations in a high-dimensional vector space. It also opens the way to the mathematical tools of these spaces (distances, transformations). The most striking example in NLP is probably the Word2Vec (MIKOLOV, SUTSKEVER et al. 2013) method, which introduced a new representation of words such that linguistic properties (particularly semantic ones) are transposed into geometric properties. Then, thanks to their homogeneous form (fixed-size real-valued vectors), the embeddings allow to simply interface many neural modules together, even if it means learning them at the same time.

Chapter 6 presents various works carried out in recent years on the encoding of linguistic information. Each time, the aim is to understand the properties of the embedding of some linguistic objects in order to exploit them according to the task in hand. I first studied the problem of language models in automatic speech recognition, in which embeddings integrate morphosyntactic and semantic information from word histories. For speech synthesis, another contribution uses embeddings to merge linguistic and acoustic information. Finally, generic methods for style transfer are presented as part of my current work. One of the questions is how the components assimilated to style and meaning are encoded. These examples show, if it were still necessary, the tremendous interest of the concept of embeddings. However, it also highlights the limitation that it is difficult to analyse their quality, or to compare several of them, without heavy evaluations of the whole system that incorporates them.

2.6 Tools and projects

The previous contributions have highlighted various research problems. Nevertheless, the field of NLP involves a lot of data manipulation and human interaction for which the development of software tools is necessary. It also gives rise to many applied projects aimed at facilitating or improving everyday life. Although these projects do not necessarily involve scientific contributions in NLP, they do work for the dissemination and transfer of knowledge.

These aspects, developped in Chapter 7, represent a significant part of my research activities, especially over the last two years. They cover tools developed to support my work (data annotation management, perceptual evaluation), industrialised research prototypes, and applied research projects focusing on speech synthesis, in which I participate or have participated and for which I have responsibility in some cases.

Although these activities produce little in terms of scientific publications, they are part of the research cycle. They are also rich in lessons about the multiple tasks and skills required to conduct a project from its scientific premises to its access by the general public.

Chapter 3
Conclusion et perspectives

This manuscript for research habilitation presented a synthesis of my research activities over the last ten years. Their common thread is the analysis of the variability inherent in natural language with a predominance on stylistic variations. Contributions on the processing of these variations concern both spoken and written language. In particular, Chapters 2 and 3 presented work on the modelling of phonetic variations and disfluences. Both activities are applied to the field of speech synthesis. We also discussed my work on casual, neutral and formal registers (Chapter 4), as well as on language for children (Chapter 5). These involve interactions with several fields of linguistics. Then, transversal work on the encoding of linguistic information *via* neural machine learning were presented (Chapter 6). Finally, Chapter 7 highlighted activities related to software development and dissemination through applied projects.

In the following, we present short-term research perspectives which are in the continuity of the automatic processing of style, and then more fundamental ones that contribute to a broader set of domains and aim at a ten year horizon.

3.1 Short-term perspectives

I am currently supervising 4 PhD theses : Antoine Perquin's (3rd year) on speech synthesis using neural networks ; Jade Mekki's (2nd year) on the automatic characterization of language registers by the extraction of language patterns ; Somayeh Jafaritazehjani's (2nd year) on unsupervised style transfer ; and Aline Étienne's (1st year) on temporality and emotions in children's texts. In addition, I supervise the work of several post-doctoral researchers and engineers in some projects. My short-term perspectives are a natural extension of these multiple works. I split them considering the oral, then written modality.

3.1.1 Modulation of synthetic speech

The field of speech synthesis has undergone a major evolution with the rise of deep learning and the research efforts of the digital giants. For example, the introduction of the WaveNet vocoder by Google (van den Oord et al. 2006) represented a significant gain in terms of rendering of parametric synthesis (until then rather mediocre) and ended up shifting the bulk of research towards this family, to the detriment of approaches based on unit selection. Thus, architectures such as Tacotron 2 (J. Shen et al. 2018), Deep Voice 3 (Ping et al. 2018) or FastSpeech (Ren et al. 2019) have emerged. These approaches tend to pose the problem of neutral speech synthesis as solved, and open the research to the integration of variability factors, such as multiple speakers, accents, emotions or even languages. These approaches rely heavily on the transformation of each of these factors into an embedding that conditions neural acoustic models.

— **Multi-factorial and linguistic conditioning.** In this context, my perspectives arise at different levels. First of all, the combination of factors and the objective of having a universal synthesis system are not yet solved problems. This raises questions about the conduct of the learning protocol as the number of parameters in the model becomes important. It is therefore necessary to find strategies that either drive the optimization (e.g., through dedicated loss functions) or that divide the training into several sub-trainings. Secondly, current work focuses mainly on the acoustic dimension of speech synthesis and few of them work are interested in the linguistic one (including speech-specific abstraction levels) and in the study of the text provided as input to the models. To begin, as we have shown, pronunciation and disfluencies are interesting elements for improving the expressivity of systems. Thus, it would be interesting to extend our previous work by the use of these new models. In particular, this seems promising because until now there have been problems in making the prosody consistent with these two phenomena, and one of the strengths of neural synthesis is precisely to treat the prosody well. In a more general manner, constraints coming from other linguistic levels could also be future topics of interest.

— **Disfluent speech synthesis.** On the specific question of the insertion of disfluences, we have highlighted the difficulty of evaluating the results. The recent possibility of using systems capable of theoretically rendering disfluences well should be a positive point. Even so, however, the use of such a synthesis would not be free of bias favouring situations observed when training the system (i.e. generally fluent speech). Thus, another avenue is the development of objective measurements correlated with perception indicators, as we have done in (PERQUIN, **LECORVÉ** et al. 2019). The production of a large set of perceptual annotations on a large volume of variants would then be necessary to search for and validate these correlations. Our work on disfluencies has also shown the limitations of insertion by a standard neural encoder-decoder. Nevertheless, I believe that this is a path to pursue because the reasons for these difficulties have been identified (no pre-trained word embedding, limited training data to train a recopy model). Then, as the output sequence is very close to the input sequence and the inserted elements must respect a certain structure, the task is actually also very close to text normalization, for which recent proposals would be tracks to be explored (H. ZHANG et al. 2019 ; J. ZHANG et al. 2020).

— **Democratization of the approaches.** Finally, the remarkable quality of the signals presented in many recent papers masks certain limitations. On the one hand, much work is done on English and is difficult to transpose to other languages because the necessary speech corpora does not exist or is of low size or even poor quality. Thus, the constitution or, more pragmatically, the pooling of resources is an objective for the future, as is the development of methods capable of working with very little data (zero-shot learning, few-shot learning) or on noisy data. On the other hand, the management of time and computing power, whether at training or inference time, is an element to be improved for ecological reasons and the democratization of these tools. In this respect, methods for adapting already trained systems (fine tuning) or transfer learning techniques seem to be good starting points for these studies.

3.1.2 Modulation of texts

Neural networks have also brought many changes in the field of text generation, mainly driven by advances in machine translation (BAHDANAU et al. 2014). In particular, the field of paraphrase generation has made significant progress (PRAKASH et al. 2016). As in the previous section, a considerable part of the current work is now focused on the inclusion of various constraints to control the generated texts, for example according to the desired output length (KIKUCHI et al. 2016) or a particular syntactic structure (IYYER et al. 2018).

— **Unsupervised approaches.** In terms of machine learning, paraphrase generation methods fall into two families. On the one hand, some works make the hypothesis of aligned

textual data, i.e., for each input sentence, the model can rely on the corresponding output ground truth. On the other hand, other works carry out an unsupervised approach where the corpora of source and target styles are not aligned (but still generally comparable). These approaches rely on adversarial architectures where the style essence of each corpus is automatically inferred. The unsupervised approach is more difficult to set up but has the advantage of being easily applicable to any situation, as long as specific corpora for each style are available. As discussed in Chapter 6, the main issue is the balance between style transfer, maintaining meaning and fluency/correctness of the output sequence. The development of architectures that better guarantee these aspects is therefore an issue. One way to do this is the use of adequate loss functions that work either from the sequences produced in the outputs, or directly on the embedding handled within the model. The difficulty here is to define these functions since the underlying notions are complex (style, meaning, fluency). Moreover, it is likely that the mere adoption of these functions is not enough because, where style and meaning can be seen as distinct in the space of concepts, they remain intertwined in the space of words. Thus, in my opinion, these efforts on evaluation functions must be coupled with the study of topologies which precisely propose to distinguish (disentangle) different dimensions of a sentence (BAO et al. 2019 ; JOHN et al. 2019).
Another approach that I am developing in the context of text modulation aims to make it possible to interact with humans to constrain or explain what is happening. This involves (1) the possibility of explicitly *characterizing* the reciprocal distinction between two styles, that is producing discriminant patterns that are understandable by humans, and (2) the possibility of *reformulating* texts under the constraint of these same patterns, in this case in order to favor some and penalize others. In this approach, which we apply to the language registers, work must first of all continue to allow the extraction of patterns under free conditions, i.e. on the basis of a large number of features and large distances between the pattern elements. Moreover, current techniques return a large number of (mostly redundant) patterns. The development of clustering methods for these patterns would be a step forward to simplify their analysis by humans and their possible transmission to other NLP modules. As for the reformulation of a text, further work is to be expected. To include patterns as constraints, a first topic, already under exploration, consists in studying the capacity of the encoder-decoder architectures used for the production of paraphrases to be conditioned in various ways (IYYER et al. 2018), i.e., to take into account these constraints without degrading the overall quality of the paraphrases produced. In addition, another aspect of the work concerns the possibility of conditioning these architectures with a variable (and potentially large) number of constraints without knowing in advance which ones will be applied to the successive input texts. This type of problem is in line with proposals made to provide models with an augmented memory (ZHAO et al. 2018).

— **Language for children.** Finally, future steps are also related to my activities on texts for children. We will, of course, focus on the study of more complex prediction models and finer linguistic characteristics, but, beyond that, it is important to consolidate the methodological framework of our work in order to offer the community the possibility of comparing and continuing it. This involves new collaborations with psycholinguists to refine the notion of "understanding", and surveys with children of different ages to validate our experiments. Increasing our amount of annotated data is also essential. We have largely begun this work, notably by completing the age annotation with a sub-categorisation in journalistic, encyclopaedic and fictional genres (novels, stories). The question of sharing this data is a delicate one because it is necessary to obtain the agreement of the authors or publishing houses, but we are also working on it. Finally, the issues addressed open up to more complex tasks of remediation or reformulation. While the task of reformulation has already been discussed above, "remediation" means accompanying authors to make them aware of good practices to adopt and errors to avoid, whether this is independent of any text or focused on texts in the process of being written. In the latter case, this requires knowing how to link neural

network predictions with linguistic interpretations. This task is part of the growing field of Explainable Artificial Intelligence (XAI) in which the justification of recommendations is an already well-known issue (Y. ZHANG, X. CHEN et al. 2020). Finally, another future application of adequacy predictions is the inclusion of our work in a children-dedicated search engine. In this case, as part of the TextToKids project, we are starting a collaboration with Qwant Junior to move in this direction.

3.2 Longer-term perspectives

Looking back over the decade 2010-2020, the main evolution of NLP is the deep learning revolution. Historically, the standard approach to solve a task was based successively on a linguistic analysis of the task, then the design of features, and their final assembling, usually using machine learning techniques. Today, in many situations, the end-to-end neural approach dominates. This approach generally skips the linguistic analysis stage and directly works on raw data, leaving the optimization algorithm in charge of determining the relevant processings to apply to them (selection, transformation). This change appears to be the simple continuation the statistical approaches, introduced in the 1990s, where expert knowledge is totally replaceable by recurrent observations of latent phenomena present in the data.

In this context, I am addressing more general perspectives, which exceed the sole issue of style and crystallize persistent gaps in current approaches or issues that today's trends suggest for tomorrow. These perspectives are organized around (i) issues related to neural network-based machine learning and (ii) the place of the humanities in NLP. Finally, I conclude with a (iii) more political questioning on the organization of research in artificial intelligence in the future.

3.2.1 Deep learning

With the almost general adoption of neural networks, the majority of recent work is based on a few architectures or principles (encoder-decoder, feed-forward/CNN/RNN, attention mechanisms, GANs). This evolution brings new issues, more fundamental than those listed in the previous section and, in this respect, also generally shared by domains other than NLP. I list below those that I consider to be the main ones based on my sensitivity, accompanied where possible by some initial thoughts.

— **Generality and multi-tasking learning.** The growth of NLP leads to a regular and fast production of new models. Nevertheless, their use in a practical case often requires to relearn them from scratch or to modify some aspects of the architectures. And this phenomenon is repeated with each major innovation in machine learning. This lack of genericity reflects both a dependence on learning data and on the task at hand. In this respect, I think that work in transfer learning is an important research direction to be developed. In particular, one of the possibilities offered by the literature is to learn multi-task models, i.e. models that share an upstream phase of encoding linguistic information and then break down into parallel branches that specialize in each respective task to be processed. These approaches are useful to produce generic embeddings and to build models that are independent of a particular task. One of its strong limitations is that corpora annotated for multiple tasks are rare, small or difficult to build. An interesting way to solve this is to learn by mixing tasks. In this approach, each task is associated to a different corpus and, during learning, the model is only asked to predict an outcome for the task to which the input example is associated.

— **Lightweight training.** As previously mentioned, the performance of current approaches lies in their ability to statistically infer latent knowledge from massive quantities of annotated examples. Faced with the objective of being able to handle any task in any language or sub-language, it seems important to decrease the weight/cost of training models. It is first of all a question of the quantity of data required. This may mean learning from scratch with

very little data or quickly adapting a generic model from a few examples of a target domain. From another point of view, this problem is partly related to the problem of knowing how to model weak signals, i.e. signals masked by other signals that are more recurrent or more salient in context, in large quantities of data. Light machine learning also goes through to the gradual lifting of supervision. For example, in generative approaches, removing the need for aligned inputs and outputs makes it possible to handle new tasks. Instead of aligning sentences, comparable corpora are compared using adversarial architectures. In the longer term, this raises the question of totally unsupervised learning where the model identifies by itself the multiple dimensions of a content and the way in which these vary or not from one sample/corpus to another. Finally, it seems important to me, as already mentioned in the short-term perspectives, to re-emphasise the objective of reducing the energy consumption of the models (learning and use) given the climate crisis and the galloping democratisation of artificial intelligence tools.

— **Analysis and formalization of latent spaces for the design of complex models.** The straightforward evolution of research is to study increasingly complex tasks, requiring analysis at multiple levels. As an illustration, in neural speech synthesis, the work initially focused on single-speaker synthesis. Then it moved to the multi-speaker case, and even more recently to multi-accent or multilingual synthesis, with different styles, etc. Each time, these changes imply to add new modules for the encoding or recognition of speakers, accents, styles, etc. As a result, synthesis models are becoming increasingly complex, both in terms of the number of parameters to be estimated and in terms of design sophistication (choice of cost functions, hyper-parameters, optimization strategies, etc.). And this observation applies to many tasks in NLP. Thus, in order to master this increasing complexity, it seems important to me to make progress on the strategies for assembling elementary models. One of the ways to do this is to study the properties of embedding spaces. For instance, it is currently difficult to determine with precision the information contained in embeddings (given an input data, what information is kept or deleted ?). Similarly, their intrinsic dimension is a property that is not very well integrated in the model design process. Instead, hyper-parameters are usually adjusted following an empirical approach. Even more so, the properties that embeddings maintain or that one wishes to maintain between them are, in my opinion, still largely left aside (orthogonality, invariance, geometric transformations). Advances in these aspects would make it possible to better interface dives of various origins. For example, they would allow the design of models where the plunge space offers topological guarantees such that the addition of new information would not present any risk of altering performance or requiring the model to be re-trained. Overall, these limits call for more interactions with information theory and algebraic topology.

— **Structured data.** Finally, one of the current limitations of today's NLP is that data is limited to relatively small units (mostly utterances, sentence. . .). However, the need to work on larger scales is present in many tasks. For instance, in automatic summarization, the global analysis of a document makes it possible to know the general framework of a discourse and to produce an approximation that maximizes the conservation of information. In question-answering, the interconnection of several sentences makes it possible to answer more complex questions. In speech synthesis, the history of a paragraph or chapter can help in situating some elements of a story (characters, plot, etc.) or to propose a less repetitive prosody when the sentences follow on from one another. In other situations, it is at the level of the collection of documents that the analysis is carried out. Once again, automatic summarization is concerned, but one can also mention fact checking, information retrieval, or authorship attribution. In addition, there may be a need to know the internal organization of the documents or collections. This organization can represent a structuring in terms of thematic proximities between texts, social or hierarchical relationships between entities or the logic of argumentation between paragraphs. Current deep learning techniques are still relatively limited for these problems. First of all, compared to usual tasks, data are struc-

turally more complex to handle (long sequences, sequences of sequences, trees, graphs...). Second, given the increased amount of information, the needs for memorization and discovery of interdependencies (attentation mechanisms) also increase. This has a direct impact on the complexity of the models and the difficulty of training them. Finally, paradoxically, as each piece of structured data is relatively large, it is difficult to build corpora with a large amount of instances. Thus, tasks beyond the sentence often rely on systems where neural networks are combined with other techniques or are inserted in *ad hoc* algorithms for information aggregation. Hence, I think that there is room for more investigation in this area.

To conclude on the perspectives related to machine learning, it is important to note that they are often intertwined. For example, the embedding space properties of a model have an impact on its genericity, or the study of complex data structures requires a better understanding of diffuse recurrences. Thus, organizationally, these insights should be part of coordinated efforts.

3.2.2 The place of the social sciences and humanities

Another impact of the success of neural approaches is the tendency to be data agnostic. In this, the prospects for cross-fertilization with the social sciences and humanities seem to be diminishing. This section shares a few thoughts that, I hope, demonstrate the future of these interactions.

— **Unified communication and interaction model.** As stated in the introduction of this manuscript, utterances are the result of a communication schema in which multiple contextual factors combine (writer/speaker, modality/media, audience, etc.). The general trend in NLP is to study one of these factors and correlate it with linguistic observations from utterances (lexicon, syntax, prosody, etc.). The objective may be either to diagnose the presence of a factor among the main causes of the utterance or, conversely, to generate a credible utterance that simulates the factor under study. Nevertheless, to my knowledge, no generalized linking of all these influencing factors is proposed in the NLP litterature. Thus, the simulation of a complete communication scheme remains a challenge. To illustrate the problems to be solved, we can take the example of a smart answering machine that would answer phone calls or text messages received when the owner of the phone is not available. In order for the smart assistant to perform its task perfectly, it would have to infer various information, such as the link between the user and the caller, the subject in question, the positioning of each stakeholder in relation to this subject, etc. This is a difficult topic which, in my opinion, cannot be solved without a multidisciplinary approach. On the one hand, some of these contextual factors in a dialogue or in a history of discussions still requires more work to be automatically recognized. For instance, detecting innuendo or irony, estimating the level of knowledge on a topic or diagnosing fine opinions are useful tasks on which work is still needed. On the other hand, the relationships between the causal factors are not obvious. For example, if the interlocutor is a friend, a relaxed style seems adequate but it is probably no longer adequate if the topic is serious. The human sciences (psychology, sociology, linguistics) should shed light on these issues. In a broader framework, these questions also open up to the consideration of other modalities for the induction of these contextual elements (facial expressions, gestures, physiological information, etc.). This would enable the analysis or simulation of multimodal behaviors, such stress or eloquence.

— **Evaluation of complex systems.** As already stated in this chapter, the evolution of NLP, and more generally of artificial intelligence, is leading research towards more and more complex tasks. Beside machine learning questions, this trend strengthens the need for elaborate experimental methodologies. First, this is observed through the increasing constraints on the data collection process as more and more information is needed. That is, the experiments require annotations on multiple dimensions, possibly organized between them, in varied situations and, *a priori*, in great quantity. Secondly, the system validation conditions must ensure that there is no bias in the assessments, which is difficult as there are many

possible risks, e.g. experiments on a too restricted domain, important dimensions have been forgotten, unbalanced situations, unsuitable metrics are used, some testers have some prior knowledge, test duration is too short, etc. This is all the more important since perceptual evaluation campaign are mandatory for some systems, especially for content generation taks, be it in fake or real conditions. Finally, the question of the sustainability of the developed systems also seems to be insufficiently studied. For example, it is a question to know whether the performance of a smart assistant that adapts to its owner's uses will not deteriorate over time because an excessive gap will gradually appear between the system's initial conditions of development and those of use. This behavior is generally difficult to apprehend because systems are made up of multiple components whose side effects are not necessarily well known. Through work on the use of the natural language, its evolution and the human behavior, the human sciences offer possible solutions or guarantees to face these multiple problems. Hence, collaboration with researchers in these fields is likely to become a necessity in the future.

— **Saving languages.** Finally, although my activities relating to low resource languages are relatively undeveloped, I personally feel that NLP should be positioned as a tool for safeguarding the cultural heritage of humanity, particularly that of languages. Indeed, while thousands of languages exist today, only a small minority dominate and benefit from advances in research. This divide contributes to the risk of the disappearance of other languages because their learning or practice is not encouraged, especially among the new generations. This is a subject where NLP and the human sciences must be able to work together to build up and process linguistic resources (texts, speech, knowledge, etc.) as well as to propose tools (computational ones or not) for their processing (automatic or not). In particular, some initiatives aimed at documenting rare or unwritten languages by computational techniques seem to me to be promising. These works seek to automatically construct elements of formal linguistics (lexicons, grammars) in a principle of exchange with linguists on these languages.

3.2.3 Which place for which research ?

Finally, I end this manuscript by putting into perspective the organizational evolutions in NLP and artificial intelligence during the last few years. The objective is to question the future, without taking any particular position.

— **The ubiquity of the digital giants.** A first observation is that many of the recent major advances have been made by the giant digital companies (Google, Baidu, Facebook, Apple, Microsoft), thanks to their financial, material and human resources. In addition, these companies naturally have at their disposal very large quantities of data through their activities (queries, messages, chat, photos, videos, friendship and preference graphs). However, this data cannot generally be disseminated because of intellectual property rights. Faced with this situation, it is legitimate to ask whether institutional players should seek to compete with the digital giants. On the "no" side, it is indeed reasonable to think that some tasks will progress anyway, even without the efforts of public research, because these tasks are at the heart of the activities of industrial actors. Moreover, due to the training role of universities, it can be risky for the success of a PhD candidate to launch him/her on a subject on which these actors are working. Thus, the efforts saved could serve other topics, with less priority (especially commercially speaking) for these actors but which still remain difficult for the state of the art (e.g. low resource languages, model explicability). Still, in the "yes" camp, it seems illusory that a few research actors can be as creative as the entire community. Multiple mediatized advances in fact result from less spectacular initial proposals. Moreover, the role of public actors in transmitting knowledge means that knowledge must not be allowed to become the exclusive property of a few. Hence, simply keeping track of the state of the art can be seen as a duty towards society, even though it may sometimes seem futile for

those who do it.

— **The acceleration of research.** A second observation is that the pace of research has considerably accelerated. Many proposals are published prior to be accepted (or not) in conferences and journals (e.g. *via* arXiv), so that new extensions sometimes appear before the initial paper gets peer-reviewed. Then, it became common—if not strongly recommended—to release the code and data of any published work. While this is undoubtedly an healthy practice, the side effect is again the acceleration of research by making it easier for other people to rerun experiments and build up on them. Combined with the competition with the digital giants, some topics are thus very difficult to follow.

— **The need for teamwork.** Finally, it is paradoxal to note that the cost of an entry ticket to a domain has sometimes become high. Despite good practice and easy access to knowledge, entering a new topic requires an increasingly large body of practical knowledge and massive material resources, particularly in deep learning. Even more so, some major works are not directly reproducible because the code is not public, some hyper-parameters are not specified or the data have not been released. Thus, the conduct of work requires teamwork for the appropriation of know-how and the conduct of experimental aspects (collection, adjustments, evaluation).

Although this panorama may seem bleak, it simply aims to underline, as a conclusion to this manuscript, the importance of encouraging collegial reflections on these assessments, their consequences (both positive and negative) and possible actions, in particular to guarantee the plurality and vitality of tomorrow's research.

Partie II

Manuscrit original

CHAPITRE 1
Introduction

L'étude du langage naturel est présente dans de multiples disciplines scientifiques. En linguistique, cette étude est l'essence même des travaux et se décline en de nombreux pans spécifiques. En neuro-sciences également, le langage est au cœur de certains travaux pour comprendre comment cette capacité se développe ou s'altère chez l'humain. En tant qu'outil de communication et de lien, les domaines anthropologiques l'étudient pour analyser la société, les relations entre humains, l'histoire ou encore l'art. Enfin, le traitement automatique du langage naturel (TALN) est un domaine historique de l'informatique. Comme le démontre l'essor actuel de nombreux produits commerciaux dits « intelligents », il s'agit d'un champ relativement mature. Cependant, deux constats soulignent, à mon sens, certains manques du TALN. Premièrement, alors que celui-ci se porte depuis longtemps sur le sens des énoncés, le traitement du style reste un aspect encore assez mal maîtrisé. Pourtant, le style d'un énoncé est le reflet de multiples facteurs utiles à l'analyse d'un contenu. Deuxièmement, le dialogue entre le TALN et les autres disciples liées au langage naturel est relativement faible et tend même à se réduire si l'on s'en réfère aux pratiques de certaines techniques d'apprentissage profond.

Sur le premier point, la prépondérance des études sur le sens relève sans doute d'une priorisation logique pour faire progresser le domaine. À l'image d'autrui apprenant une langue, il est en effet plus important de comprendre ou de se faire comprendre que de maîtriser les différentes manières qui le permettent. Par ailleurs, le traitement automatique du style contribue davantage aux tâches de production de contenus langagiers qu'à celles d'analyse et de reconnaissance. Or, la popularité de ces tâches de production est relativement récente (synthèse de la parole, les *chatbots*, la production de paraphrases...).

De son côté, la faiblesse du dialogue interdisciplinaire s'explique probablement, outre la difficulté classique à s'entendre sur le vocabulaire, par la prédominance de l'approche statistique prédominante en TALN, issue des années 1990, selon laquelle la résolution des tâches s'appuie majoritairement sur l'estimation de phénomènes moyens à partir de volumes importants de données. Cette approche est en décalage avec les autres disciplines qui visent souvent, à l'inverse, à exhiber de multiples dimensions explicatives à partir d'un phénomène moyen initial. Par ailleurs, le passage au paradigme de l'apprentissage profond pour la quasi totalité des tâches en TALN accentue le manque de dialogue. En effet, pour un problème particulier à traiter, ce paradigme tend à rendre inutile la phase de conception des descripteurs pertinents et, par effet de bord, le recours à des experts linguistes, psychologues, sociologues, etc. Dorénavant, les modèles travaillent souvent directement sur des données brutes et intègrent des transformations préliminaires à leur appliquer en remplacement des descripteurs experts.

Ce manuscrit d'habilitation à diriger des recherches est une synthèse de mes activités de recherche depuis l'obtention de mon doctorat en 2010. Ces activités, qui s'inscrivent dans le cadre d'encadrements et de collaborations, reflètent mes évolutions de parcours à travers les domaines de la reconnaissance automatique de la parole (2007-2012), de la synthèse de la parole (depuis

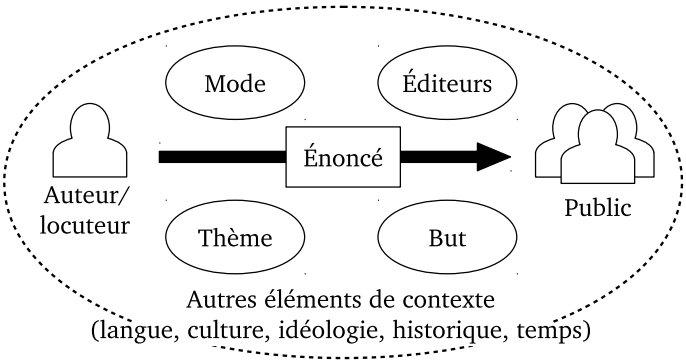


FIGURE 1.1 – Facteurs d'influence dans une communication verbale (adapté de (ARGAMON 2019)).

2012) et du traitement de l'écrit (depuis 2017). Sans prétendre infléchir les constats dressés à l'instant, mes travaux ont comme point commun de contribuer à la question du *traitement automatique du style* dans le langage naturel, qu'il soit oral ou écrit. Par ailleurs, bien que l'apprentissage automatique occupe une place primordiale dans mes travaux, l'intégration de considérations ou problématiques issues de la linguistique est un autre aspect qui, je crois, donne une cohérence à mon travail. Avant de détailler l'organisation de ce manuscrit en section 1.4, cette introduction pose le cadre général des travaux à suivre. Ainsi, la section 1.1 discute les mécanismes de construction d'un énoncé. La section 1.2 présente quelques exemples qui illustrent l'impact de ces mécanismes sur les observations linguistiques des énoncés produits. De là, nous analysons en section 1.3 des éléments de définition de la notion de style.

1.1 Construction d'un énoncé

Dans sa vision stratifiée de la langue, la linguistique définit de multiples niveaux d'abstraction (phonologie, lexique, morphologie, morphosyntaxe, syntaxe...). Ces niveaux offrent des outils de description et d'analyse depuis la forme physique d'un énoncé, c'est-à-dire son signal ou ses caractères jusqu'à son sens (niveau sémantique) et la manière dont celui-ci s'inscrit dans un contexte plus général (niveau pragmatique). Lorsque l'on compare des énoncés, ces niveaux permettent de caractériser des différences. Par exemple, l'emploi de structures syntaxiques plus ou moins simples, le recours à un vocabulaire plus ou moins riche ou encore la fluidité du discours sont des caractéristiques observables.

Néanmoins, ces niveaux d'abstraction n'expliquent pas les causes de ces observations, si ce n'est à travers l'existence du niveau pragmatique. Ces causes apportent pourtant une cohérence qui, dans une perspective de traitement automatique, éclaire souvent la résolution d'un problème. Comme le résume le schéma de communication de la figure 1.1, la forme finale d'un énoncé est conditionnée par de multiples facteurs. Tout d'abord, elle dépend de l'*énonciateur* (son état d'esprit, opinion, âge, origine sociale, niveau d'expertise...), du *public visé* (*idem*) et de leur relation (lien social, lien affectif, connaissances partagées, état d'avancement d'une discussion, unité de temps et de lieu...). Le *thème* aussi est prépondérant. Il influence évidemment le niveau sémantique mais peut en outre appeler des réactions chez l'énonciateur ou son public, et donc impacter le style de la communication. Le *but communicationnel* (par ex., informer, dicter, divertir...), le *mode* (parole, roman, blog, email, émission TV...) et d'éventuels mécanismes d'*édition* imposent généralement des codes quant à la forme du message. Par exemple, le discours d'un journal télévisé aura tendance à avoir un ton neutre et un niveau de langue courant, caractérisé par une syntaxe et un lexique non fautifs et généralement abordables par tous. Enfin, beaucoup d'autres facteurs plus généraux influencent l'énoncé : la langue ; des facteurs sociétaux comme la culture, les idéologies, etc. ; ou l'emprise temporelle ou spatiale d'un discours.

	Énoncé A	Énoncé B	Variations observables	Causes potentielles
(1)	Les fleurs éclosent au printemps.	Le parlement a voté le texte.	Lexique, syntaxe, sémantique, pragmatique	Thèmes différents
(2)	J'adore cette histoire.	Je déteste ce scénario.	Lexique, syntaxe, prosodie, sémantique	Sentiment, Émotion
(3)	Avant de s'en aller, il me salua.	Il m'a salué et il est parti.	Syntaxe, Lexique	Public visé, maîtrise de la langue par l'auteur
(4)	Elle bosse dur.	Elle œuvre avec abnégation.	Syntaxe, Lexique	Origine sociale, Niveau d'éducation, public visé
(5)	J'ai une rhinopharyngite.	J'ai un rhume.	Lexique	Public visé, expertise de l'auteur
(6)	Elle est assidue.	Elle est zélée.	Lexique, sémantique	Sentiment
(7)	Je le veux.	C'est ça que je veux.	Syntaxe, prosodie	Degré de volonté, état du dialogue
(8)	« Comme euh... Comme tu veux. »	« Comme tu veux. »	Syntaxe, lexique, prosodie	Émotion, intention
(9)	« Comment ? »	« Co- co- comment ? »	Morphologie, phonologie, prosodie	Émotion, pathologie
(10)	« Je ne le savais pas. »	« J' le savais pas. »	Phonologie, phonétique, prosodie	Lien locuteur/public, niveau d'éducation, émotion
(11)	« Bonjour maman »	« Bon.jour ma.man » (*surarticulation*)	Prosodie	But (dictée ?), Environnement sonore, pathologie
(12)	« Ça va ? »	« Ça va. »	Prosodie, sémantique	État du dialogue
(13)	<*Signal de parole grave*>	<*Signal de parole aigu*>	Acoustique	Âge, sexe du locuteur, émotion
(14)	tkt c bi1! ;-)	Ne t'inquiète pas, c'est bien ! Haha !	Lexique, morphologie, syntaxe	Média

TABLE 1.1 – Exemples de variations entre paires d'énoncés.

1.2 Exemples de variations

Les liens de causalité entre facteurs contextuels et observations linguistiques sont divers. Pour situer les travaux développés dans ce manuscrit, la table 1.1 liste des paires d'énoncés A et B qui varient à différents niveaux linguistiques. Pour chaque variation, des causes contextuelles possibles sont données. Tout d'abord, les exemples (1) et (2) montrent des situations où le sens diffère. Les autres niveaux linguistiques varient également car ceux-ci sont en partie l'instrument du niveau sémantique. Dans les autres exemples, le sens varie peu. Les paires d'énoncés 3 à 7 sont des re-

formulations syntaxiques et lexicales. Certaines varient en terme de complexité (3-5), peut-être en raison d'un énonciateur ou d'un public avec une maîtrise plus ou moins élevée de la langue. D'autres (6-7) reflètent possiblement l'état d'esprit du locuteur (un sentiment, une envie...). Les exemples 8 à 12 rassemblent des variations spécifiques à la parole, par exemple des disfluences au niveau lexical (8) ou infra-lexical (9), des variantes de prononciation (10), d'articulation (11) ou d'intonation (12). Là encore, les causes peuvent être variées, bien qu'aisément imaginables. Enfin, d'autres variations sont spécifiques le mode de communication (signal de parole, texte manuscrit, tapé sur un téléphone). Par exemple, la fréquence fondamentale d'un signal de parole peut traduire des informations sur le locuteur (13). Sous une autre forme encore, l'exemple 14 montent que les codes graphiques de la langue peuvent être modifiés (SMS, Twitter, forums...).

Dans ce manuscrit, le périmètre d'étude va des exemples (3) à (10), c'est-à-dire que nous ne nous intéressons pas aux variations sémantiques ni à celles liées au média. Par ailleurs, l'objectif n'est pas non plus de construire un modèle de causalité général. Nous nous contentons, plus modestement, d'étudier des cas particuliers, déjà mis en lumière par la littérature linguistique, et d'en proposer des modélisations informatiques. L'ambition d'un modèle unifié de ces différents cas relève des perspectives que nous aborderons en fin de manuscrit.

1.3 Qu'est-ce que le style ?

Bien que l'enjeu de ce document ne soit pas de proposer une définition du style, ni d'en défendre une approche particulière, il est intéressant d'observer les interprétations qui en sont faites dans la littérature en sciences humaines et en TALN afin de positionner les différents travaux que nous exposerons par la suite.

Dans les sciences humaines, la notion de style n'est pas uniformément établie. Il est ainsi sans doute préférable de définir le style comme l'objet de l'analyse stylistique. Sous un angle littéraire, cette analyse porte sur les efforts et stratégies artistiques d'un auteur (SEBEOK 1960). En psychologique, elle étudie la perception d'un message (RIFFATERRE 1961). En particulier, la rhétorique s'intéresse à l'efficacité de ce message. Le style est alors souvent lié à la modalité orale et à une logique de prise de position (JOHNSTONE 2009). Dans les domaines sociologiques, il peut être associé à l'usage cohérent de la langue au sein d'une communauté linguistique, concept que nous qualifierions plutôt de registre dans ce manuscrit (*cf.* chapitre 4), comme le notent d'ailleurs ECKERT et RICKFORD (2001). Ainsi, citant (CRESSOT et GALLO 1969), « *le fait stylistique est donc d'ordre à la fois linguistique, psychologique et social : il faut que nous soyons compris.* » En cela, une vision linguistique consensuelle consiste à aborder le style comme l'ensemble des choix faits par un énonciateur d'utiliser certains outils offerts par la langue (1) afin de faciliter la compréhension de son énoncé et, donc, (2) en fonction du contexte d'énonciation.

Dans la perspective d'un traitement automatique, le style peut plus pragmatiquement se définir à travers son association au sens, le premier modulant le second par des informations qui relèvent du contexte d'énonciation. Parmi les questions autour de cette association, l'une majeure est alors de savoir si style et sens peuvent être considérés comme indépendants ou intriqués. La conséquence pratique d'une hypothèse d'indépendance est qu'un style donné peut être décomposé en un ensemble d'éléments linguistiques explicites comme des mots spécifiques, des marqueurs ou des structures syntaxiques (LEEFTINK et SPANAKIS 2019 ; LI et al. 2018). Dans l'autre, le style est considéré comme un concept holistique, c'est-à-dire une composante implicite et intégrale d'un texte. Pour pouvoir distinguer style et sens, les modèles doivent alors passer de l'espace linguistique, où les deux concepts ne peuvent être séparés, à des espaces latents (TIKHONOV et YAMSHCHIKOV 2018). Ce concept présente l'intérêt de sa généralisation aisée à de nombreux cas pratiques. En effet, étant donné deux corpus supposés de styles différents, le style peut y être vu comme la dimension qui discrimine de manière invariante les textes de ces deux corpus.

1.4 Organisation du manuscrit

Le présent document reflète mon exploration en largeur de la question du style dans le langage et en TALN au cours de ces dix dernières années. Il s'organise en courts chapitres dont l'exposé prend la forme de résumés successifs regroupés autour de questionnements communs. Précisément, les chapitres 2 à 5 présentent des travaux sur des variations stylistiques bien identifiées :

— Le chapitre 2 traite de la modélisation et de l'adaptation de séquences de phonèmes pour intégrer de la variabilité et s'adapter à des styles de prononciations différents.
— Le chapitre 3 décrit des travaux réalisés sur l'insertion de disfluences dans un texte afin de lui apporter une dimension plus naturelle, plus humaine.
— Le chapitre 4 aborde la question des registres (ou niveaux) de langue familier, courant et soutenu.
— Le chapitre 5 donne un aperçu de travaux sur la modélisation du langage à destination des enfants.

Les chapitres 6 et 7 abordent, quant à eux, des sujets plus transversaux :

— Le chapitre 6 aborde des travaux liés à l'encodage de l'information linguistique par des réseaux de neurones.
— Le chapitre 7 présente mes activités de développement logiciel pour la recherche et de transfert technologique.

Enfin, le chapitre 8 rassemble les perspectives de recherche issues de ces différents travaux à un horizon d'une nouvelle dizaine d'années.

CHAPITRE 2

Variantes de prononciation

La prononciation associée aux mots d'une langue est l'élément-clé qui lie le monde de l'écrit et celui de l'oral. En TALN, la phonétisation est la tâche qui a pour objectif de prédire la prononciation de mots ou d'énoncés en associant une séquence de phonèmes à une séquence de graphèmes. Cette étape peut être difficile car la prononciation est sujette à beaucoup de variations, par exemple en fonction des habitudes ou d'une pathologie du locuteur, du degré de familiarité avec le public ou encore d'un accent spécifique (régional ou étranger). Ces variations se traduisent par des insertions, suppressions et susbstitutions de phonèmes par rapport à la prononciation standard, *canonique*.

La plupart des outils de phonétisation reposent principalement sur des lexiques de prononciation construits manuellement pour les mots communs de la langue et sur une conversion graphèmes-phonèmes automatique pour les autres. Pour la plupart des langues, la phonétisation d'un énoncé se limite alors à la concaténation des prononciations individuelles de ses mots. Cette approche n'est cependant pas viable pour certaines langues (par exemple, le français) où les transitions entre mots provoquent des modifications phonétiques.

La production de variantes de prononciation peut servir plusieurs objectifs. Il peut s'agir de modéliser de manière compacte et exhaustive toutes les variantes possibles dans une langue donnée ou de préserver un maximum d'information en vue de post-traitements. Alternativement, l'objectif peut être d'introduire de nouvelles prononciations à partir de celles canoniques pour imiter ou s'adapter à un contexte particulier. La tâche s'apparente alors à une conversion phonèmes-phonèmes.

Ce chapitre présente différentes contributions à ces problèmes dans le cadre de la synthèse de la parole. Leur finalité globale est de pouvoir contrôler l'étape de phonétisation afin d'imiter certains traits d'expressivité. Dans la section 2.1, nous abordons la question de la phonétisation d'énoncés en français et présentons une méthode de production de variantes sous la forme de treillis de phonèmes. Dans la section 2.2, nous traitons de la conversion phonèmes-phonèmes et l'illustrons sur plusieurs cas.

2.1 Conversion graphèmes-phonèmes

Collaborations : Damien Lolive (IRISA)
Références :

— LECORVÉ, G. & LOLIVE, D. (2015). Adaptive statistical utterance phonetization for French. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE
— LECORVÉ, G. & LOLIVE, D. (2016). Phonétisation statistique adaptable d'énoncés pour le français. *Actes des Journées d'Études sur la Parole (JEP)*

Dans (**LECORVÉ** et LOLIVE 2015, 2016), nous avons présenté une nouvelle méthode de phonétisation du français. Cette méthode apporte trois contributions : (i) elle introduit la notion de modèle d'élision pour modéliser les variantes intra-mots ; (ii) elle intègre les contextes phonologiques pour modéliser les variantes inter-mots ; (iii) elle permet de générer des treillis probabilistes de phonèmes à partir d'énoncés. Pour cela, cette méthode repose sur des champs aléatoires conditionnels (*Conditional Random Fields*, CRF) pour estimer les probabilités des phonèmes sur les mots isolés [1] et sur des transducteurs finis pondérés pour traiter les transitions entre mots.

Cette section présente tout d'abord le cas des mots isolés, puis celui des énoncés.

2.1.1 Phonétisation de mots isolés

Dans ce travail, la prédiction des prononciations possibles pour un mot isolé est effectuée par deux CRF consécutifs, l'un pour la séquence de graphèmes du mot en une séquence de phonèmes, l'autre pour prédire d'éventuelles élisions sur ces phonèmes.

Le modèle de conversion graphèmes-phonèmes est appris sur un corpus aligné de graphèmes et de phonèmes issus d'un lexique de prononciations. Ces alignements sont effectués entre blocs de graphèmes et phonèmes de taille maximale fixée (JIAMPOJAMARN et al. 2007). Le CRF prédit zéro, un ou plusieurs phonèmes pour chaque graphème en fonction de ces alignements. Cette prédiction s'appuie sur un voisinage de $\pm N$ graphèmes autour du graphème en cours d'examen. Par ailleurs, comme le français contient de nombreux homographes aux prononciations différentes [2], la classe grammaticale du mot est également prise en compte. Comme proposé par (ILLINA et al. 2011), cette information peut se résumer à la simple distinction verbe/non verbe pour le français. Sur le fond, d'autres caractéristiques comme l'étymologie du mot ou un accent à considérer pourraient être utilisées mais ce n'est pas l'objectif du présent travail. Formellement, le CRF de conversion graphèmes-phonèmes prédit donc chaque phonème p à partir d'un n-gramme de graphèmes $\mathbf{g}$ et de caractéristiques $\mathbf{o}$ dérivées du mot à phonétiser. Ce CRF est capable de produire la ou les meilleures hypothèses de phonétisation du mot et la probabilité *a posteriori* de chaque phonème.

Ensuite, certains phonèmes peuvent être élidés. Ces élisions dépendent d'informations variées, comme le contexte phonologique, le type de parole, les règles ou exceptions liées à la grammaires, etc. En français, le phénomène le plus courant pour illustrer cette variabilité est le cas du schwa (/ə/). Par exemple, le mot *semaine* peut être prononcé /səmɛnə/, /səmɛn/, /smɛnə/ ou /smɛn/. Notons, au passage, que toutes ces prononciations ne portent pas la même expressivité. Des phénomènes similaires existent pour d'autres phonèmes, en particulier les liaisons lorsque l'on considère les liens entre mots consécutifs. Ainsi, nous avons proposé d'entraîner un modèle d'élision. Pour chaque phonème dans la prononciation d'un mot, ce modèle prédit une étiquette *optionnel*, c'est-à-dire que le phonème peut être prononcé ou non, ou *obligatoire*. Ce modèle est appris grâce aux informations fournies par le lexique de prononciations.

Une fois combinées les prédictions des deux modèles CRF, la probabilité d'un phonème peut être reformulée par la probabilité $\Pr(p|\mathbf{g},\mathbf{o})$ que le phonème p se réalise et celle $\Pr(\epsilon|p,\mathbf{g},\mathbf{o})$ qu'il ne se réalise pas, ϵ signifiant l'absence de phonème. En utilisant ces probabilités, un treillis de phonèmes peut être créé pour une séquence quelconque de graphèmes donnée en entrée. L'architecture d'un tel treillis est illustré par la figure 2.1.

2.1.2 Phonétisation d'énoncés

Comme l'illustre la figure 2.2, la prononciation d'un énoncé s'appuie sur la notion de contexte phonologique, c'est-à-dire qu'un mot w_i influence la prononciation des mots précédent et suivant w_{i-1} et w_{i+1}. Réciproquement, la prononciation du mot w_i dépend de celle des mots w_{i-1}

1. Le choix des CRF est lié à la popularité de cette approche pour la conversion graphèmes-phonèmes avant les succès récents des réseaux de neurones (ILLINA et al. 2011 ; D. WANG et KING 2011).
2. Par exemple, la graphie *président* se prononce /pʁezidɑ̃/ s'il s'agit du nom ou /pʁezid/ s'il s'agit du verbe *présider*.

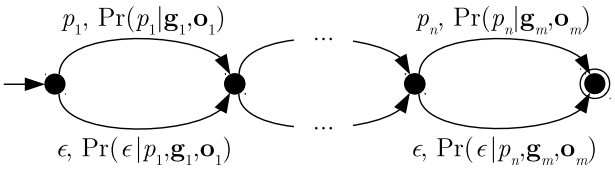


FIGURE 2.1 – Treillis de phonèmes pour un mot isolé.

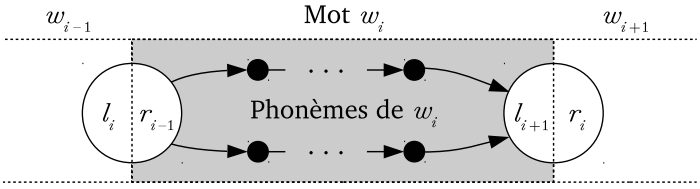


FIGURE 2.2 – Principe de contextualisation d'un mot w_i.

et w_{i+1}. Ainsi, nous pouvons considérer l_i, l'information transmise par w_i vers la gauche, c'est-à-dire vers w_{i-1}, et r_i l'information transmise vers la droite, vers w_{i+1}. De manière symétrique, la prononciation de w_i dépend de r_{i-1} et l_{i+1}. Nous avons proposé, d'une part, d'intégrer r_{i-1} et l_{i+1} comme nouveaux descripteurs dans le processus d'apprentissage des CRF de conversion graphèmes-phonèmes et d'élision et, d'autre part, de mettre en lien toutes les prononciations contextualisés d'un énoncé au sein d'un transducteur. La constructeur d'un treillis de phonèmes pour l'énoncé se fait alors composant le transducteur de prononciations contextualisés avec un autre représentant l'énchainement des mots de l'énoncé.

Le transducteur des prononciations contextualisées est présenté par la figure 2.3. Pour chaque mot w_i et ses paramètres $\mathbf{o}_i$, plusieurs phonétisations peuvent être acceptables selon le contexte phonologique d'usage du mot. Ces contextes phonologiques sont représentés comme des nœuds (a, b) à partir desquels et vers lesquels chaque phonétisation est reliée. Entre ces nœuds, le tranducteur consomme $(w_i, \mathbf{o}_i)$ et génère les phonèmes $p_{i,j}$ ou des ϵ-transitions pour les élisions. Finalement, les transitions entre mots sont traitées de la manière suivante : chaque prononciation contextualisée $(r_{i-1}, [p_{i,1}, \ldots, p_{i,n}], l_{i+1})$ est liée à tous les nœuds possibles de contexte $(*, r_{i-1})$ et $(l_{i+1}, *)$ reflétant les informations transmises par la prononciation $[p_{i,1}, \ldots, p_{i,n}]$. Toutes les prononciations sont également liées à un nœud de repli pour autoriser des transitions théoriquement interdites [3]. Enfin, en fonction de leur contexte phonologique, certains nœuds de contexte sont définis comme terminaux.

2.1.3 Mise en application

La méthode de conversion proposée a été entrainée sur le corpus MHATLex (PÉRENNOU et DE CALMES 2000) qui comprend $450\,000$ mots avec un total de $710\,000$ prononciations. Chaque prononciation inclut des possibilités d'élision ainsi que les contextes phonologiques pour lesquels chaque prononciation s'applique. Pour mesurer la qualité de l'approche, $1\,400$ énoncés ont été tirés d'un corpus de parole manuellement phonétisés. Les expériences ont montré un taux d'erreurs comparable, quoique légèrement moins bon, à des méthodes de l'état de l'art mais qui ne produisent pas de variantes. Une importante proportion d'erreurs vient d'élisions mal prédites. Ceci s'explique par le fait que les treillis de phonèmes recensent de nombreux chemins équiprobables, notamment engendrés par des possibilités d'élision ou de liaison, car le modèle ne favorise aucune stratégie phonétique particulière. Lors de l'intégration de la méthode dans un système de synthèse de la parole du français, ce manque de stratégie cohérente a été compensé par une étape

3. Par exemple, si l'on souhaite imiter un usage fautif de la langue

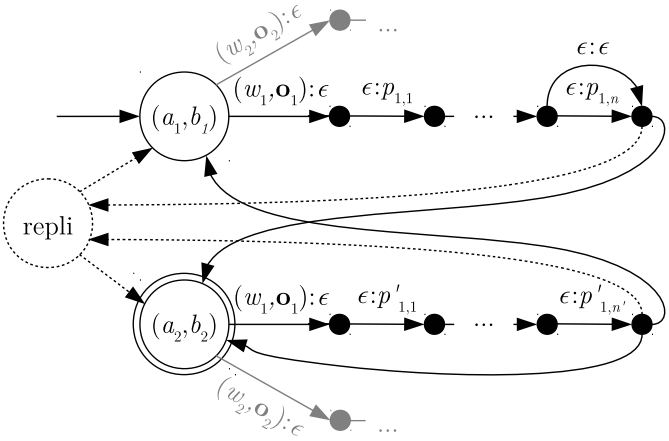


FIGURE 2.3 – Transducteur des prononciations contextualisées. Les arcs sans étiquette sont des transitions $\epsilon:\epsilon$. Les probabilités ne sont pas affichées par souci de clarté.

de réévaluation *a posteriori* des treillis de phonèmes. Cette réévaluation se fait par un modèle *n*-gramme de phonèmes appris sur des données de parole dont on souhaite adopter le style. Il ressort de ce post-traitement que le taux d'erreurs est largement réduit.

2.2 Adaptation de séquences phonémiques

Projets : SynPaFlex (responsable de tâche – ANR – 2015-2019 – IRISA, LLF, ATILF)
Encadrements⋆/collaborations : ⋆Raheel Qader (doctorat), ⋆Marie Tahon (postdoctorat), Damien Lolive (IRISA), Pascale Sébillot (IRISA)
Références :

— QADER, R., **LECORVÉ**, G., LOLIVE, D. & SÉBILLOT, P. (2014). *Phonology Modelling for Expressive Speech Synthesis : a Review* (Rapport de recherche N° PI-2020). Rapport de recherche. IRISA
— QADER, R., **LECORVÉ**, G., LOLIVE, D. & SÉBILLOT, P. (2015). Probabilistic Speaker Pronunciation Adaptation for Spontaneous Speech Synthesis Using Linguistic Features. *Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP)*
— QADER, R., **LECORVÉ**, G., LOLIVE, D. & SÉBILLOT, P. (2016). Adaptation de la prononciation pour la synthèse de la parole spontanée en utilisant des informations linguistiques. *Actes des Journées d'Études sur la Parole (JEP)*
— TAHON, M., QADER, R., **LECORVÉ**, G. & LOLIVE, D. (2016a). Improving TTS with corpus-specific pronunciation adaptation. *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*
— TAHON, M., QADER, R., **LECORVÉ**, G. & LOLIVE, D. (2016b). Optimal feature set and minimal training size for pronunciation adaptation in TTS. *Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP)*. Springer
— QADER, R. (2017). *Pronunciation and disfluency modeling for expressive speech synthesis* (thèse de doct., University of Rennes 1)
— LOLIVE, D., ALAIN, P., BARBOT, N., CHEVELU, J., **LECORVÉ**, G., SIMON, C. & TAHON, M. (2017). The IRISA Text-To-Speech System for the Blizzard Challenge 2017. *Proceedings of the Blizzard Challenge Workshop*

— QADER, R., **LECORVÉ**, G., LOLIVE, D., TAHON, M. & SÉBILLOT, P. (2017). Statistical pronunciation adaptation for spontaneous speech synthesis. *Proceedings of the International Conference on Text, Speech, and Dialogue (TSD)*. Springer
— TAHON, M., **LECORVÉ**, G., LOLIVE, D. & QADER, R. (2017). Perception of expressivity in TTS : linguistics, phonetics or prosody ? *Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP)* (T. 10583). Springer
— TAHON, M., **LECORVÉ**, G. & LOLIVE, D. (2018). Can we Generate Emotional Pronunciations for Expressive Speech Synthesis ? *IEEE Transactions on Affective Computing*

Plutôt que de produire des treillis de phonèmes à partir d'une séquence de graphèmes, il est également possible d'imiter un style en adaptant une séquence phonémique issue d'un phonétiseur standard. Cette adaptation consiste en une conversion phonèmes-phonèmes où chaque phonème d'une prononciation canonique peut être supprimé, remplacé par un autre, gardé tel quel ou complété par des phonèmes à insérer. Dans le contexte de la synthèse de la parole, nous avons travaillé à la mise en place d'une telle méthode d'adaptation. Comme dans la section précédente, la méthode repose sur des CRF. Après avoir donné le principe général de l'adaptation, cette section décline son utilisation sur trois problèmes que nous avons étudiés : le style de la parole spontanée (QADER 2017 ; QADER, **LECORVÉ**, LOLIVE et SÉBILLOT 2014, 2015, 2016 ; QADER, **LECORVÉ**, LOLIVE, TAHON et al. 2017), celui d'une voix enregistrée pour la synthèse (TAHON, QADER et al. 2016a,b) et les prononciations dans un contexte émotionnel (TAHON, **LECORVÉ** et LOLIVE 2018 ; TAHON, **LECORVÉ**, LOLIVE et QADER 2017)

2.2.1 Principe général

Nous formalisons l'adaptation de la prononciation comme la prédiction d'une séquence de phonèmes réalisés à partir d'une séquence de phonèmes canoniques. La qualité d'une adaptation se mesure alors à son taux d'erreurs entre les phonèmes prédits, dits *adaptés*, et ceux effectivement réalisés dans le corpus dont le style a pour but d'être reproduit. La figure 2.4 en présente le protocole expérimental. Nous considérons un ensemble d'énoncés dont on dispose du signal, de la transcription orthographique et de deux transcriptions phonétiques, l'une correspondant aux phonèmes canoniques qui auraient dû être prononcés si le style avait été neutre, l'autre aux phonèmes effectivement réalisés par le locuteur. Ces informations peuvent être enrichies par des descripteurs linguistiques ou articulatoires. De plus, comme il a été démontré que la prosodie influence la prononciation (K. CHEN et HASEGAWA-JOHNSON 2004), il est intéressant de considérer des descripteurs acoustico-prosodiques. Dans un système idéal, ceux-ci pourraient être prédits depuis le texte. Néanmoins, cette tâche étant toujours un sujet de recherche, nous considérons des descripteurs extraits de manière oracle depuis le signal réalisé.

Dans ce cadre, l'essentiel de nos travaux a consisté en l'étude de l'incidence de différentes configurations d'apprentissage du modèle d'adaptation sur le taux d'erreurs entre les phonèmes adaptés et réalisés, ainsi que sur la perception auditive finale par des testeurs. Les éléments-clés de ces configurations sont illustrés par la figure 2.5 et détaillés ci-dessous.

1. Principalement, chaque phonème p_i à prédire dépend de n descripteurs $\{d_i^1, \ldots, d_i^n\}$, par exemple le phonème canonique à adapter, sa position dans la syllabe ou la fréquence du mot qui le contient. La question est alors de savoir quels descripteurs parmi tous ceux considérés sont pertinents et quels autres dégradent l'adaptation.
2. Ensuite, le panel des informations peut être étendu au voisinage de p_i, par exemple en considérant également le phonème canonique précédent et le suivant, ainsi que les descripteurs qui leur sont associés. Cette notion de voisinage se définit en pratique par une fenêtre de taille W autour de p_i. Le choix de cette taille est un point que nous avons étudié.
3. De manière analogue, une dépendance entre p_i et les prédictions p_{i-1} et p_{i+1} peut être considérée. Cela doit notamment permettre d'éviter des enchaînements de phonèmes possiblement non articulables. Nous avons cherché à savoir si cette dépendance est utile.

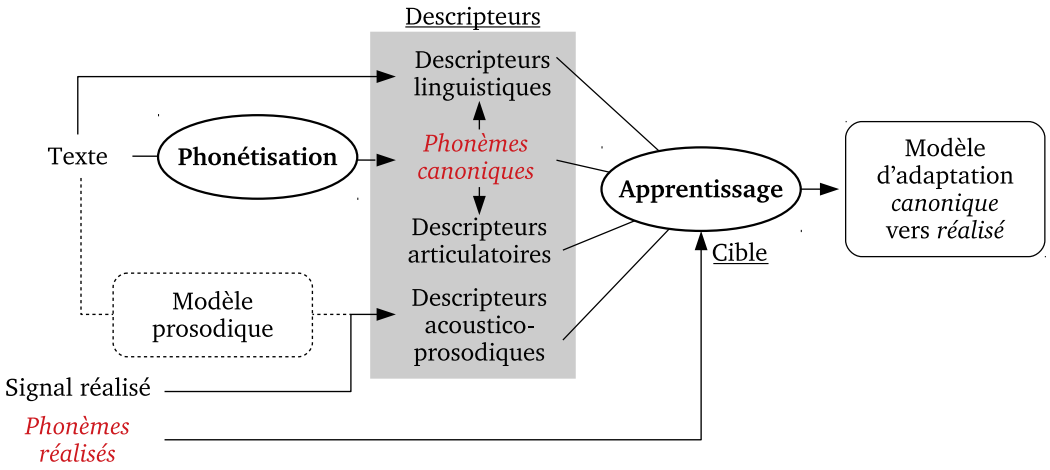


FIGURE 2.4 – Protocole d'apprentissage d'un modèle d'adaptation.

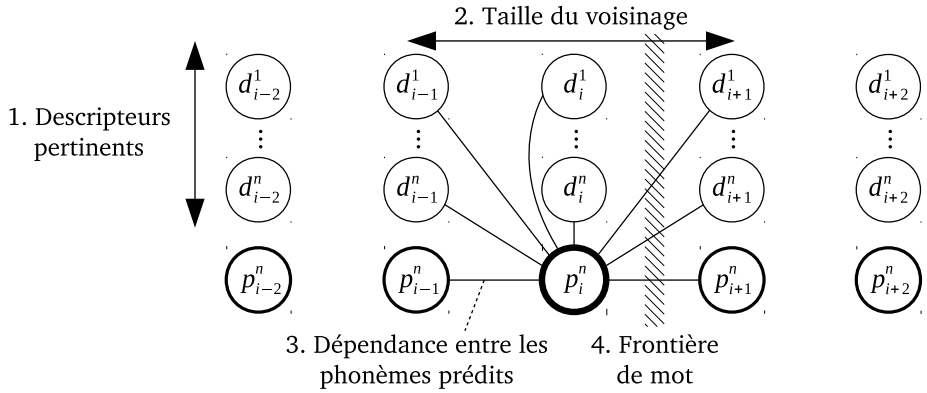


FIGURE 2.5 – Paramètres et liens de dépendance pour l'apprentissage du modèle d'adaptation.

4. Enfin, il est possible d'interdire ou autoriser la propagation des dépendances par-delà les frontières de mots. Nous avons également étudié ces deux options.

Les prochaines sections donnent les principales conclusions à ces questions sur les trois situations de mise en application précédemment listées.

2.2.2 Adaptation à un style de parole spontanée

La parole spontanée est un style de parole très expressif dans lequel les orateurs n'ont pas préparé leur discours auparavant et où la conversation évolue naturellement. Sur le plan phonétique, les variations de la prononciation y sont très présentes, comme en donne un aperçu la table 2.1. Il ressort de la littérature que celles-ci sont corrélées à des phénomènes observables à d'autres niveaux linguistiques. Par exemple, les durées d'énonciation plus courtes ou plus longues que la moyenne sont des indicateurs de variations phonétiques particulières (FOSLER-LUSSIER et MORGAN 1999 ; WIGHTMAN et OSTENDORF 1994), tout comme l'accentuation – notamment lorsqu'elle est lexicale, comme en anglais – et l'énergie (BATES et OSTENDORF 2002 ; K. CHEN et HASEGAWA-JOHNSON 2004). La richesse du langage, la prévisibilité et l'importance des mots affectent également la prononciation dans la parole spontanée. Les locuteurs ajustent leurs articulateurs pour tenir compte de l'importance des informations dans leur discours (BATES et OSTENDORF 2002) ou de la répétition de certains termes (FOWLER et HOUSUM 1987). Ce phénomène est alors compensé par le

Nature	Texte	Prononciation canonique	Variante
Assimilation	*can be*	/kæ**n** biː/	/kæ**m** biː/
Élision	*last month*	/læs**t** mənθ/	/læs mənθ/
Épenthèse	*vanilla ice cream*	/vənɪlə aɪskriːm/	/vənɪlə **r**aɪskriːm/
Réduction	*and*	/**æ** nd/	/ənd/
Haplologie	*library*	/laɪ.brə.ri/	/laɪ.bri/
Combinaison	*bread and butter*	/brɛ**d ænd** bʌtə/	/brɛ**b m** bʌtə/

TABLE 2.1 – Exemples de variations rencontrés dans de la parole spontanée.

			Mono-locuteur	Multi-locuteur
Pas d'adaptation			28,3 %	
Adaptation avec	Phonèmes canoniques	Sans voisinage	30,0 %	–
		Avec voisinage	24,2 %	24,8 %
	+ Descripteurs acoustico-prosodiques		21,5 %	20,6 %
	+ Descripteurs linguistiques		21,1 %	20,3 %
+ Ré-ordonnancement			20,6 %	20,1 %

TABLE 2.2 – Taux d'erreurs sur les phonèmes de l'adaptation pour la parole spontanée.

fait que les auditeurs perçoivent mieux les mots dans des contextes prévisibles (LIEBERMAN 1963).

Pour produire des prononciations adaptées au style spontané, nous avons travaillé avec le corpus de parole conversationnelle Buckeye (PITT et al. 2005). Ce corpus, en anglais, consiste en de longs entretiens non préparés avec des locuteurs de l'Ohio, aux États-Unis. Après alignement des phonèmes canoniques et réalisés, il s'avère que $30\,\%$ des phonèmes et $57\,\%$ des mots sont prononcés différemment de ce qui était attendu.

Dans ce travail, nous associons à chaque phonème canonique 42 descripteurs aux niveaux linguistiques (fréquences des mots dans le discours et dans la langue, position dans l'énoncé, catégorie grammaticale, racine, position du phonème, structure de la syllabe...), articulatoires (voisement, place/manière de l'articulation...) et acoustico-prosodiques (énergie/F0 de la syllabe, taux d'élocution, distance à la prochaine pause...). Les expériences, dont les principaux résultats sont rapportés par la table 2.2, montrent globalement que l'utilisation d'un voisinage autour de chaque phonème canonique à adapter est nécessaire (fenêtre de quelques phonèmes) mais la transmissions d'informations entre mots consécutifs ne semble pas nécessaire. En termes de descripteurs, les descripteurs acoustiques ont la plus grande influence sur l'adaptation de la prononciation mais les descripteurs linguistiques apportent un gain complémentaire. Au contraire, les descripteurs articulatoires n'apportent pas d'information additionnelle par rapport aux seuls phonèmes canoniques. Bien qu'il soit raisonnable de penser que chaque locuteur a ses propres habitudes phonologiques, il ressort également des expériences qu'un modèle d'adaptation multi-locuteurs est plus performant que des modèles propes à chacun. Tout comme pour la section 2.1, nous avons montré qu'un post-traitement pénalisant les hypothèses contenant certains enchainements de phonèmes peu probables dans la langue conduit à des améliorations. Enfin, il ressort que les taux d'erreurs restent élevés même après adaptation. Nous expliquons ceci par le fait qu'il est sans doute illusoire d'espérer atteindre un taux d'erreurs nul car les prononciations réalisées par les locuteurs ne sont pas systématiques. L'étude des N meilleures hypothèses montre que le modèle d'adaptation prévoit les bonnes possibilités de variation mais qu'il est incapable de savoir précisément laquelle sera réalisée, sauf à prédire la plus fréquente d'entre elles. Pour faire écho aux discussion du chapitre introductif, quand bien même certains éléments contextuels permettent d'expliquer cette non-systématicité apparente, ces éléments sont hors de portée du modèle que nous apprenons faute d'outils d'analyse contextuelle.

La spontanéité et l'intelligibilité a été évaluée perceptuellement par un test de préférence auprès

de locuteurs anglais natifs à qui des signaux de parole synthétisés avec diverses variantes de prononciations ont été présentés. Il en ressort que les prononciations adaptées et réalisées sont effectivement jugées plus spontanées que les prononciations non adaptées. Tout particulièrement, il apparaît même que la prise en compte des informations linguistiques pour l'adaptation aboutit à des prononciations qui, une fois synthétisées, sont jugées plus spontanées que les prononciations réalisées. En terme d'intelligibilité, les prononciations non adaptées sont meilleures mais les prononciations adaptées sont, là encore, bien meilleures que les prononciations réalisées. Ces résultats sur l'intelligibilité s'expliquent par le fait que le système de synthèse utilisé ici est construit sur une voix de style lu, avec une prononciation appliquée. Cela souligne le fait que l'adaptation des prononciations à fournir au moteur de synthèse devrait s'accompagner d'une adaptation de la voix qui sous-tend le système de synthèse. Comme cette solution est coûteuse, une alternative plus légère est de prendre en compte la stratégie de prononciation de la voix pour trouver un compromis entre le style désiré et celui de la voix. La prochaine section présente des travaux en ce sens.

2.2.3 Adaptation à une voix de synthèse

En synthèse de la parole, étant donnée une séquence de phonèmes d'entrée, le signal de parole synthétique correspondant est généré en interrogeant un modèle génératif (synthèse paramétrique) ou une base de segments de parole (synthèse par sélection d'unités). Dans les deux cas, le système a été construit en utilisant un corpus de parole dans lequel les phonèmes réalisés ont été soigneusement étiquetés et segmentés acoustiquement. Par conséquent, les systèmes de synthèse dépendent fortement de la cohérence entre les phonèmes étiquetés dans leur corpus de parole sous-jacent et ceux générés par le phonétiseur pendant la synthèse. Dans le cas de la sélection d'unités, les incohérences se traduisent par une irrégularité du signal et un nombre élevé de (mauvaises) concaténations. Dans des systèmes paramétriques, les modèles ont du mal à généraliser leur prédictions acoustiques à des situations rarement vues lors de l'apprentissage et produisent des signaux étouffés. Pour résoudre ce problème, nous avons proposé d'exploiter notre méthode d'adaptation pour adapter les phonèmes générés par le phonétiseur à ceux du corpus de la parole du système de synthèse.

Les expériences ont été réalisées sur un corpus en français dédié à la construction d'un système de synthèse vocale pour un serveur vocal interactif. À ce titre, ce corpus couvre tous les diphones du français et comprend les mots les plus utilisés dans le domaine des télécommunications. Il dispose d'une voix féminine neutre. Le corpus est composé de 7 208 énoncés, contenant 225K phonèmes pour un total d'environ 7 heures de parole. Les prononciations réalisées ont été contrôlées pendant le processus d'enregistrement de la voix.

Les prononciations canoniques des énoncés ont été générées avec Liaphon (BÉCHET 2001). Leur taux d'erreurs est de $11,2\,\%$ par rapport aux phonèmes réalisés. La plupart des confusions concerne des allophones : /o/ ≒ /ɔ/, /e/ ≒ /ɛ/ et /ɛ̃/ ≒ /œ̃/. De telles confusions ne peuvent pas être considérées comme des erreurs sur le plan phonologique en français. Elles dépendent du style de parole. De même, beaucoup d'insertions sont des schwas (/ə/) dont l'élision ou la présence est mal prédite. Beaucoup de suppressions sont des liaisons réalisées mais non prédites, telles que /t‿/ et /z‿/. Les autres différences concernent des stratégies d'annotations et choix d'alphabet, par exemple /ɲ/ ≒ /nj/, ou /ə/ ≒ /∅/.

En utilisant les familles de descripteurs linguistiques, articulatoires et prosodiques présentés précédemment, l'évaluation objective aboutit à des conclusions comparables. À nouveau, l'évaluation perceptuelle de la méthode a été réalisée par un test de préférence auprès de locuteurs natifs (du français cette fois-ci). Comme l'illustrent les préférences des testeurs sur la figure 2.6, l'imitation du style phonétique de la voix amène des gains significatifs. La qualité des signaux issus de prononciations adaptées n'est pas aussi bonne que celle obtenue avec les prononciations réalisées mais elle s'en approche car les utilisateurs n'ont pas de préférence dans la majorité des cas. Ce résultat est d'autant plus intéressant que nous avons montré que 5 minutes de parole annotées en phonèmes réalisés suffisent pour apprendre un bon modèle d'adaptation. Ainsi, cette contribu-

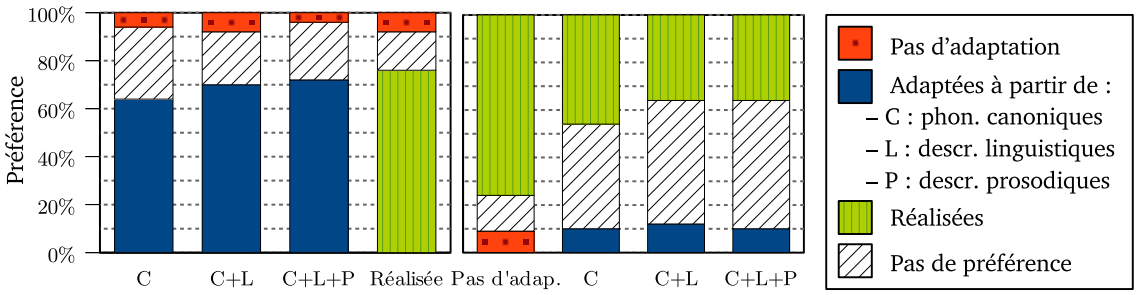


FIGURE 2.6 – Comparaisons entre les prononciations canoniques, réalisées et adaptées en terme de préférence de signaux synthétisés par un système paramétrique (HTS).

tion a été intégrée dans le système que nous avons proposé au challenge annuel international de synthèse de la parole *Blizzard Challenge* (LOLIVE et al. 2017).

2.2.4 Adaptation aux émotions

Parmi les problématiques actuelles en synthèse de la parole, la production de signaux de parole reflétant divers états émotionnels est un thème de recherche actif, que nous avons également développé dans le cadre de nos travaux sur les variantes de prononciation. Pour cela, par le biais de notre méthode d'adaptation, nous avons exploité des prononciations transcrites à partir d'un corpus de parole émotionnelle. Ce corpus est composé de signaux de parole porteurs des émotions du *big six* (colère, dégoût, joie, peur, surprise, tristesse). Ces énoncés sont joués par un acteur avec un haut degré d'activation et sur des énoncés spécifiques à chaque émotion. En parallèle, chaque énoncé a également été joué de manière neutre. Sur cette base, nous avons expérimenté l'apprentissage de plusieurs types de modèles d'adaptation et leurs combinaisons. En particulier, nous avons appris des modèles spécifiques à chaque émotion, un modèle global pour toute les émotions, un modèle pour les signaux neutres et un modèle adapté à la voix de notre système de synthèse.

En terme de taux d'erreurs, la comparaison de ces multiples modèles et des prononciations de référence (neutres ou émotionnelles) montre que les adaptations émotionnelles permettent bien de se rapprocher de la prononciation émotionnelle jouée. Sur le plan de la perception, il ressort que la double adaptation à la voix et à une prononciation émotionnelle améliore simultanément la qualité et l'expressivité avec des phrases expressives, notamment par rapport à une simple adaptation à l'un ou l'autre des deux facteurs.

Ces travaux ont également abordé la question de la caractérisation des prononciations émotionnelles. Nous avons étudié les variations de prononciation dans l'expression de différents états émotionnels, d'abord sur la parole naturelle, puis sur les prononciations générées. La première conclusion est que les descripteurs prosodiques jouent un rôle important dans la suppression et l'assimilation des phonèmes. L'étude montre également que la distinction de styles de prononciation propres à chaque émotion est difficile à établir mais que celle d'un style émotionnel globale apparaît clairement par rapport aux phonétisations neutres. Enfin, il ressort de nos expériences que la prononciation émotionnelle n'est pas perçue si elle n'est pas accompagnée d'une cohérence lexicale et prosodique.

Conclusion

Dans ce chapitre, nous avons présenté différentes contributions autour de la notion de variabilité et de style dans la prononciation d'énoncés. La première porte sur la construction de treillis de phonèmes qui représentent des possibilités de variantes autorisées dans la langue. L'adaptation à

différents styles peut alors se traduire par la repondération des arcs, sans risque de produire des variantes incorrectes. Alternativement, nous avons vu que la production de variantes spécifiques à un style peut être effectué par un modèle d'adaptation. Dans ce cas, les variantes autorisées sont plus libres. Nous avons montré que ce procédé est efficace et générique puisque nous l'avons appliqué avec succès à trois styles différents. Ces travaux ont été intégrés au sein du système de synthèse de la parole de l'IRISA. La principale limite observée est la difficulté du système à produire le signal de prononciations différentes de celles utilisées pour sa construction. Malgré les avancées des approches par réseaux de neurones, cela reste un problème et donc une perspective majeure à développer (*cf.* chapitre 8).

CHAPITRE 3
Insertion automatique de disfluences

Les disfluences sont un phénomène qui interrompt le discours sans ajouter aucun contenu propositionnel (TREE 1995). Elles apparaissent principalement lorsque l'élocution va plus vite que le processus de pensée, ce qui est particulièrement fréquent lorsque le locuteur n'a pas préparé son discours. Les disfluences jouent un rôle important dans le discours. Elles en améliorent la compréhension par des auditeurs, signalent la complexité de propos à venir (ROSE 1998 ; TREE 2001) et, dans le cas d'un dialogue, facilitent la synchronisation entre interlocuteurs (CLARK 2002). De nombreux travaux se retrouvent – quoique la terminologie puisse différer – sur une catégorisation des disfluences en trois grandes familles : les *pauses*, les *répétitions* et les *révisions* (BOULA DE MAREÜIL et al. 2005 ; E. E. SHRIBERG 1999 ; TSENG 1999).

En TALN, la majorité des travaux sur les disfluences se situe dans le domaine de la reconnaissance automatique de la parole (HASSAN et al. 2014 ; KAUSHIK et al. 2010 ; LIU et al. 2006 ; STOLCKE et E. SHRIBERG 1996 ; STOLCKE, E. SHRIBERG et al. 1998). Ces travaux ont pour principal objectif d'intégrer ce phénomène dans le modèle de langage des systèmes ou de produire des transcriptions automatiques nettoyées de toutes éventuelles disfluences. Ainsi, ils se sont davantage intéressés à la forme de surface des disfluences qu'au processus qui conduit à leur production. Les études en synthèse de la parole se sont faites plus rares (ADELL, BONAFONTE et ESCUDERO 2007 ; ADELL, BONAFONTE et MANCEBO 2008 ; ADELL, ESCUDERO et al. 2012 ; DALL et al. 2014 ; SUNDARAM et NARAYANAN 2003). ADELL, BONAFONTE et MANCEBO (2008) l'expliquent en partie par l'absence de disfluences dans les corpus de parole sur lesquels les système de synthèse s'appuient et la difficulté accrue pour les pré-traitements linguistiques à travailler sur des textes disfluents. Dans l'ensemble, les travaux sur l'ajout de disfluences se focalisent sur l'insertion de pauses, éventuellement catégorisées en types (BETZ et al. 2015), mais ils ne proposent rien pour les autres disfluences. Ainsi, à moins que l'utilisateur n'explicite des disfluences dans le texte à synthétiser, les signaux produits sont généralement perçus comme peu spontanés.

Ce chapitre porte sur les travaux que j'ai menés sur l'ajout automatique de disfluences dans une perspective de synthèse de la parole avec un style plus spontané. La section 3.1 présente une première proposition fondée sur un principe de composition de disfluences élémentaires. La section 3.2 présente une autre étude, moins aboutie, portant sur l'utilisation de réseaux de neurones séquence-à-séquence. Enfin, les discussions sur l'évaluation de ces propositions sont rassemblées au sein de la section 3.3.

3.1 Composition de disfluences élémentaires

Encadrements*/collaborations : *Raheel Qader (doctorat), Damien Lolive (IRISA), Pascale Sébillot (IRISA)
Références :

— QADER, R. (2017). *Pronunciation and disfluency modeling for expressive speech synthesis* (thèse de doct., University of Rennes 1)
— QADER, R., **LECORVÉ**, G., LOLIVE, D. & SÉBILLOT, P. (2017). Ajout automatique de disfluences pour la synthèse de la parole spontanée : formalisation et preuve de concept. *Traitement automatique du langage naturel (TALN)*. **Prix du meilleur article**
— QADER, R., **LECORVÉ**, G., LOLIVE, D. & SÉBILLOT, P. (2018). Disfluency Insertion for Spontaneous TTS : Formalization and Proof of Concept. *Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP)*. Springer

Plusieurs études ont montré que les disfluences présentent des regularités structurelles (CLARK 1996 ; LEVELT 1983 ; E. E. SHRIBERG 1994). Le schéma dominant, proposé par Shriberg, décrit la structure d'une disfluence comme une suite de mots dont certaines sections jouent un rôle particulier. Ainsi, une portion de texte disfluente peut s'écrire comme $\langle RM, IM, RR \rangle$ où :

— RM est une séquence de mots erronée appelée *reparandum* ;
— RR est la séquence de mots corrigée correspondant à RM, cette nouvelle séquence étant dénommée *réparation* dans la suite ;
— IM est l'espace dit *interregnum* qui marque l'interruption du flot de parole par un silence ou des marqueurs spécifiques.

Dans cette structure, la frontière entre le *reparandum* et l'espace *interregnum* est appelée *point d'interruption* (PI). Par exemple, la phrase « *Show me flights from Boston on uh from Denver on Monday.* » peut ainsi être découpée comme suit :

$$\textit{Show me flights } \overbrace{\textit{from Boston on}}^{RM} \underset{PI}{\uparrow} \overbrace{\textit{uh}}^{IM} \overbrace{\textit{from Denver on}}^{RR} \textit{ Monday.} \qquad \text{(Exemple 1)}$$

Dans la première méthode que nous avons proposée, une disfluence est vue comme le résultat d'une fonction de transformation d'une phrase fluide. Ainsi, un énoncé avec de multiples disfluences résulte d'une successions de transformations atomiques. Pour cela, nous divisons le schéma générique de Shriberg en sous-schémas, chacun dédié à une famille de disfluences (pause, répétition ou révision) et adapté à ses spécificités structurelles, et définissons une fonction de transformation f_T pour chaque famille T. En nous inspirant des travaux de l'état de l'art sur l'insertion de pauses, ces fonctions de transformation déterminent la position du point d'interruption à considérer, puis insère les mots de la disfluence à l'endroit choisi.

3.1.1 Sous-schémas élémentaires et composition

Pour chaque famille de disfluences, une instanciation particulière du schéma de Shriberg est formulée. Nous résumons ici ces instanciations, ainsi que le principe de composition des disfluences. Pour illustrer nos propos, nous prenons l'exemple de la séquence de mots fluide $\mathbf{w}$ ci-dessous :

$$\mathbf{w} \ : \quad \textit{je souhaite que tu viennes.} \qquad \text{(Exemple 2)}$$

Pauses. Syntaxiquement, les pauses sont de simples interruptions dans un énoncé. Elles ne contiennent aucun *reparandum* ni donc aucune réparation, et se réduisent alors à leur seul segment *interregnum* dont les valeurs possibles sont généralement un silence, un silence rempli (en français, « euh », « hmm ») ou des marqueurs lexicaux (par exemple, « enfin », « je veux dire »).

$$f_{pause}(\mathbf{w}) \ : \quad \textit{je souhaite que } \underset{PI}{\uparrow} \overbrace{\textit{euh}}^{IM} \textit{ tu viennes.} \qquad \text{(Exemple 3)}$$

Répétitions. Les répétitions sont la duplication d'une portion de texte. Leurs *reparandum* et réparation sont donc identiques. Par ailleurs, en raison du mécanisme de composition proposé, nous considérons que l'espace *interregnum* entre ces deux régions est vide, charge à la fonction d'insertion de pauses d'en ajouter éventuellement une (ou plusieurs).

$$f_{répétition}(\mathbf{w}) \ : \quad \textit{je souhaite } \overbrace{\textit{que}}^{RM} \ \underset{\uparrow}{\underset{PI}{}} \ \overbrace{\textit{que}}^{RR} \textit{ tu viennes.} \qquad \text{(Exemple 4)}$$

Révisions. Dans le même esprit, la fonction de révision $f_{révision}$ positionne le PI, délimite la zone de réparation, puis produit un *reparandum* lui correspondant, sans aucun espace *interregnum*. À la différence des répétitions, la production du *reparandum* n'est pas directe car il s'agit de produire un énoncé factice à réparer, généralement sous des contraintes de proximité sémantique ou phonétique avec la réparation.

$$f_{révision}(\mathbf{w}) \ : \quad \overbrace{\textit{je veux que}}^{RM} \ \underset{\uparrow}{\underset{PI}{}} \ \overbrace{\textit{je souhaite que}}^{RR} \textit{ tu viennes.} \qquad \text{(Exemple 5)}$$

Composition. Notre processus de production de disfluences est présenté par le graphe de la figure 3.1. Partant d'une phrase fluide, les transformations de chaque famille peuvent s'appliquer zéro, une ou plusieurs fois. Par conséquent, les fonctions de transformations doivent être capables de traiter en entrée des énoncés fluides comme disfluents. Un ordre de précédence entre familles est imposé afin de désambiguïser la production de certains énoncés contenant de multiples disfluences. L'exemple ci-dessous présente un déroulé possible de l'insertion de disfluences par de multiples compositions successives :

$$\begin{aligned}
& \textit{Je souhaite que tu viennes.} \\
f_{révision} \quad & [\textbf{Je veux que } \textit{je souhaite que}]_{\text{rév.}} \textit{ tu viennes.} \\
\circ \ f_{répétition} \quad & \Big[\textit{Je veux } [\textbf{que } \textit{que}]_{\text{rép.}} \textit{ je souhaite que}\Big]_{\text{rév.}} \textit{ tu viennes.} \qquad \text{(Exemple 6)} \\
\circ \ f_{pause} \quad & \Big[\textit{Je veux } [\textit{que que}]_{\text{rép.}} \, [\textbf{euh}]_{\text{pause}} \textit{je souhaite que}\Big]_{\text{rév.}} \textit{ tu viennes.} \\
\circ \ f_{pause} \quad & \Big[\textit{Je veux } [\textit{que que}]_{\text{rép.}} \, [\textit{euh}]_{\text{pause}} \, [\textbf{enfin}]_{\text{pause}} \textit{je souhaite que}\Big]_{\text{rév.}} \textit{ tu viennes.}
\end{aligned}$$

3.1.2 Implémentation

Nous avons proposé de modéliser la prédiction des PI comme une tâche d'étiquetage automatique que nous traitons par des CRF, un pour chaque famille T. L'insertion de nouveaux mots est, quant à elle, traitée comme l'énumération de toutes les hypothèses disfluentes possibles pour T et la sélection parmi celles-ci de la plus probable d'après un modèle de langage. Ces différents modèles sont mis en musique par un algorithme qui passe en revue les révisions, les répétitions, puis les pauses. Pour chaque famille, une décision est prise de rester à la famille courante ou de passer à la suivante d'après un critère défini par l'utilisateur. Ce critère peut être le taux ou le nombre de disfluences présentes dans l'énoncé en cours de transformation, ou encore un seuil sur la probabilité donnée par le CRF chargé de positionner le PI à venir.

Notre implémentation a été testée sur le corpus Buckeye (PITT et al. 2005), déjà présenté dans le chapitre 2. Sur 20 heures de parole spontanée, $20\,264$ pauses, $2\,714$ répétitions et 550 révisions ont été annotées. Pour chaque famille de disfluences, une version spécifique du corpus est dérivée en ne retenant que les phrases contenant au moins une disfluence de cette famille, dans le respect

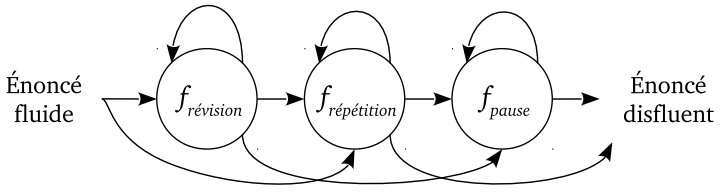
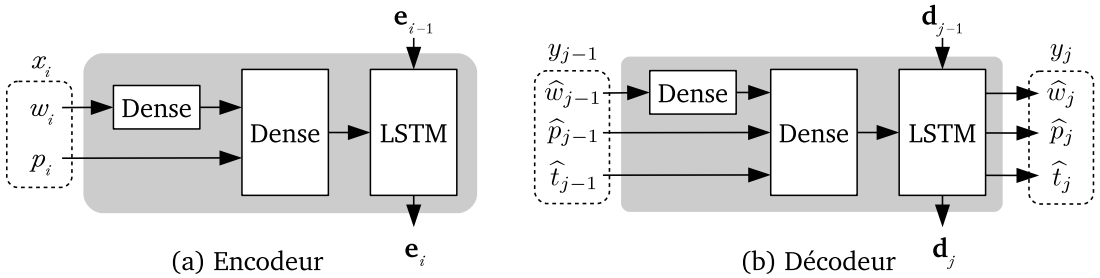


FIGURE 3.1 – Processus complet de production des disfluences.



(a) Encodeur (b) Décodeur

FIGURE 3.2 – Topologie de l'encodeur et du décodeur pour l'insertion de disfluences.

de l'ordre de précédence. Par exemple, le corpus dédié aux répétitions a été nettoyé de toutes les pauses mais il contient des révisions. Les prédictions des PI s'appuient sur les mots, leur catégorie grammaticale ainsi qu'une information sur leur nature fluide ou disfluente, par exemple pour indiquer au modèle propre aux répétitions que tel mot fait partie d'une révision.

3.2 Insertion par un modèle séquence-à-séquence

Encadrements/collaborations : ⋆Henri Lasselin (stage de master)
Référence :
— LASSELIN, H. (2018). *Make text look like speech : disfluency generation using sequence-to-sequence neural networks* (Rapport de recherche, Rapport de recherche. IRISA)

Nous avons exploré l'utilisation de réseaux de neurones séquence-à-séquence pour l'insertion de disfluences. À l'inverse de la précédente approche, toutes les disfluences sont ici ajoutées en même temps. Pour cela, nous avons adopté une architecture encodeur-décodeur, telle qu'introduite par (SUTSKEVER et al. 2014). Pour une séquence d'entrée $\mathbf{x} = [x_1, \ldots, x_n]$ et une séquence de sortie $\mathbf{y} = [y_1, \ldots, y_m], m \geq n$, son principe est le suivant :

Encodeur : L'encodeur E lit chaque élément de la séquence d'entrée et accumule progressivement les informations jusqu'à finalement produire un plongement de toute la séquence. L'accumulation d'information est effectuée par une couche récurrente qui, à la position $i \in [\![1, n]\!]$, prédit un nouvel état interne $\mathbf{e}_i \in \mathbb{R}^P$ à partir de celui de la position précédente et du nouveau mot lu, soit :

$$\mathbf{e}_i = E(x_i, \mathbf{e}_{i-1}) \tag{3.1}$$

et avec P une dimension fixée par le modèle, 256 dans notre cas. Ainsi, le plongement de la séquence d'entrée n'est autre que le dernier état interne $\mathbf{e}_n$.

Décodeur : Le décodeur D produit des symboles de sorties y_j. Il recourt également à une couche récurrente où, à chaque étape $j \in [\![0, m]\!]$, le décodeur prédit le mot y_j à partir du mot de la précédente prédiction y_{j-1} et de l'état interne du décodeur $\mathbf{d}_j$. L'état interne du décodeur est initialisé avec le plongement de la séquence d'entrée et le processus de décodage est amorcé par un symbole factice de début, soit formellement :

$$\begin{cases} (y_0, \mathbf{d}_0) &= (\texttt{<début>}, \mathbf{e}_n) \\ (y_i, \mathbf{d}_j) &= D(y_{j-1}, \mathbf{d}_{j-1}) \,. \end{cases} \tag{3.2}$$

Le processus de décodage s'arrête lorsqu'un symbole spécifique de fin est prédit ($y_{m+1} = \texttt{<fin>}$).

Le modèle construit, détaillé sur la figure 3.2, représente les entrées x_i par des couples contenant le mot w_i et sa catégorie grammaticale p_i. Les éléments y_j du décodeur portent sur des mots

prédits $\widehat{w}_j$ et leur catégorie grammaticale estimée $\widehat{p}_j$. Ils exploitent en plus une étiquette $\widehat{t}_j$ indiquant si l'élément y_j est une disfluence d'un type donné ou non. Cette information est à la fois utile lorsque l'élément y_j est utilisé comme entrée du décodeur à l'étape $j+1$ et pour aider le modèle à cerner les propriétés de chaque famille de disfluences lors de l'apprentissage. L'encodage de ces étiquettes suit une stratégie BIO (*Beginning-Inside-Outside*) déclinée pour chaque famille de disfluences. Les mots passent dans une couche de plongement dont la sortie est fusionnée avec les autres informations, le tout étant finalement transmis à la couche récurrente, en l'occurence un LSTM pour l'encodeur comme pour le décodeur.

Pour valider l'approche dans un contexte maîtrisé et plus simple, nous avons expérimenté cette approches sur des corpus de textes artificiels. Des énoncés fluides ont été générés *via* une grammaire hors-contexte probabiliste sur un vocabulaire fortement limité. Des variantes disfluentes en sont dérivés selon des règles *ad hoc* probabilisées. 50 000 énoncés très disfluents sont ainsi générés, contenant 15 000 disfluences de chaque famille. Nous avons également travaillé sur des données réelles du français grâce au corpus RATP-DECODA qui recense des conversations téléphoniques collectées par le centre d'appel de la RATP à Paris. Il représente plus de 60 heures de parole et contient 42 000 pauses, 6 000 répétitions et 2 000 révisions.

3.3 Évaluation d'énoncés disfluents

L'insertion de disfluences n'est pas une tâche-phare pour laquelle les protocoles d'évaluation (métriques, corpus) sont bien tout établis. Par ailleurs, comme dans d'autres tâches de génération de contenu, il existe de multiples variantes disfluentes acceptables pour un énoncé fluide donné mais les corpus d'apprentissage n'en donnent qu'une seule pour l'évaluation. Cette dernière section vise donc à partager quelques réflexions pour les évaluations objectives et perceptuelles.

3.3.1 Mesures objectives

Nous avons proposé plusieurs métriques d'évaluation. Celles-ci découlent de deux propriétés à respecter et de deux hypothèses qui sous-tendent les méthodes développées.

Propriété 1 (intégrité) : *Un énoncé disfluent est constitué de son énoncé fluide d'origine et de portions disfluentes.*
Pour mesurer le respect de cette propriété, il suffit de nettoyer des énoncés disfluents et de les aligner avec la séquence fluide initiale. Les métrique qui en découlent s'appuient sur des statistiques de nombres d'insertion, suppression et substitution de mots.

Propriété 2 (conformité) : *Chaque disfluence est conforme au schéma de la famille à laquelle elle appartient.*
Suite aux choix de la section 3.1.1, les pauses produites doivent appartenir au vocabulaire des pauses (certains mots ne peuvent pas constituer une pause), les répétitions doivent avoir un *reparandum* identique à la réparation et enfin, le *reparandum* d'une révision doit corriger sa réparation (cela suppose que ces deux parties doivent être différentes). La conformité d'un énoncé peut donc facilement s'implémenter pour chaque famille de disfluences comme un critère binaire. À l'échelle d'un corpus, nous avons proposé d'utiliser la conformité moyenne des disfluences générées comme indicateur de qualité.

Hypothèse 1 : *Il existe des positions de préférences pour insérer les disfluences. Ces positions dépendent du type de disfluence et de son contexte.*
Cette hypothèse justifie le repérage de PI par un modèle statistique. Elle a été étayée sur des données réelles (DALL et al. 2014). Il en découle des mesures classiques de rappel, précision et F-mesure par rapport à la position dans les énoncés disfluents de référence. À l'échelle d'un corpus, nous avons également proposé d'observer le nombre de disfluences prédites rapporté à leur nombre dans la référence, indépendamment de leur position. Ceci permet de savoir si la méthode testée n'insère pas trop ou trop peu de disfluences.

Hypothèse 2 : *Le contexte d'une disfluence joue un rôle sur le contenu de la disfluence.*
Cette seconde hypothèse sous-tend, quant à elle, l'utilisation de modèles de langage pour sélectionner une proposition de disfluences parmi plusieurs. Il s'agit de la dimension la plus difficile à évaluer automatiquement. Une approche prudente est d'utiliser des mesures propres aux modèles de langage, par exemple la perplexité des énoncés fluides et disfluents.

Ces approches ont comme principal intérêt de permettre la comparaison de modèles lors des réglages inhérents à toute approche. Notamment, des bornes basses peuvent être calculées par comparaison à des approches aléatoires (position et/ou contenu des disfluences).

L'emploi de ces métriques nous a permis de constater que l'approche par compositions successives d'insertions a l'avantage de garantir, par construction, les propriétés d'intégrité et de conformité. À l'inverse, l'approche neuronale séquence-à-séquence proposée peut induire des altérations de la séquence d'origine. Les expérimentations montrent que l'apprentissage d'un simple auto-encodeur, sans insertion de disfluences, conduit à une distorsion de l'énoncé initial. Ensuite, l'étude des PI montre que leurs précision et rappel sont faibles en raison de la non systématicité des disfluences dans le langage naturel mais meilleurs que l'aléatoire. Les pauses sont les plus faciles à prédire, avec une F-mesure de $0,25$ sur le corpus en anglais avec l'approche par composition et de $0,21$ sur le corpus français pour l'approche neuronale. Par nature, les répétitions et révisions sont, quant à elles, plus rares et leurs prédictions est beaucoup plus hasardeuse.

3.3.2 Évaluation perceptuelle

Compte tenu des faibles performances objectives de l'approche neuronale en termes de préservation de la séquence d'origine, seule l'approche par composition a été évaluée à ce jour auprès de testeurs. À partir de textes fluides, les testeurs devaient imaginer comment ceux-ci pourraient être énoncés lors d'une conversation spontanée et donner leur opinion sur plusieurs propositions présentées, selon une échelle allant de 0 (énoncé impossible) à 10 (parfaitement possible). Le choix de présenter des textes, et non des signaux de parole, a été fait car aucun système de synthèse n'est en capacité de synthétiser convenablement des disfluences et que l'enregistrement par un acteur de tous les énoncés et de leurs variantes était inconcevable. Globalement, les résultats des tests nous montrent que les énoncés produits par notre méthode sont acceptables en comparaison à des textes fluides ou disfluents de référence. Néanmoins, il ressort également que les différences sont faibles et rendent impossible toute analyse appronfondie. Ceci s'explique sans doute par une difficulté des testeurs à se projeter dans la tâche demandée, notamment en imaginant un discours oral à partir d'un texte.

Conclusion

Ce chapitre a retracé différents travaux et plusieurs réflexions menées autour de la notion de disfluences, composante importante de l'oral mais pourtant globalement négligée. Nous avons présenté deux techniques d'insertions de disfluences, conçues dans la perspective d'apporter un pouvoir d'expressivité plus grand à des systèmes de synthèse de la parole. La première approche s'est appuyée sur un cadre linguistique issu de la littérature du domaine, alors que la seconde a laissé libre cours au pouvoir de modélisation des réseaux de neurones. Dans un dernier temps, nous avons fait ressortir qu'un aspect commun à ces travaux est la difficulté d'évaluer cette tâche, la solution la plus fiable qui consiste à enregistrer un acteur pour jouer des énoncés étant également la plus coûteuse. De manière analogue aux conclusions du précédent chapitre, le principal obstacle à la poursuite de ces travaux est la difficulté encore actuelle à à faire restituer les disfluences par des systèmes de synthèse de la parole. Notons néanmoins que l'insertion automatique de disfluences est dores et déjà parfaitement envisageable dans un contexte purement textuel, par exemple pour humaniser des *chatbots*. Dans ce cas, il est probable qu'il faille toutefois adapté les familles considérés aux spécificités de l'écrit numérique (autres marqueurs, fautes de frappe...).

CHAPITRE 4

Registres familier, courant et soutenu

La notion de registre renvoie à la manière dont les productions linguistiques sont évaluées et catégorisées au sein d'une même communauté linguistique. Elle est abordée dans de multiples travaux en linguistique comme en sociolinguistique. Ainsi, FERGUSON (1982) définit les registres comme une variation « *dans laquelle la structure linguistique varie en fonction des occasions d'utilisation* ». URE (1982) associe cette variation aux activités humaines : « *chaque communauté linguistique a son propre système de registres... correspondant à l'éventail des activités que ses membres exercent normalement* ». Selon l'angle d'étude privilégié, on observe dans la littérature linguistique diverses manières de partitionner l'espace linguistique en différents registres. Par exemple, il peut s'agir de distinguer les registres familier, populaire et vulgaire dans des journaux satiriques (ILMOLA 2012), l'influence de différents médias de communication (CHARAUDEAU 1997), des degrés de spécialisation (BORZEIX et FRAENKEL 2005 ; MOIRAND 2007) ou encore l'opposition entre communication fonctionnelle et relationnelle (BORZEIX et FRAENKEL 2005). Cette diversité laisse par ailleurs apparaître une difficulté terminologique puisque les dénominations « niveau », « style » ou encore « genre » co-existent avec celle de « registre » et font l'objet de débats (BELL 1984 ; BIBER 1991 ; BIBER et FINEGAN 1994 ; GADET 1996a,b). Notre travail ne visant pas des contributions sur ces aspects définitoires, nous adoptons arbitraitement le terme de registre, issu de la tradition britannique (SANDERS 1993 ; URE 1982).

Ce chapitre retranscrit des travaux effectués sur le cas des registres familier, courant et soutenu. Ce découpage est avant tout motivé par le pragmatisme car il est relativement consensuel et peu sujet à ambiguïté pour l'étiquetage manuel d'un ensemble de données initial, tout en n'interdisant pas d'éventuels raffinements pour l'avenir.

Les travaux résumés dans ce chapitre s'intègrent dans le cadre du projet exploratoire ANR TREMoLo, dont je suis le coordinateur scientifique, qui vise l'étude des registres de langue et le développpément de méthodes automatiques de transformation de textes d'un registre vers un autre. Ce travail s'appuie sur l'extraction de patrons linguistiques spécifiques à des registres donnés et sur leur prise en compte dans un processus de production automatique de paraphrases. Dans ce chapitre, nous parcourons les travaux déjà réalisés sur la caractérisation linguistique des registres (section 4.1), la constitution d'un corpus textuel annoté en registres (section 4.2) et l'extraction automatique de motifs langagiers discriminants (section 4.3).

4.1 Caractérisation linguistique des registres

Projets : TREMoLo (coordinateur – ANR – 2017-2021 – IRISA, MoDyCo)
Encadrements⋆/collaborations : ⋆Jade Mekki (stage de master, doctorat), Delphine Battistelli (MoDyCo), Nicolas Béchet (IRISA)

Références :

— MEKKI, J., BATTISTELLI, D., BÉCHET, N. & **LECORVÉ**, G. (2017). « *Nous nous arrachâmes promptement avec ma caisse » : quels descripteurs linguistiques caractérisent les registres de langue ?* Rapport de recherche. IRISA, MoDyCo

— MEKKI, J., BATTISTELLI, D., **LECORVÉ**, G. & BÉCHET, N. (2018). Identification de descripteurs pour la caractérisation de registres. *Actes des Rencontres Jeunes Chercheurs (RJC) de la conférence CORIA-TALN*

Dans un travail préliminaire (MEKKI, BATTISTELLI, BÉCHET et al. 2017 ; MEKKI, BATTISTELLI, **LECORVÉ** et al. 2018), nous avons cherché à produire une vérité terrain des caractéristiques discriminantes des registres familier, courant et soutenu les uns par rapport aux autres. Pour cela, nous avons choisi de partir de caractéristiques linguistiques identifiées dans la littérature, que nous avons catégorisées, puis complétées, et de les valider ou invalider en corpus.

Nous avons réalisé des comparaisons de fréquences d'une caractéristique linguistique entre corpus associés à chaque registre. Pour un corpus donné, chaque caractéristique étudiée (par exemple, l'emploi du mot « ça ») est décrit par sa fréquence d'apparition relative, c'est-à-dire normalisée par la longueur en mots du corpus. Considérant trois corpus textuels, chacun spécifique à un registre, nous posons alors un descripteur comme valide pour un registre donné si, parmi les différents corpus, la valeur du descripteur est significativement [1] supérieure pour le corpus dédié à ce registre à celle des autres corpus. Cette approche est volontairement simpliste pour rester indépendante d'un maximum d'hypothèses. Notre travail ne prétend ainsi pas statuer de manière absolue sur la validité de tel ou tel descripteur mais dresse un panorama du niveau de saillance d'un large panel de descripteurs. Cet étalonnage a pour finalité d'offrir un point de départ et de comparaison à nos travaux d'extraction automatique de motifs discriminants par des méthodes de fouille.

Nous avons dressé une liste de 72 descripteurs, émanant soit de la littérature (66), soit d'une analyse préliminaire conduite par nos soins pour les différents registres (6 descripteurs supplémentaires ont ainsi été identifiés pour le registre familier). Ces descripteurs sont généralement soit des motifs « fautifs » (au sens d'un écart à la norme), soit des motifs corrects (toujours selon la norme) mais rares. Ces descripteurs couvrent divers niveaux d'abstraction de la langue que nous regroupons sous les catégories lexicale (16 descripteurs), morphologique (16), syntaxique (38) et phonétique (2). En voici quelques exemples :

— Niveau lexique : emploi de « ça », d'onomatopées, d'éléments ponctuants ;
— Niveau morphosyntaxique : contraction « cela est »/« c'est », terminaison en « -ou », verbe du premier groupe ;
— Niveau syntaxique : inversion sujet-verbe dans une question, proposition relative en « que » ;
— Niveau phonétique : élision du « e » , apocope du « r ».

Dans l'absolu, l'appartenance explicite de certains lexèmes à un registre donné est également très discriminante. Nous n'avons cependant pas traité cet aspect car il suppose des lexiques suffisamment exhaustifs [2]. De plus, nous n'avons pas eu recours à une mesure de richesse lexicale car cette notion nous a semblé délicate à traiter. De fait, plus il y a de différents termes lexicaux employés, plus nous pouvons supposer que le vocabulaire est riche donc soutenu. Toutefois, le registre familier est également reconnu pour sa créativité. Pour être efficace, la mesure lexicale devrait ainsi s'appuyer sur une distinction entre termes standards et exotiques, par exemple sur la base d'un dictionnaire à nouveau. Enfin, notons que l'étude de descripteurs phonétiques fait sens y compris pour une analyse de textes écrits car l'usage écrit de formes orales est désormais répandu à travers des modes de communications connectés (chats, messageries, textos...).

Nous avons examiné 30 descripteurs parmi les 72 répertoriés, les 42 restants nécessitant soit davantage de ressources textuelles (par extension du corpus initial), soit le recours à des outils d'extraction plus ou moins complexes. Les trois types de corpus considérés sont composés d'écrits

1. S'agissant d'un travail préliminaire, ce critère de significativité n'a pas été formalisé mathématiquement.
2. La construction de ces lexiques est néanmoins réaliste car les dictionnaires renseignent souvent leurs entrées en terme de registre de langue.

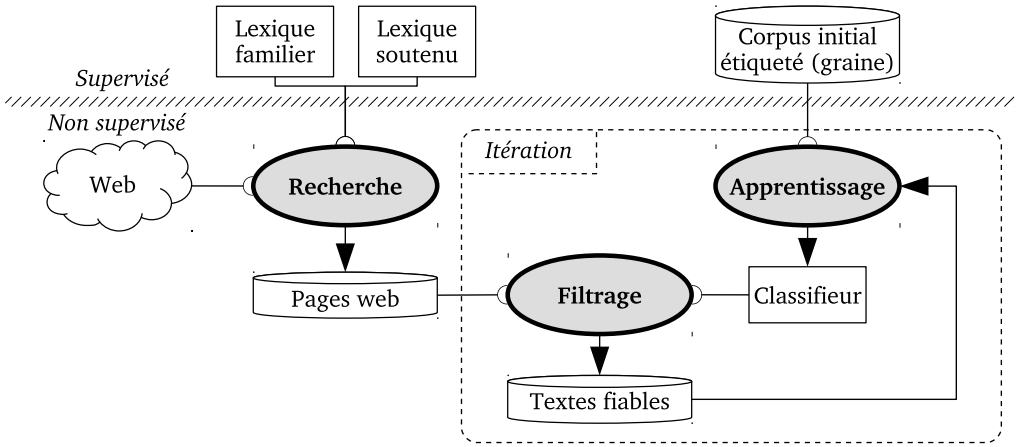


FIGURE 4.1 – Vue d'ensemble du processus semi-supervisé.

français : *Albertine disparue* de Proust pour le registre soutenu, des archives de *L'Humanité* et *Le Monde* pour le registre courant et *Kiffe Kiffe Demain* de Guene, *L'Assommoir* de Zola et *Voyage au bout de la nuit* de Céline pour le registre familier. Il ressort que la majorité des descripteurs validés concernent le registre familier. Ce constat s'explique par la créativité de ce registre vis à vis de la norme, qu'il s'agisse de la richesse lexicale puisant dans les lexiques populaires, argotiques voire vulgaires, ou de la multiplicité de manière de s'écarter de la norme (et donc de produire des ≪ fautes ≫). Cette première étude en corpus a également mis en lumière la nécessité de traitements spécifiques pour les discours rapportés (identifiés *via* la présence de verbes de parole ou de marques de ponctuation telles que les guillemets). En effet, il existe une certaine corrélation entre l'oralité et le registre familier. L'analyse automatique de textes courants ou soutenus peut être biaisée par ce phénomène.

4.2 Constitution d'un corpus

Projets : TREMoLo (coordinateur – ANR – 2017-2021 – IRISA, MoDyCo)
Encadrements⋆/collaborations : ⋆Hugo Ayats (stage), ⋆Benoît Fournier (stage), ⋆Jade Mekki (stage de master, doctorat), Delphine Battistelli (MoDyCo), Nicolas Béchet (IRISA), Jonathan Chevelu (IRISA)
Références :

- LECORVÉ, G., AYATS, H., FOURNIER, B., MEKKI, J., CHEVELU, J., BATTISTELLI, D. & BÉCHET, N. (2018). Construction conjointe d'un corpus et d'un classifieur pour les registres de langue en français. *Traitement automatique du langage naturel (TALN)*
- LECORVÉ, G., AYATS, H., FOURNIER, B., MEKKI, J., CHEVELU, J., BATTISTELLI, D. & BÉCHET, N. (2019). Towards the Automatic Processing of Language Registers : Semi-supervisedly Built Corpus and Classifier for French. *International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*

Pour développer des approches de caractérisation automatique et de production de paraphrases pour les registres de langue, nous avons travaillé sur la constitution d'un corpus de grande taille annotés en registres familier, courant et soutenu. Nous avons développé une approche semi-supervisée pour y parvenir (LECORVÉ, AYATS et al. 2018, 2019).

Comme illustré sur la figure 4.1, le processus est amorcé par une étape de collecte de données sur le web (partie gauche). Cette collecte s'appuie sur deux lexiques spécialisés, l'un pour le registre familier, l'autre pour le soutenu, à partir desquels des requêtes familières ou soutenues

sont formées, puis soumises à un moteur de recherche. Après nettoyage automatique, les pages récupérées sont regroupées au sein d'un unique corpus dont on cherche à extraire les plus pertinentes pour chaque registre. Cette extraction se fait par le biais d'un classifieur probabiliste (un réseau de neurones) prédisant la probabilité d'appartenance à chaque registre. Pour résoudre l'interdépendance selon laquelle le classifieur nécessite des données d'entraînement étiquetées et l'étiquetage des données nécessite un classifieur, l'approche procède par itérations. Ainsi, un premier classifieur est initialement entraîné sur une graine, c'est-à-dire un faible ensemble initial de données annotées manuellement et indépendant des pages web récupérées. Ce classifieur permet de sélectionner les textes dont l'appartenance à l'un des registres est considérée comme fiable, à savoir dans notre cas si la probabilité d'appartenance à un registre est supérieure à un seuil donné. Ces textes sont ensuite ajoutés à ceux déjà étiquetés, puis une nouvelle itération démarre. Ce processus semi-supervisé permet en fin de processus d'obtenir conjointement un ensemble de textes catégorisés et un classifieur.

En pratique, les lexiques sur lesquels s'appuie la collecte de pages web sont constitués de mots et expressions automatiquement récupérés à partir d'une sauvegarde de la version française de Wiktionary[3]. Pour un registre donné, seuls les mots sans ambiguïté d'appartenance à un registre sont considérés, c'est-à-dire les termes ayant toutes leurs acceptions annotées comme appartenant à un même registre. La graine de textes manuellement étiquetés rassemblent 435 textes (environ $440\,000$ mots) répartis entre les registres familier, courant et soutenu et issus de romans, journaux et sites web. Enfin, le classifieur est un réseau de neurones multi-couches alimenté, pour chaque texte, par un vecteur de 46 descripteurs globaux, principalement issus des travaux de la section 4.1. Le choix de recourir à des vecteurs globaux et à une architecture relativement simple (réseau non récurrent) est justifiée par le fait que la graine est de taille modeste pour apprendre une architecture plus complexe.

Le résultat de l'approche est, d'une part, un corpus constitué de $800\,000$ textes représentant un total d'environ 750 millions de mots et, d'autre part, un classifieur dont le taux de bonne classification, mesurés sur différents corpus de test, oscille entre $60\,\%$ et $85\,\%$. Parmi les différents enseignements de ce travail, il ressort tout d'abord que le corpus final n'est pas équilibré alors que la graine l'était. Le registre courant prédomine ($48\,\%$ de textes) et les deux autres sont en proportions quasi égales. Cette prédominance du registre courant est cohérente avec son interprétation comme usage neutre de la langue. Ensuite, il apparaît que certains textes issus des requêtes familières soient finalement catégorisés comme soutenus, et réciproquement. Bien que cela soit légitime dans certaines situations, ce grand écart en terme de registre correspond généralement à de mauvaises prédictions. L'analyse montre que l'une des raisons à cela est, à nouveau, la présence de portions de discours rapportés ou directs. Une autre cause remarquable est la grande importance associés à la présence de termes spécifiques à un registre et dont la présence incohérente au milieu d'un texte peut pousser le classifieur à mal étiqueter un texte. Dans l'ensemble cependant, l'analyse des textes catégorisés montre que leur exploitation est possible pour des tâches de TALN.

4.3 Extraction de motifs discriminants

Projets : TREMoLo (coordinateur – ANR – 2017-2021 – IRISA, MoDyCo)
Encadrements⋆/collaborations : ⋆Inès Dabbebi (stage de master), ⋆Jade Mekki (doctorat), Delphine Battistelli (MoDyCo), Nicolas Béchet (IRISA)
Référence :

— DABBEBI, I. (2015). *Emerging Pattern Mining to Characterize Language Registers in French* (mém. de mast., Université de Tunis)

— MEKKI, J., BÉCHET, N., BATTISTELLI, D. & **LECORVÉ**, G. (2020). Caractérisation de re-

3. http://fr.wiktionary.org

gistres de langue par extraction de motifs séquentiels émergents. *Actes des Journées Internationales d'Analyse statistique des Données Textuelles (JADT)*

Dans le cadre du projet ANR TREMoLo, l'un des objectifs est de produire automatiquement un ensemble de motifs langagiers permettant de discriminer un registre par rapport à un autre. Outre le cas des registres de langue, l'un des objectifs de ces travaux est de développer une méthodologie applicable à n'importe quelle variation stylistique. Pour cela, nous travaillons sur l'emploi de techniques d'extraction de motifs séquentiels. Avant de présenter le travail réalisé, nous commençons par introduire quelques notions théoriques.

4.3.1 Définitions et concepts

L'extraction de motifs séquentiels, introduite par (AGRAWAL, SRIKANT et al. 1995), permet d'identifier des régularités dans des séries temporelles symboliques multivaluées, c'est-à-dire dans des bases de *séquences* d'ensemble de symboles. Formellement, les éléments d'une séquence sont appelés *itemsets*, notés I, et sont composés d'un ensemble de valeurs symboliques appelés *items*, soit $I = \{i_1, i_2, \ldots, i_n\}$. Une *séquence* S est alors une liste ordonnée d'itemsets, notée $\langle I_1 \ldots I_m \rangle$, la cardinalité des itemsets n'étant pas fixe. Par exemple, la séquence $\langle (a,b,c)(a,d)(a,b) \rangle$ est une séquence de trois itemsets, chacun composé respectivement de trois, deux et deux items.

Une séquence $S_1 = \langle I_1, I_2, ...I_n \rangle$ est qualifiée de sous-séquence de $S_2 = \langle I'_1, I'_2, ...I'_m \rangle$, noté $S_1 \leqslant S_2$, s'il existe des entiers $1 \leq j_1 < \ldots < j_n \leq m$ tels que $I_1 \subseteq I'_{j1}, \ldots, I_n \subseteq I'_{jn}$. Par exemple, $\langle (a)(d) \rangle \leqslant \langle (a,b,c)(a,d)(a,b) \rangle$. La tâche d'extraction de motifs séquentiels consiste alors à trouver parmi toutes les séquences d'une base l'ensemble des sous-séquences qui satisfont un certain critère, le plus souvent une fréquence minimale d'apparition. La résolution de cette tâche est un problème complexe sur le plan calculatoire en raison du nombre de sous-séquences à examiner à partir de chaque séquence. Plusieurs extensions au critère de fréquence existent pour juguler cette complexité, par exemple *via* l'addition d'un critère sur la longueur des sous-séquences extraites (YAN et AL 2003) ou sur l'éloignement maximal entre itemsets au sein de celles-ci (DONG et PEI 2007).

Enfin, lorsque deux bases de séquences sont considérées, leurs spécificités mutuelles peuvent être caractérisées par l'extraction de motifs discriminants, dits *émergents*. Pour cela, les motifs issus de chaque base sont passés au crible d'un critère d'émergence calculé comme le rapport de leur fréquence d'apparition dans les deux bases. Si ce taux d'émergence est supérieur à un seuil (*a minima*, 1), le motif est sélectionné.

4.3.2 Application aux registres

Pour extraire des motifs langagiers propres à chaque registre, nous considérons un énoncé comme une séquence de mots et chaque mot est vu comme un itemset dont les items sont des descripteurs linguistiques. Nous avons appliqué une extension de l'algorithme CloSpec (BÉCHET et al. 2015) pour extraire les motifs émergents d'un registre par rapport à un autre.

À l'image de nos travaux sur les disfluences (chapitre 3), nous avons travaillé sur des textes artificiels aux propriétés connues pour valider l'intérêt des techniques d'extraction de motifs émergents, puis évolué vers des données réelles. Dans les deux cas, nous connaissons les motifs caractéristiques à retrouver, soit grâce à la grammaire générative des textes artificielles, soit à travers nos travaux préliminaires de la section 4.1. Ainsi, nous avons cherché à savoir :

— Si les motifs que nous savons caractéristiques d'un registre sont effectivement extraits et seulement cela (à moins de trouver de nouveaux motifs dans le cas des données réelles) ;
— Si la complexité algorithmique permet le passage à de grandes quantités de données et à un grand nombre de descripteurs (taille des itemsets).

En l'état de nos travaux, chaque mot est associé à son lemme, sa catégorie grammaticale et sa fonction syntaxique. Sur les données artificielles, nous avons utilisé des métriques issues du domaine

Motif	Exemples
Motifs connus	
⟨ (*pos:auxiliaire*) , (*syntax:advmod, pos:adverbe, lemme:pas*) ⟩	• *Hé ! dis, vieux, je l'ai pas refroidie, au moins ?* • *c'est pas non plus ton frometon à toi, béby !*
⟨ (*pos:ponctuation, syntax:ponctuation*) , (*pos:ponctuation*) , (*pos:ponctuation*) ⟩	• *Et c'est 80 euros d'ailleurs (... ahahahaha)* • *ne le laissont pas filer ! ! !*
Motifs nouveaux	
⟨ (*pos:pronom, mot:se*) , (*pos:verbe*) ⟩	• *pour pas se faire chopper* • *[...] aux chinois de se magner à fabriquer [...]*
⟨ (*syntax:auxiliaire*) , (*pos:adverbe*) ⟩	• *c'est mal foutu cette affaire...* • *elle a pleuré super fort*

TABLE 4.1 – Quelques exemples de motifs discriminants du registre familier par rapport au registre soutenu.

de la recherche d'information pour mesurer la qualité du classement des motifs extraits et nous avons pu vérifier la validité de l'algorithme mis en place. Sur les données réelles, les résultats de ces expériences valident la capacité de ces algorithmes à faire extraire les motifs de la littérature ainsi que d'autres, à notre connaissance nouveaux. La table 4.1 liste quelques exemples de motifs connus et nouveaux. Ces exemples nous permettent de noter qu'un intérêt important de l'extraction de motifs est qu'elle permet de combiner différents niveaux linguistiques au sein d'un même motif. Sur le plan calculatoire, il ressort néanmoins des expériences que la complexité algorithmique est un verrou majeur. En particulier, l'augmentation du nombre de descripteurs de chaque mot induit une combinatoire que l'algorithme d'extraction subit, ce qui se traduit par des temps et besoins en mémoire irréalistes. L'un des objectifs à terme de ce travail étant sa généralisation à des styles non encore caractérisés linguistiquement, l'emploi d'un maximum de descripteurs est malheureusement une nécessité. De même, nos expériences montrent qu'il est, en l'état, difficile d'espérer obtenir des motifs d'une portée supérieure à 2 ou 3 mots. Ces limites sont bien connues dans le domaine de l'extraction de motifs séquentiels mais peu de solutions ont jusqu'à présent été proposées (RAÏSSI et PONCELET 2007). Pour progresser sur cette question, nous travaillons actuellement à la construction de versions approchées des algorithmes classiques par des mécanismes d'échantillonnage des données.

Conclusion

Dans ce chapitre, nous avons parcouru différents travaux contribuant à une modélisation de la notion de registre de langue en TALN. Nous avons présenté des travaux sur la validation en corpus de descripteurs issus de la littérature linguistique et sociolinguistique, la construction automatique d'un vaste corpus textuel annoté et l'extraction automatique de motifs discriminant. Ces différents travaux se placent dans une visée cohérente de réplicabilité à d'autres styles textuels. L'une de leur originalité est le positionnement à l'intersection de la linguistique et de la fouille de données, chacun apportant ses verrous. D'un côté, il est difficile de circonscrire les registres de langue familier, courant et soutenu comme objet d'étude par comparaison avec les autres notions de la linguistique (genre, niveau, style, mode...). De l'autre, la complexité des algorithmes de fouille classiques rend leur application difficile. Enfin, notons également que, dans le cadre du projet ANR TREMoLo, ces travaux doivent s'interfacer avec des problématiques de production automatique de paraphrases. Ces travaux ont récemment débuté et font l'objet d'une discussion dans le chapitre 8.

Chapitre 5
Textes pour les enfants

La façon dont un individu comprend un texte dépend des caractéristiques du texte et des capacités de l'individu. C'est particulièrement vrai en fonction de l'âge puisque, pendant l'enfance, les capacités cognitives, linguistiques et culturelles évoluent beaucoup. En écho à cela, il est préférable d'adapter son discours lorsque l'on s'adresse à des enfants. Sur ce principe, de multiples médias offrent aux enfants un meilleur accès aux événements de l'actualité et aux informations du web dans le respect de leur capacité cognitives. Il peut s'agir de portails dédiés (Qwant Junior en France, Yahoo ! Kids au Japon), de contenus textuels (encyclopédies Wikimini, Vikidia, Simple Wikipedia ; journaux Le P'tit Libé, Le Petit Quotidien...) ou de contenus multimédia (pastilles vidéo 1 jour 1 question, 1 jour 1 actu, déclinaisons "Kids" de Youtube et Netflix...). Malheureusement, ces initiatives reposent sur le travail d'experts et souffrent donc d'un pouvoir d'impact limité. Les techniques d'intelligence artificielle sont donc un levier important pour développer ces initiatives.

En psycholinguistique, la compréhension du langage par les enfants est un domaine largement étudié. Les travaux mettent en avant des facteurs tels que le rôle de la mémoire phonologique à court terme (Gathercole 1999), l'acquisition des notions temporelles (Hickmann 2012 ; Tartas 2010) et des constructions linguistiques qui les accompagnent (Vion et Colas 1999), ou encore la cohérence et la complexité des émotions présentes dans un récit (Blanc 2010 ; Davidson 2006 ; Mouw et al. 2019). De leur côté, les études en apprentissage de la lecture apportent des éléments d'analyse quant à la facilité que peuvent avoir les enfants à déchiffrer un texte. Frith (1985) fait valoir que la lecture est acquise en trois étapes principales liées à la reconnaisance de mots-formes, puis de graphèmes et, enfin, de morphèmes. Par la suite, la compréhension du langage s'accompagne de l'acquisition de connaissances sur la langue (construction syntaxique, vocabulaire...). Sur le plan oral, d'autres travaux notent que l'intonation lors de la lecture d'un texte – induite par le lexique, la ponctuation, la syntaxe,... – influence la perception d'un texte et que la compréhension de cette intonation évolue avec l'âge (Aguert et al. 2009). Enfin, des approches calculatoires existent depuis longtemps pour lier la lisibilité d'un texte à un niveau d'étude. Historiquement, celles-ci se fondent sur les complexités lexicale et syntaxique, à l'image de l'indice Flesch-Kincaid (Flesch 1948), ou de la formule Dale-Chall qui considère en plus la notion de mots « difficiles » (Dale et Chall 1948). Plus récemment et plus généralement en TALN, des travaux sur la simplification du texte pour les enfants (De Belder et Moens 2010 ; Gala et al. 2018) ou sur la lisibilité de textes (François 2015 ; François et Fairon 2012) ont été proposés.

Dans ce chapitre, nous présentons des travaux sur les textes adressés aux enfants. Ceux-ci s'intègrent dans le cadre de collaborations avec l'université de Paris-Nanterre, qui ont récemment donné lieu au démarrage du projet ANR TextToKids. Cet axe de recherche mèle des problématiques de TALN et de psycholinguistique. Ainsi, nous abordons dans la section 5.1 un premier modèle développé pour prédire des recommandations d'âge pour des textes et, dans la section 5.2, des études linguistiques menées sur la temporalité et les émotions dans les récits pour enfants.

5.1 Prédiction d'une recommendation d'âge

Projets : TextToKids (coordinateur – PEPS CNRS – 2018 – IRISA, MoDyCO, LLing), TextToKids (coordinateur adjoint, responsable local – ANR – 2019-2023 – MoDyCO, IRISA, Qwant, Synapse Développement, Le P'tit Libé)
Encadrements⋆/collaborations : ⋆Alexis Blandin (stage de master), ⋆Aline Étienne (stage de master, doctorat), ⋆Md Rashedur Rahman (postdoctorat), Delphine Battistelli (MoDyCo), Nicolas Béchet (IRISA), Jonathan Chevelu (IRISA)
Références :

- BLANDIN, A. (2019). *Prédiction de recommandations d'âge pour l'accès à des enfants à des textes* (mém. de mast., Université de Rennes 1)
- BLANDIN, A., **LECORVÉ**, G., BATTISTELLI, D. & ÉTIENNE, A. (2020a). Age recommendation for texts. *Proceedings of the Language Resources and Evaluation Conference (LREC)*
- BLANDIN, A., **LECORVÉ**, G., BATTISTELLI, D. & ÉTIENNE, A. (2020b). Recommendation d'âge pour des textes. *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (JEP-TALN)*

Nous avons travaillé sur une méthode de prédiction automatique des recommandations d'âge pour des textes en français dans le but de maximiser leur compréhension par des enfants. Bien que la question puisse être catégorisée comme relevant de l'analyse de la lisibilité d'un texte, la formulation sous la forme d'une prédiction d'âge nous semble nouvelle. En cela, nous avons contribué sur de multiples points. En premier lieu, nous avons listé un ensemble de descripteurs potentiellement pertinents et avons proposé différentes façons de formaliser la prédiction de l'âge comme un problème de régression. Ensuite, nous avons construit manuellement un corpus textuel annotées en âges en exploitant les recommandations fournies par les auteurs ou éditeurs de textes. Enfin, nous avons validé expérimentalement nos propositions. Notamment, nous avons montré que l'hypothèse selon laquelle toutes les phrases d'un même texte partagent la même recommandation d'âge est un simplification acceptable du problème. Par ailleurs, nous avons montré que nos modèles étaient acceptables car la marge d'erreur de leurs recommandations semble meilleure que celle d'experts psycholinguistes.

Dans la suite, nous revenons sur les aspects liés à la modélisation, puis donnons davantage de détails sur les expérimentations.

5.1.1 Modélisation

En nous appuyant sur la littérature linguistique et psycholinguistique, nous avons listé des descripteurs portant sur de multiples niveaux d'abstraction. Leur liste exhaustive est la suivante :

- **Lexique (5 descripteurs) :** log-probabilité des mots en français ; diversité des mots/lemmes.
- **Graphie/typographie (6) :** Score de confusion graphique des mots [1] ; longueur des mots ; ratio de caractères par mot ; ratio de ponctuations par mot.
- **Morphosyntaxe (7) :** classes grammaticales ; mots-outils.
- **Temps verbaux (24)** : diversité des temps verbaux ; proportions de 14 temps (simples et composés) ; modes ; systèmes temporels (passé, présent, futur).
- **Personne et forme verbale (5) :** proportion des première/deuxième/troisième personnes ; proportion des formes singulier/pluriel.
- **Syntaxe (8) :** mots par phrase ; distances moy. et max. entre un mot et ses dépendances syntaxiques ; nombre de dépendances entrantes/sortantes par mot ; profondeur de l'arbre de dépendances.
- **Connecteurs logiques (16) :** addition ; temps ; but ; cause ; comparaison ; concession ; conclusion ; condition ; conséquence ; énumération ; explicat. ; illustrat. ; justificat. ; opposit. ; res-

1. Inspiré de (GEYER 1977).

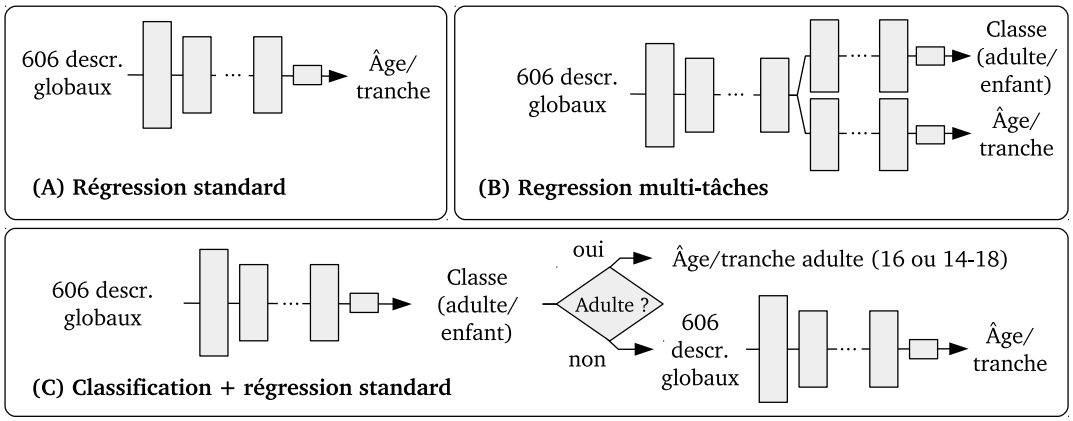


FIGURE 5.1 – Architectures des approches étudiées.

trict. ; exclusion.

— **Phonétique (9) :** longueur de la phrase en phonèmes ; nombre de phonèmes par mot ; diversité des phonèmes dans le texte / dans les mots ; scores d'ordinarité phonétique [2].

— **Sentiments/émotions (26) :** Scores de subjectivité et de polarité ; mots identifiés comme déclencheurs d'une parmi 24 émotions [3].

En parallèle, comme souvent en TALN désormais, nous considérons aussi des plongements de mots :

— **Plongements (1 descripteur de dimension 500) :** plongement (*skip-gramme*) moyen des mots du texte.

Dans nos modèles, nous considérons l'âge soit comme une valeur réelle unique soit comme une tranche d'âges (un âge bas et un âge haut, comme le font souvent les auteurs et éditeurs). La tranche correspondant à l'âge « adulte » est fixée à 14-18 ans. La figure 5.1 présente les 3 architectures de modèles neuronaux que nous avons étudiées. Toutes sont de type *feed-forward*, s'appuyant sur un vecteur d'entrée de dimension 606 et produisant soit un âge soit une tranche d'âges recommandé. Le modèle A est un modèle de régression standard. Le modèle B est un modèle multi-tâches où la prédiction de l'âge est augmentée d'une classification binaire adulte/enfant. Enfin, le modèle C enchaîne un classifieur et un modèle de régression si la classe prédite est « enfants ». L'idée est que la régression est inutile pour les textes considérés comme « adultes » car l'âge associé est fixe. Le modèle de régression est le même que A mais estimé sur un ensemble d'apprentissage restreint aux seuls textes pour enfants.

5.1.2 Expérimentations

Nous avons collecté un ensemble de 632 textes, dont 543 sont destinés aux enfants de 0 à 14 ans, les 89 autres étant pour les adultes. Les textes pour enfants proviennent de contes, romans, magazines et journaux. Ces textes sont annotés avec les indications des éditeurs ou des auteurs sous la forme d'une tranche d'âge A-B que nous interprétons comme les âges bas et haut à partir desquels le texte peut être compris. Les textes pour adultes sont d'un niveau difficile pour des enfants, par exemple des romans avec un langage soutenu, des articles Wikipedia et de journaux sur des sujets avancés (capitalisme, génétique, diplomatie...). Pour augmenter le nombre d'exemples d'apprentissage, les textes sont découpés en phrases, chacune partageant la même annotation que son texte d'origine. Il s'agit d'une hypothèse forte mais nos résultats montrent qu'elle est fonctionnelle. Le corpus ainsi obtenu est composé d'environ 30 000 phrases et 450 000 mots.

2. Calculé comme la probabilité moyenne de chaque phonème en français, comme indiqué dans (GROMER et M. WEISS 1990).

3. Les mots et les émotions sont issus d'un raffinement du dictionnaire EMOTAIX (PIOLAT et BANNOUR 2009).

Les grandes conclusions de nos expérience sont les suivantes :

— Les différences entre les modèles A, B et C sont non significatives. L'approche C tend néanmoins à être la plus mauvaise. De même, la prédiction directe d'un âge ou le recours intermédiaire à une tranche d'âge conduisent à des résultats proches.
— Les modèles atteignent une erreur d'estimation de l'âge recommandé d'environ 2,5 ans, ce qui est nettement meilleur qu'une prédiction naïve fondée sur la moyenne observée dans notre corpus (erreur d'environ 4,5 ans).
— La fusion des prédictions au niveau des phrases d'un texte permet d'affiner la prédiction au niveau du texte.
— Les recommandations de nos modèles semblent plus précises que celles de psycholinguistes (erreur d'environ 3 ans).
— Les plongements lexicaux sont les descripteurs les plus utiles pour les prédictions. L'ensemble des autres descripteurs apportent néanmoins un gain léger.

Ces conclusions positives sont à relativiser par différents biais qu'il est difficile de lever en l'état. Tout d'abord, notre vérité terrain s'appuie sur les recommandations des auteurs et éditeurs. Or, leurs précision et bien-fondé scientifique sont sans doute parfois discutables car probablement sujets à des impératifs autres (thématiques, commerciaux...). Par ailleurs, l'analyse des prédictions faites par les psycholinguistes montrent qu'ils ont tendance à utiliser un intervalle plus restreint que celui autorisé (4-13 contre 0-18 dans nos données). Ceci est probablement lié à l'interprétation de la notion de compréhension. Le projet dans lequel ce travail s'insère prévoit une campagne d'évaluation auprès d'enfants (en classe) pour consolider ces analyses.

5.2 Temporalité et émotions dans les textes pour enfants

Projets : TextToKids (coordinateur – PEPS CNRS – 2018 – IRISA, MoDyCO, LLing), TextToKids (coordinateur adjoint, responsable local – ANR – 2019-2023 – MoDyCO, IRISA, Qwant, Synapse Développement, Le P'tit Libé)
Encadrements⋆/collaborations : ⋆Charlotte Bourgoin (stage de master), ⋆Aline Étienne (stage de master, doctorat), Delphine Battistelli (MoDyCo)
Références :

— BOURGOIN, C., BATTISTELLI, D. & **LECORVÉ**, G. (2018). *Les notions temporelles dans la mise en récit d'événements dans le discours journalistique enfantin*. Rapport de recherche. IRISA, MoDyCo
— ÉTIENNE, A. (2019). *Compréhension de textes par les enfants et émotions : point(s) de vue psycholinguistique(s) et leur mise en œuvre en TAL* (mém. de mast., Université de Paris-Nanterre)
— ÉTIENNE, A., BATTISTELLI, D. & **LECORVÉ**, G. (2020a). Apports de la linguistique et du TAL à l'analyse des émotions dans les textes pour enfants. *Actes de colloque Langage et éMOTions*
— ÉTIENNE, A., BATTISTELLI, D. & **LECORVÉ**, G. (2020b). L'expression des émotions dans les textes pour enfants : constitution d'un corpus annoté. *Actes de la Conférences sur le Traitement Automatique du Langage Naturel (TALN)*

Sur le plan linguistique, nous avons conduit deux études sur des articles tirés du journal Le P'tit Libé, dédié aux enfant de 7 à 12 ans et avec qui nous collaborons. La première étude porte sur la temporalité, la seconde sur les émotions. Leur objectif est la confrontation de la littérature sur la compréhension de ces dimensions avec la pratique d'auteurs telle qu'illustrée par ces articles. Par là, l'objectif à terme est l'identification de pratiques recommandables ou problématiques.

5.2.1 Temporalité

Nous avons étudié la mise en récit d'événements dans le discours journalistique à destination d'enfants à travers l'analyse des trois composantes fondamentales des notions temporelles et de la mise en récit : les temps verbaux ; les connecteurs temporels (*avant*, *après*, *puis*, *ensuite*, *alors*, *quand*, *pendant* et *et*) ; et les adverbiaux temporels.

Les résultats de cette analyse ont démontré un emploi diversifié. Si l'utilisation du système de temps du récit (passé composé / imparfait) est comparable à celles des enfants en général (BRONCKART et BOURDIN 1993), celle des compléments adverbiaux et des connecteurs temporels tend plutôt à se rapprocher de celle des enfants de 9 ans et plus. En effet, suivant les conclusions de (FAVART 2005), l'utilisation de diverses formes des connecteurs temporels (par exemple, *avant*, *avant que*, *avant tout*, *avant de*) peut provoquer, dans une certaine mesure, un allongement de leur temps de compréhension. De plus, l'analyse des types de repères utilisés dans les articles du P'tit Libé montre que les repères absolus, qui reposent sur le temps conventionnel et calendaire correspondant à l'utilisation de dates, ou heures, sont les plus fréquents et représentent $42\,\%$ de l'ensemble des occurrences. Ainsi, la notion de temps n'atteignant le stade de la généralisation qu'à environ 10-11 ans (DROIT-VOLET 2000 ; QUARTIER 2008), l'utilisation des repères temporels absolus dans le P'tit Libé est sans doute difficile pour une partie de leur public.

Certains mécanismes comme l'apposition de marqueurs explicitent les repères absolus ou la simplification lexicale et permettent de rendre plus accessibles les articles aux enfants de 7 à 9 ans. Ce souci de clarification étant cependant une caractéristique partagée par le discours journalistique destiné aux adultes, nous pouvons nous demander si cette stratégie de clarification est typique du discours journalistique enfantin ou du discours journalistique en général. Une perspective de recherche serait donc de comparer l'utilisation d'adverbiaux apposés dans les deux types de discours. De plus, les marqueurs temporels ayant parfois été analysés en tant que marques de cohésion semblables aux pronoms et aux références anaphoriques (CROWHURST 1987), il serait pertinent d'étudier l'utilisation de ces derniers afin de poursuivre la caractérisation du discours journalistique enfantin.

5.2.2 Émotions

L'objectif de ce travail était de trouver des critères pour caractériser l'expression linguistique des émotions. Idéalement, ces critères devraient être implémentables dans une chaîne de reconnaissance automatique des émotions.

Afin d'étudier la dimension émotionnelle de nos textes, nous avons retenu de la littérature psycholinguistique deux types d'émotions (de base et complexe) et dix catégories émotionnelles (colère, dégoût, joie, peur, surprise, tristesse, culpabilité, embarras, fierté et jalousie) utilisés dans les recherches sur la compréhension de textes par les enfants. Nous avons également constitué notre propre typologie des modes d'expression linguistique des émotions en combinant celle proposée en psycholinguistique par (BLANC 2010 ; CREISSEN et BLANC 2017) et celle proposée en linguistique par (MICHELI 2014). Notre typologie comporte quatre modes d'expression des émotions : les émotions désignées (termes du lexique émotionnel), les émotions comportementales (descriptions de manifestations physiques et comportementales des émotions), les émotions montrées (caractéristiques des énoncés indiquant l'état émotionnel de l'énonciateur) et les émotions étayées (description d'un situation conventionnellement associée à une émotion).

Dans un premier temps, l'analyse linguistique fine d'exemples issus du P'tit Libé a confirmé la présence dans le corpus, et donc la pertinence, des types d'émotions, catégories émotionnelles et modes d'expression des émotions que nous avions retenus. En illustrant différents critères hors-lexique émotionnel, ces premiers exemples ont permis de formuler plusieurs hypothèses pour l'analyse automatique des émotions. Par exemple, certaines structures syntaxiques typiques des émotions montrées (dislocations, énoncés averbaux…) seraient reconnaissables automatiquement par un analyseur syntaxique. Mais encore, une collection de termes pourrait être constituée pour

le repérage des émotions comportementales (par exemple « sourire », « crier »...).

L'analyse linguistique a ensuite été systématisée grâce à l'élaboration d'un schéma d'annotation intégrant les différentes catégories et les critères mis en œuvre lors de la première étape d'analyse. Son application sur notre corpus via la plateforme Glozz (WIDLÖCHER et MATHET 2009) a conduit au repérage de 2 043 unités exprimant des émotions dans les textes du P'tit Libé. Malgré l'aspect théoriquement plus « descriptif » et « explicatif » des textes journalistiques, le nombre d'unités émotionnelles annotées dans les textes du P'tit Libé laisse supposer que les émotions sont en fait très présentes. Au niveau lexical, en dehors des termes du lexique des émotions, certains mots indiqueraient la présence de passages émotionnels : les termes décrivant une manifestation physique ou comportementale d'une émotion (par exemple « crier », « pleurer », « manifester contre »...) ; ceux entrant dans la description d'une situation conventionnellement associée à une émotion (par exemple « guerre », « harceler », « gagner »...) ; et enfin les interjections (par exemple « ouf », « ah »...) et certains adverbes (par exemple « même », « décidément »...). Certaines structures syntaxiques sont aussi mobilisées pour exprimer les émotions, comme les dislocations, les énoncés averbaux ou les exclamations. Certains critères relèvent du niveau discursif, plus difficile à analyser automatiquement, notamment lorsque la description d'une situation à même de suggérer une émotion est réalisée à l'échelle de plusieurs phrases, d'un paragraphe entier ou de l'ensemble du document. Par ailleurs, l'observation quantitative des unités d'émotion de notre corpus suggère l'utilisation privilégiée de certains modes d'expression pour réaliser certaines émotions. Les émotions montrées expriment en effet majoritairement de la surprise, de la joie et de la colère, tandis que l'expression comportementale est le plus souvent associée à de la colère, de la peur et de la tristesse. Les émotions désignées et étayées traduisent quant à elles principalement la peur, la colère et la joie, ce qui correspond aux trois catégories émotionnelles les plus observées sur l'ensemble du corpus. Enfin, l'observation linguistique menée sur le corpus du P'tit Libé nous semble souligner l'importance de la notion de prise en charge énonciative pour l'étude de l'expression des émotions.

Conclusion

Ce chapitre a présenté mes activités de recherches liées à l'analyse de textes à destination des enfants. Nous avons abordé un premier travail sur la prédiction de recommandation d'âge. Il s'agit d'une tâche singulière dans la littérature pour laquelle ces premiers résultats sont très encourageants. D'autres travaux se sont concentrés sur les particularités linguistiques de la temporalité et des émotions dans des textes journalistiques pour enfants. Ces travaux complètent les premiers et dans le sens où ils ouvre la possibilité à l'avenir d'expliquer en langage naturel à des auteurs des éléments inadéquats de leur texte en cours de rédaction. Au delà, de nombreux travaux sont encore à prévoir pour affiner les analyses linguistiques, les généraliser à d'autres genres et dimensions. De plus, les modèles de prédiction présentés ici sont encore rudimentaires et nous en expérimentons actuellement des plus élaborés.

CHAPITRE 6

Approches neuronales pour l'encodage de l'information linguistique

En une décennie de travaux en TALN, les réseaux de neurones sont passés du statut d'outil de mise en œuvre à objet central des avancées proposées. Ce changement a considérablement modifié l'approche des problèmes. La tendance actuelle, souvent appuyée par les résultats, consiste à l'agnostisme quant à la nature des données et aux propriétés des phénomènes à traiter. Il est ainsi relativement similaire de construire un système de synthèse de la parole ou un modèle de simplification de textes. L'expertise linguistique y est remplacée par une expertise sur les techniques d'apprentissage profond, nécessaire pour comprendre et entraîner les architectures parfois complexes de l'état de l'art.

L'apprentissage profond a également fait émerger un nouvel usage des réseaux de neurones : celui d'encoder l'information grâce au concept de plongement. Cet encodage a l'intérêt de convertir des ensembles d'informations hétérogènes (symboliques ou numériques, de multiples dimensions, séquentielles ou non...) en des représentations latentes homogènes dans d'un espace vectoriel de grande dimension et d'ouvrir la voie aux outils mathématiques de ces espaces (distances, transformations...). L'exemple le plus marquant en TALN est probablement la méthode Word2Vec (MIKOLOV, SUTSKEVER et al. 2013) qui a introduit une nouvelle représentation des mots telle que des propriétés linguistiques (notamment sémantiques) se retrouvent transposer en propriétés géométriques. Ensuite, par leur forme homogène (des vecteurs de réels de taille fixe), les plongements permettent d'interfacer simplement ensemble de nombreux modules neuronaux, quitte même à les apprendre en même temps, à l'image des architectures encodeur-décodeur déjà abordées dans ce manuscrit.

Ce chapitre présente différents travaux menés ces dernières années sur l'encodage de l'information linguistique. À chaque fois, il s'agit de comprendre les propriétés des plongements de certains objets linguistiques donnés pour les exploiter en fonction de la tâche traitée. Ainsi, la section 6.1 présente une contribution sur les modèles de langue par réseaux de neurones récurrents. La section 6.2 aborde l'utilisation des plongements pour la synthèse de la parole. Enfin, la section 6.3 présente des travaux sur le transfert de style appliqué à des textes.

6.1 Discrétisation de modèles de langage neuronaux

Projets : TAO CSR (CTI, Suisse – 2011-2012 – IDIAP, Koemei)
Collaborations : Petr Motlicek (IDIAP)
Références :

— **LECORVÉ**, G. & MOTLICEK, P. (2012). Conversion of recurrent neural network language

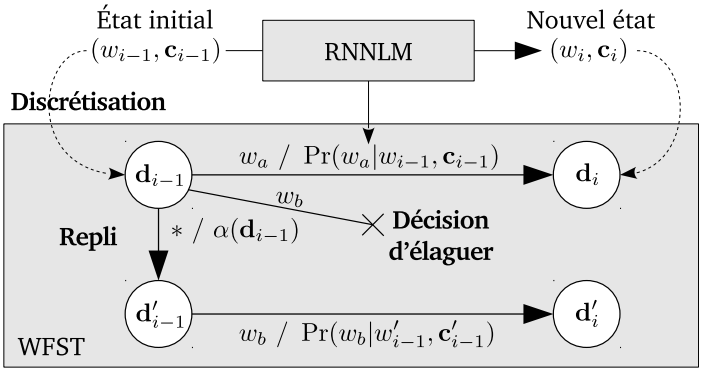


FIGURE 6.1 – Vue d'ensemble de la conversion d'un RNNLM en un WFST.

models to weighted finite state transducers for automatic speech recognition. *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*
Logiciel : https://github.com/glecorve/rnnlm2wfst

Les modèles de langage par réseau de neurones récurrents (*Recurrent Neural Network Language Model*, RNNLM) ont été popularisés par MIKOLOV, KARAFIAT et al. (2010). Ils marquent un tournant important en modélisation du langage, longtemps dominé par les vénérables modèles *n*-grammes. En reconnaissance automatique de la parole, les RNNLM sont difficiles à utiliser directement pour décoder un signal de parole [1] car ils reposent sur des représentations continues de l'historique des mots tandis que les algorithmes de décodage nécessitent de gérer des représentations discrètes (MOHRI et al. 2008 ; NEY et ORTMANNS 1999). Ainsi, les RNNLM sont souvent utilisés en post-traitement pour réévaluer des listes d'hypothèses décodées par un modèle *n*-gramme. Par conséquent, le pouvoir de prédiction des RNNLM n'est utilisé que sur des sous-ensembles de toutes les hypothèses de transcription. Dans (**LECORVÉ** et MOTLICEK 2012), nous avons défini une nouvelle stratégie pour transformer les RNNLM en un transducteur à états fini pondéré (*Weighted Finite State Transducer*, WFST) qui peut être directement utilisé dans le processus de décodage d'un système de reconnaissance automatique de la parole (MOHRI et al. 2008).

Précisément, le but d'un modèle de langage employé dans le système de reconnaissance est de fournir la probabilité conditionnelle d'un mot w_i étant donné un historique h des mots précédents. Comme détaillé dans (MIKOLOV, KARAFIAT et al. 2010), cet historique est représentée dans les RNNLM par le mot précédent le plus récent w_{i-1} et un plongement $\mathbf{c}_{i-1}$ du contexte accumulé jusque là. La topologie du réseau neuronal utilisé pour calculer les probabilités conditionnelles $\Pr(w_i|w_{i-1}, \mathbf{c}_{i-1})$ est organisée en plusieurs couches. La couche d'entrée lit un mot w_{i-1} et un historique continu $\mathbf{c}_{i-1}$. Les couches cachées compressent les informations de ces deux entrées et calculent une nouvelle représentation $\mathbf{c}_i$. La valeur $\mathbf{c}_i$ est ensuite passée à la couche de sortie qui, après normalisation, fournit la probabilité conditionnelle du mot w_i.

Comme illustré dans la figure 6.1, la méthode proposée de conversion d'un RNNLM en un WFST consiste principalement à lier des états discrets $\mathbf{d}_k$ aux états d'entrée $(w_k, \mathbf{c}_k)$ du RNNLM et à utiliser ces états discrets comme nœuds d'un WFST. Les arcs entre les nœuds sont ensuite étiquetés avec des transitions de mots et leurs probabilités estimées par le RNNLM. Il y a deux aspects-clés pour accomplir cette tâche. Premièrement, une fonction de discrétisation doit être définie pour transformer les représentations continues. Deuxièmement, la taille du WFST construit doit être maîtrisée car l'énumération de tous les états discrets possibles peut rapidement être impossible à traiter dès que le vocabulaire devient grand et que la discrétisation devient précise. En

1. Sauf à utiliser une architecture neuronale de bout en bout qui intègre aussi l'équivalent de l'extraction des descripteurs, du modèle acoustique et du lexique phonétisé, comme cela a été proposé depuis que mes travaux ont été menés.

conséquence, un critère d'élagage doit être défini afin d'éliminer les états discrets non intéressants, ainsi qu'une stratégie de repli des états élagués $\mathbf{d}_k$ vers d'autres plus simples $\mathbf{d}'_k$.

Nous avons décliné les propriétés que doivent satisfaire ces différents éléments et avons proposé une première implémentation de la méthode de conversion. Celle-ci s'appuie sur l'algorithme des *K-means* pour discrétiser l'espace continu des plongements. Ce choix exploite l'idée que des plongements proches représentent des contextes linguistiques également proches. Quant à l'élagage, il s'appuie sur un critère entropique sur la distribution de probabilités du WFST en cours de construction et ramène les discrétisations non retenues (élaguées) vers le barycentre globale de l'espace continu.

Deux séries d'expériences ont été menées pour évaluer l'approche proposée : des expériences sur le corpus de Penn Treebank pour étudier le comportement du processus de conversion et des expériences sur des enregistrements de parole pour vérifier la validité du WFST généré dans un système de reconnaissance de la parole à l'état de l'art. Les résultats de ces expériences ont respectivement montré la viabilité de la proposition sur le plan calculatoire, puis la compétitivité de l'implémentation proposée par rapport à un modèle de langage *n*-gramme pour le décodage.

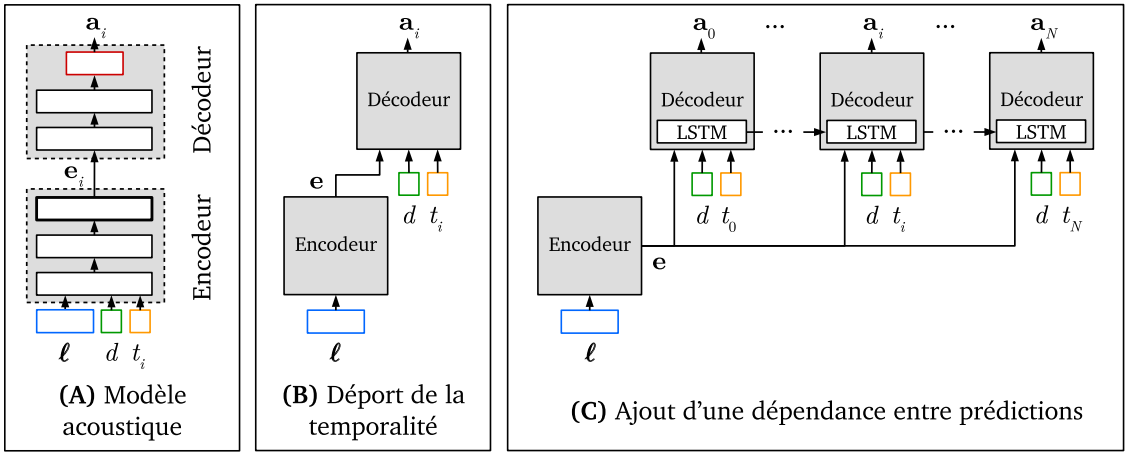
6.2 Plongements de phonèmes pour la synthèse de la parole

Projets : SynPaFlex (ANR – 2015-2019 – IRISA, LLF, ATILF)
Encadrements⋆/collaborations : ⋆Antoine Perquin (stage de master, doctorat), ⋆David Guennec (ingénieur de recherche), Laurent Amsaleg (IRISA), Damien Lolive (IRISA)
Références :

- PERQUIN, A. (2017). *Big deep voice : indexation de données massives de parole grâce à des réseaux de neurones profonds* (mém. de mast., University of Rennes 1)
- PERQUIN, A., **LECORVÉ**, G., LOLIVE, D. & AMSALEG, L. (2018). Phone-Level Embeddings for Unit Selection Speech Synthesis. *Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP)*. Springer
- ALAIN, P., **LECORVÉ**, G., LOLIVE, D. & PERQUIN, A. (2018). The IRISA Text-To-Speech System for the Blizzard Challenge 2018. *Proceedings of the Blizzard Challenge 2018 Workshop*
- PERQUIN, A., **LECORVÉ**, G., LOLIVE, D. & AMSALEG, L. (2019). Évaluation objective de plongements pour la synthèse de parole guidée par réseaux de neurones. *Actes de la conférences sur le Traitement automatique du langage naturel (TALN)*

En synthèse de la parole, la méthode historique dite par sélection d'unités consiste à transformer un texte en entrée en une séquence de phonèmes, puis à concaténer des unités de paroles pré-enregistrées correspondant à ces phonèmes pour obtenir un signal de sortie (HUNT et BLACK 1996). Les unités à concaténer sont choisies au sein d'une base de données telles qu'elles minimisent un coût de sélection indiquant à quel point l'unité s'éloigne de l'idéal demandé par le texte. Pour cela, tout phonème (de la base de parole ou dérivé du texte) est représenté par un vecteur d'informations linguistiques, par exemple, son identité, celle de ses proches voisins, sa position dans la syllabe, le mot ou l'énoncé auquel il appartient, etc. Ainsi, le coût de sélection est une distance construite entre deux vecteurs de descripteurs. Ce coût est habituellement défini par des experts et est fastidieux à régler. À l'image d'autres travaux à la même période (MERRITT et al. 2016 ; WAN et al. 2017), nous avons proposé de remplacer le coût de sélection expert par une distance euclidienne entre plongements (PERQUIN 2017 ; PERQUIN, **LECORVÉ** et al. 2018). Pour cela, nous cherchons à représenter chaque phonème comme un plongement tel que la distance entre deux plongements reflète la distance acoustique entre les réalisations théoriques de leur phonème sous-jacent.

Comme illustré par le modèle A de la figure 6.2, la construction des plongements reprend le principe des modèles acoustiques tel que celui proposé dans (WU et KING 2016). Pour un



(A) Modèle acoustique

(B) Déport de la temporalité

(C) Ajout d'une dépendance entre prédictions

FIGURE 6.2 – Architecture d'un modèle acoustique (A) et deux modèles plongements de phonèmes (B et C).

phonème donné représenté par son vecteur linguistique $\boldsymbol{\ell}$, un tel modèle prédit les coefficients acoustiques $\mathbf{a}_i$ de la i-ème trame d'un phone correspondant [2]. Des informations de synchronisation sont adjointes au modèle pour prendre en compte la dynamique des coefficients acoustiques. Celles-ci sont la durée d du phone et la position relative $t_i \in [0, 1]$ de la i-ème trame à l'intérieur du phone. Ce modèle permet de fournir des plongement de trames $\mathbf{e}_i$ comme la sortie de l'une des couches cachées. Notons que le réseau de neurones s'explique alors plus clairement comme l'enchainement d'un encodeur et d'un décodeur. Pour construire des plongements de *phonèmes*, et non de trames, nous avons déporté les informations temporelles après l'encodeur (modèle B). Une fois le modèle appris, seul l'encodeur est utile au moment de la synthèse. Nous avons également étudié l'introduction d'une dépendance temporelle à l'apprentissage par l'utilisation d'une couche LSTM dans le décodeur (modèle C). Celle-ci permet alors de prédire la séquence entière des coefficients $[\mathbf{a}_i]_{i \in [\![0,N]\!]}$ pour un phonème en entrée. Enfin, nous avons aussi étudié l'impact de la dimension du plongement *via* une topologie en goulot d'étranglement (non représentée sur le schéma).

Nous avons expérimenté ces méthodes sur deux corpus de parole en français, l'un de style neutre, l'autre plus expressif (livre audio). Les modèles ont été évalués pour la tâche de modélisation acoustique (comparaison des trames prédites) et pour la synthèse (test perceptif). Les expériences ont mis en évidence que l'intégration tardive du temps et de l'information n'influe pas sur la qualité de la prédiction des coefficients acoustiques, même si l'utilisation du LTSM ne semble pas encore concluante. Ensuite, les expériences sur la synthèse par sélection d'unités ont montré que le remplacement du coût de sélection expert par celui fondé sur les plongements se traduit par une amélioration de la qualité perçue par les évaluateurs. En outre, les plongements proposés semblent bien se généraliser d'une voix de synthèse à l'autre. La solution a ainsi été intégrée dans le système que nous avons proposé à l'édition 2018 du challenge international de synthèse de la parole *Blizzard Challenge* (ALAIN, **LECORVÉ** et al. 2018).

Pour progresser dans l'élaboration de modèles de plongement, nous avons également étudié une méthodologie pour déterminer la qualité d'un plongement (PERQUIN, **LECORVÉ** et al. 2019). Nous avons proposé une méthode générique consistant à comparer un plongement dont on cherche à évaluer la qualité avec d'autres issus d'entraînement volontairement défavorables en guise de bornes basses, par exemple des cas de sur- et sous-apprentissage, voire des plongements aléatoires. Nous avons montré que la comparaison visuelle de ces plongements par des techniques de compression de dimension permet d'obtenir de premiers critères distinctifs et que des mesures objec-

2. Un phone est la réalisation acoustique d'un phonème.

tives correspondant à ces critères rend possible des comparaisons entre plongements quelconques. Appliqué à la synthèse de la parole, nous avons montré que les mesures objectives corroborent les résultats de nos tests perceptifs.

6.3 Transfert de style par réseaux de neurones

Projets : TREMoLo (coordinateur – ANR – 2016-2021 – IRISA, MoDyCo)
Encadrements⋆/collaborations : ⋆Somayeh Jafaritazehjani (doctorat), ⋆Nazanin Dehghani (postdoctorat), John D. Kelleher (Technical University of Dublin, ADAPT Research Centre), Jonathan Chevelu (IRISA), Damien Lolive (IRISA)
Références :
— Travail en cours

Le transfert de style peut être défini comme une tâche de production de langage naturel où une séquence de mots d'entrée (par exemple, un texte, une phrase, un groupe de souffle...) d'un certain style est reformulée afin d'imiter un autre style voulu. Il s'agit d'un problème multi-objectifs où l'adoption du style cible est contrebalancée par les besoins de produire une séquence de sortie linguistiquement correcte et de conserver le sens d'origine.

Nous travaillons actuellement sur ce sujet avec l'optique de développer des approches génériques. En cela, nous adoptons ici une définition du style volontairement très large (et éventuellement contestable) qui laisse émerger le style de la confrontation entre deux corpus textuels. Ainsi, le style est l'ensemble des traits linguistiques qui diffèrent avec constance entre chaque corpus. Par opposition, les aspects communs, invariants, à ces deux distributions du langage peuvent être abusivement désignés par le terme *sens*. Bien que ces travaux ne soient pas encore publiés, je les mentionne dans ce manuscrit car ils reflètent un aspect important des perspectives que je souhaite soutenir pour l'avenir.

Dans ces travaux, nous étudions la tâche de transfert de style de manière non supervisée, c'est-à-dire que les énoncés représentatifs des styles source et cible n'ont pas à être alignés. Précisément, nous travaillons sur des extensions du modèle proposé dans (T. SHEN et al. 2017). Ce modèle, représenté par la figure 6.3, a une sur architecture encodeur-décodeur antagoniste. En supposant deux styles, notés 1 et 2, le modèle de transfert de style repose sur 4 blocs :

— Un encodeur E qui lit une séquence de mots $\mathbf{x}$ du style $s \in \{1, 2\}$, noté $\mathbf{x}^{(\mathbf{s})}$, et produit un plongement $\mathbf{z}$.
— Un générateur G qui, outre $\mathbf{z}$, prend en entrée la consigne du style désiré en sortie. En l'occurrence, à l'apprentissage, le générateur est utilisé deux fois pour produire une nouvelle séquence de chaque style $\widetilde{\mathbf{x}}^{(1)}$ et $\widetilde{\mathbf{x}}^{(2)}$.
— Deux discriminateurs, un pour chaque style, D_1 et D_2. Chaque discriminateur du style $t \in \{1, 2\}$ lit une séquence générée $\widetilde{\mathbf{x}}^{(t)}$ et prédit la probabilité que t soit le style initial s ou qu'il y ait eu un transfert. Ainsi, chaque discriminateur est un classifieur binaire sur les classes « style conservé » et « style transféré ».

Cette architecture est apprise de manière antagoniste par l'entremise de fonctions de perte spécifiques et de rétropropagations conduites en parallèle. Le générateur G est ainsi incité à produire des séquence capables de duper les discriminateur D_1 et D_2 alors que chaque discriminateur se

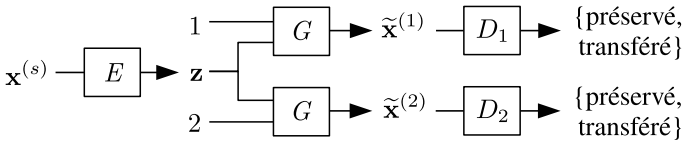


FIGURE 6.3 – Modèle de base introduit par (T. SHEN et al. 2017)

doit de distinguer les cas de transfert du mieux possible. Par ailleurs, dans le cas où la consigne de style en entrée de G est le style d'origine de l'énoncé (s), le générateur est évalué sur sa capacité à reconstuire la séquence d'origine, c'est-à-dire $\widetilde{\mathbf{x}}^{(s)} = \mathbf{x}^{(s)}$.

Nos travaux actuels visent à comprendre comment et par quels modules le sens et le style sont traités dans une telle architecture. D'une part, s'agissant d'une tâche de génération de langage naturel, cela nous amène à réfléchir à des méthodes d'évaluation objective dont l'absence freine le développement de nouvelles solutions. Ainsi, nous étudions des possibilités d'évaluation automatique sur la base de 3 dimensions jugées nécessaires : (1) la qualité du transfert de style (« est-ce que le style de $\widetilde{\mathbf{x}}^{(t)}$ est bien t ? ») ; (2) la conservation du sens de la phrase d'origine (« est-ce que $\widetilde{\mathbf{x}}^{(s)}$ et $\widetilde{\mathbf{x}}^{(t)}$ partagent bien la même trame sémantique ? ») ; (3) de la fluidité de la phrase dans la langue (« est-ce que $\widetilde{\mathbf{x}}^{(t)}$ est correcte dans la langue ? »). D'autre part, nous étudions différentes variantes du modèle afin de déterminer le rôle de chaque élément dans le résultat final. Ainsi, nos travaux se portent sur :

— le remplacement d'un encodage déterministe par un encodage variationnel (HU et al. 2017 ; JOHN et al. 2018), censé mieux généraliser des données lues en entrée ;
— la comparaison avec un encodage reconnu pour sa préservation du sens, en l'occurence ELMo (PETERS et al. 2018) ;
— le renforcement du conditionnement par le plongement $\mathbf{z}$ sur la phase de génération ;
— le remplacement d'un unique générateur conditionné par une consigne de style par deux générateurs respectivement dédiés à chaque style.

Ces différents travaux sont expérimentés sur une tâche de transfert d'opinion dans des commentaires d'utilisateurs (corpus Yelp et Amazon), où la polarité de l'opinion est, dans la définition que nous avons posée, définie comme le « style ». Par exemple, la phrase « *Nous avons été bien accueillis et le reste des prestations était très agréables.* » pourrait être transférée vers la phrase « *L'accueil était mauvais, tout comme le reste.* ». En l'état, il ressort de ces travaux qu'il est difficile d'améliorer une dimension sans dégrader l'une des deux autres. Chaque variation de l'architecture produit ses effets propres. Par exemple, l'utilisation d'ELMo comme encodeur stimule la préservation du sens mais dégrade totalement la qualité finale du style. À l'inverse, un encodage variationnel atténue l'information sémantique dans $\mathbf{z}$ et conduit à des énoncés excellents sur le plan du style mais incohérent avec la phrase d'origine. Ainsi, les prochaines étapes sont de comprendre comment concilier au mieux ces différents aspects. L'une des pistes privilégiées pour cela est l'intégration de nos mesures d'évaluation objective des dimensions étudiées comme des fonctions de pertes dans l'apprentissage antagoniste.

Conclusion

Dans ce chapitre, nous nous sommes intéressés à la manière dont les réseaux de neurones permettent d'encoder l'information linguistique et à l'utilisation qui peut être faite des plongements qui en découlent. Nous avons tout d'abord étudié cela à travers la problématique des modèles de langage en reconnaissance automatique de la parole, au sein desquels les plongements intègrent des informations morphosyntaxiques et sémantiques à partir d'historiques de mots. Pour la tâche de synthèse de la parole, nous avons présenté une contribution où les plongements visent à fusionner des informations d'ordres linguistique et acoustique. Enfin, nous travaillons actuellement sur des méthodes génériques de transfert du style au sein duquel l'une des questions est de savoir comment les composantes assimilées au style et au sens sont encodées. Ces travaux montrent, s'il était encore besoin, l'intérêt formidable du concept de plongement. Ils soulignent néanmoins également la limite selon laquelle il est difficile d'analyser la qualité d'un plongement ou d'en comparer plusieurs, sans passer par de lourdes évaluations du système entier qui les intègre.

CHAPITRE 7

Outils et projets appliqués

Les précédents chapitres ont mis en avant des contributions sur divers problèmes de recherche. Néanmoins, le domaine du TALN implique de nombreuses manipulations de données et interactions avec l'humain pour lesquelles le développement d'outils logiciels est nécessaire. Il donne également lieu à de nombreux projets appliqués visant à faciliter ou améliorer le quotidien. Bien que ces projets n'impliquent pas nécessairement de contributions scientifiques en TALN, ils œuvrent pour la diffusion et le transfert des connaissances.

Ces aspects représentent un temps conséquent de mes activités de chercheur, notamment depuis ces deux dernières années. Ce chapitre vise donc à les mettre à l'honneur. La section 7.1 présentent des outils développés comme support à mes travaux et la section 7.2 présente des projets de recherche appliquée, axés sur mes activités en synthèse de la parole, auquel je participe ou ai participé et dont j'assure pour certains une responsabilité.

7.1 Outils

7.1.1 Gestion de données multi-niveaux

Collaborations : Jonathan Chevelu (IRISA), Sébastien Le Maguer (IRISA), Damien Lolive (IRISA)
Références :

- CHEVELU, J., **LECORVÉ**, G. & LOLIVE, D. (2014a). ROOTS : un outil pour manipuler facilement, efficacement et avec cohérence des corpus annotés de séquences. *Journées d'Etude sur la Parole (JEP)*
- CHEVELU, J., **LECORVÉ**, G. & LOLIVE, D. (2014b). ROOTS : a toolkit for easy, fast and consistent processing of large sequential annotated data collections. *Proceedings of Language Resources and Evaluation Conference (LREC)*

Logiciel : https://gitlab.inria.fr/expression/tools/roots

Mettre au point de nouvelles méthodes de TALN implique la résolution de divers problèmes de gestion des données liés à la diversité des descripteurs à intégrer ainsi qu'à celle des formats utilisés par les outils qui les fournissent. Plusieurs propositions existent dans la communauté pour standardiser cette gestion et les chaînes de TALN auxquelles les données contribuent, par exemple, GATE (CUNNINGHAM et al. 2002), NXT (CALHOUN et al. 2010 ; CARLETTA et al. 2005) ou encore UIMA (FERRUCCI et LALLY 2004 ; FERRUCCI, LALLY et al. 2006). Pour le besoin des activités de recherche en synthèse de la parole au sein de l'IRISA, j'ai participé au développement d'un tel outil, nommé ROOTS.

ROOTS permet de représenter des données comme de multiples niveaux d'annotations, modélisés comme des séquences d'items, qu'il est possible de lier entre eux par des relations (*cf.* figure 7.1).

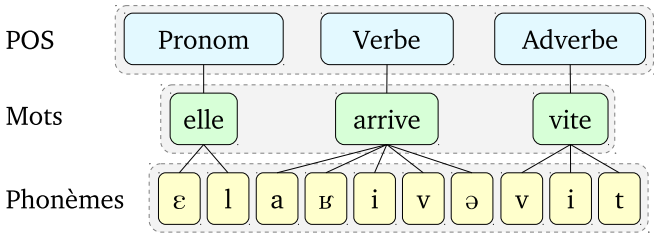


FIGURE 7.1 – Exemple de 3 séquences de types différents et mises en relation.

L'ajout d'informations tout au long d'une chaîne de traitement se fait alors par l'ajout de nouvelles séquences et de relations qui les relient aux séquences déjà existantes. L'outil s'appuie sur une bibliothèque écrite en C++ (environ 35 000 lignes de code) et rassemble de nombreux utilitaires pour le parcours et modification de corpus de données en ligne de commandes. ROOTS s'accompagne d'une API riche, largement documentée ainsi que de surcouches en Python et Perl. Son développement initial s'est étalé sur plusieurs années et ses mises à jour perdurent.

7.1.2 Plateforme d'évaluation perceptive

Collaborations : Sébastien Le Maguer (Trinity College Dublin, ADAPT Research Centre), Jonathan Chevelu (IRISA), Cédric Fayet (IRISA), Damien Lolive (IRISA), Claude Simon (IRISA)
Références :

— FAYET, C., BLOND, A., COULOMBEL, G., SIMON, C., LOLIVE, D., **LECORVÉ**, G., ... LE MAGUER, S. (2020). FlexEval, création de sites web légers pour des campagnes de tests perceptifs multimédias. *Actes de la conférence conjointe sur le Traitement Automatique du Langage Naturel et des Journées d'Étude sur le Parole (JEP-TALN) – session démo.*

Logiciel : https://gitlab.inria.fr/expression/tools/flexeval

De nombreux domaines de recherche requièrent la récolte d'avis d'utilisateurs, en particulier en TALN pour la validation de travaux. Comme déjà vu, il peut s'agir d'avis sur des signaux de parole ou sur des énoncés textuels mais il pourrait tout aussi bien s'agir de qualifier l'expressivité d'avatars signant en langue des signes ou encore la pertinence de résultats d'un moteur de recherche. Face à ce besoin redondant, nous avons développé un outil, FlexEval, pour créer et déployer de tels tests sans requérir de développements informatiques lourds ou de configurations complexes.

FlexEval modélise un test perceptif de manière générique comme l'enchaînement de différentes phases que le concepteur décide d'agencer selon son besoin. La principale phase consiste pour un utilisateur donné en une succession d'*étapes* identiques lui présentant un ou plusieurs échantillons de données avec une ou plusieurs questions et sollicitant ses réponses. Une fois terminée la campagne de test, le concepteur peut récupérer les réponses de tous les utilisateurs sous différents formats. Les autre phases permettent de gérer l'identification des testeurs, présenter des questionnaires indépendants de tout échantillon (par exemple, pour récolter l'âge et la localité de l'utilisateur, des commentaires en fin de test...), ainsi que des phases explicatives ne nécessitant aucune action de l'utilisateur (typiquement pour inclure un tutoriel). Le concepteur est complètement libre d'agencer ces phases, de les paramétrer et de personnaliser les pages web qui les joueront, notamment pour modifier leur identité visuelle. Enfin, FlexEval est facile à installer car il s'appuie sur des logiciels et bibliothèques standards de développement web.

7.1.3 Industrialisation d'un système de synthèse de la parole

Projets : Kaligo DYS (responsable local, Région Bretagne/FEDER), P2IA (responsable local,

Ministère de l'Éducation nationale), SPAM (SATT Ouest Valorisation)
Encadrements⋆/collaborations : ⋆Quentin Di-Fant (ingénieur d'étude), ⋆Simon Giddings (ingénieur d'étude), ⋆Waseem Safi (ingénieur de recherche), ⋆Aghilas Sini (ingénieur de recherche), ⋆Gaëlle Vidal (ingénieure d'étude), Pierre Alain (IRISA), Jonathan Chevelu (IRISA), Damien Lolive (IRISA)
Logiciel : Dépôts de logiciels en cours

Dans le cadre de multiples projets appliqués (*cf. infra*), je travaille sur le transfert technologique du moteur de synthèse de la parole de mon équipe de recherche à l'IRISA. Initialement, en tant qu'outil de recherche, ce système était constitué d'un ensemble de briques hétérogènes assemblées par différents scripts et la bibliothèque ROOTS. Cela pose différents problèmes dans une perspective de mise en production : lenteur, gestion anarchique des entrées-sorties, accès direct au code source. Par ailleurs, les outils de recherche ne couvrent généralement que les situations intéressantes scientifiquement et ne sont pas nécessairement robustes à des usages aussi diversifiés que des utilisateurs finaux peuvent avoir.

Ainsi, une partie considérable de mes activités depuis fin 2018 est lié à des questions d'industrialisation de ces multiples briques. Ceci consiste en la transposition en C++ des multiples briques de traitement du texte à synthétiser. En particulier, les briques de normalisation textuelle (réécriture des nombres, abbréviations, unités...) et de conversion graphèmes-phonèmes, dont je suis l'auteur initial[1]. Cela inclut également des problématiques de gestion humaine (synchronisation des tâches, livrables, service utilisateur...) et logicielle (versionnement, protection des logiciels et données, déploiement, compatibilité entre architectures...). Aujourd'hui, le système par sélection d'unités est entièrement opérationnel sur serveur Linux et mobile (Android et iOS). L'une des grandes perspectives à venir est l'intégration de méthodes de synthèse par réseaux de neurones avec des temps de rendu acceptables.

7.2 Projets appliqués

7.2.1 Adaptation de systèmes de reconnaissance automatique de la parole

Projets : TAO CSR (CTI, Suisse – 2011-2012 – IDIAP, Koemei)
Collaborations : John Dines (IDIAP, Koemei), Thomas Hain (University of Sheffield, Koemei), Petr Motlicek (IDIAP)
Références :

— LECORVÉ, G., DINES, J., HAIN, T. & MOTLICEK, P. (2012a). Impact du degré de supervision sur l'adaptation à un domaine d'un modèle de langage à partir du Web. *Actes des Journées d'Études sur la Parole et de la conférences sur le Traitement Automatique du Langage Naturel (JEP-TALN)*

— LECORVÉ, G., DINES, J., HAIN, T. & MOTLICEK, P. (2012b). Supervised and unsupervised Web-based language model domain adaptation. *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*

Dans le cadre de mon travail de postdoctorat, j'ai collaboré avec une entreprise commercialisant une solution de reconnaissance automatique de la parole. Je me suis intéressé à des questions déjà développées pendant ma thèse, à savoir l'adaptation des composants linguistiques d'un système de reconnaissance de la parole aux différents thèmes abordés dans les discours à transcrire. Cette tâche comprend la caractérisation du thème, la modification de la distribution de probabilités portées par le modèle de langage d'un tel système (renforcement des termes et tournures spécifiques au thème) et celle de son vocabulaire (ajout de termes spécifiques).

1. Le normaliseur de texte d'origine est publiquement disponible *via* https://github.com/glecorve/irisa-text-normalizer et le convertisseur graphèmes-phonèmes, dénommé *Grumph*, est celui issu des travaux présentés en section 2.1

7.2.2 Apprentissage du français par les enfants

Projets : Kaligo DYS (responsable local – Région Bretagne/FEDER – 2018-2020 – Learn&Go, IRISA, LOUSTIC, KARDI, HOALI), P2IA (responsable local – Ministère de l'Éducation nationale – 2019-2021 – Learn&Go, IRISA, LOUSTIC, INSPÉ Bretagne, Académie de Rennes, Académie de Caen)
Encadrements⋆/collaborations : ⋆Quentin Di-Fant (ingénieur d'étude), ⋆Simon Giddings (ingénieur d'étude), ⋆Gaëlle Vidal (ingénieure d'étude), Damien Lolive (IRISA)

Dans un contexte d'innovation numérique au service de la pédagogie, je participe à deux projets dont l'objectif est de permettre à des enfants de s'entraîner à l'écriture et à la lecture. Il s'agit de mettre en place des solutions d'apprentissage innovantes et pertinentes (exercices, jeux, parcours, explications...) par l'utilisation de tablettes munies d'un stylet et d'entrées-sorties audio. Les enfants ont ainsi une entrée intuitive, identique à leurs habitudes sur papier, qui ne va pas venir interférer avec leur processus de réflexion contrairement à la manipulation d'un clavier et d'une souris.

Dans ce projet, je coordonne le développement de trois fonctionnalités liées à la parole. La première est une brique de synthèse de la parole afin de dispenser les consignes et retours d'analyse aux enfants de manière orale. Celle-ci met notamment en avant la nécessité d'avoir une voix acceptable pour les enfants, avec un style proche de celui d'un.e enseignant.e. La seconde fonctionnalité est l'analyse de prononciations de mots faites par les enfants (alignement entre un signal audio et séquence phonétique cible). Enfin, la troisième brique vise à proposer automatiquement aux enfants des mots phonétiquement proches d'un autre afin de les exercer à les distinguer.

Outre l'entreprise qui commercialise l'application finale, ce travail se fait en interaction avec des spécialistes des usages de technologies de l'informations et de la communication et des experts dans l'apprentissage des langues. En particulier, l'un des projets (Kaligo DYS) vise la population des enfants souffrant de difficultés pathologiques dans cet apprentissage.

7.2.3 Intégration des migrants et réfugiés

Projets : NADINE (H2020 – 2018-2021 – 12 partenaires européens, coordinateur Script&Go)
Encadrements⋆/collaborations : ⋆Cédric Fayet (ingénieur de recherche), ⋆Gaëlle Vidal (ingénieure d'étude), ⋆Waseem Safi (ingénieur de recherche), Arnaud Delhay-Lorrain (IRISA), Damien Lolive (IRISA)

Le projet H2020 NADINE se place dans une perspective de faciliter l'accès à l'emploi par des migrants et réfugiés, car il s'agit d'un ingrédient-clé d'une intégration réussie dans les sociétés d'accueil. Cet accès est rendu difficile par les différences socioculturelles et linguistiques entre les pays d'accueil et d'origine, ne serait-ce que pour mettre en adéquation des attentes du marché avec des capacités des potentiels candidats.

Pour résoudre ce problème – devenu particulièrement important depuis les récentes vagues d'immigration en Europe, le projet NADINE, d'une part, à utiliser les données ouvertes pour mieux comprendre les réalités des migrants et les défis associés à leur intégration et, d'autre part, répondre aux besoins sociaux, éducatifs et culturels des migrants.

La contribution de l'IRISA dans ce projet est le développement de briques analogues à celles déjà développées pour les projets liées à l'apprentissage des langues mais dans plusieurs langues : celles des pays d'accueil (en l'occurrence, français, anglais, espagnol et grec) et d'origine (arabe et farsi).

7.2.4 Langues peu dotées

Projets : Synthèse du breton (co-responsable, Office public de la langue bretonne – 2019-2020 – IRISA, Skol Vreizh)
Encadrements⋆/collaborations : ⋆Hassan Hajipoor (ingénieur de recherche), ⋆Pascal Lintanf (linguiste contractuel), ⋆Gaëlle Vidal (ingénieure d'étude), ⋆Xuyang Zhang (stage), Elisabeth Delais-Roussarie (LLing), Anaïd Donabédian (Inalco), Tabita Toparlak (Inalco), Damien Lolive (IRISA)
Références :

— ALAIN, P., CHEVELU, J., GUENNEC, D., **LECORVÉ**, G. & LOLIVE, D. (2015). The IRISA Text-To-Speech System for the Blizzard Challenge 2015. *Proceedings of the Blizzard Challenge 2015 Workshop*
— TOPARLAK, T., DONABÉDIAN, A., LOLIVE, D., **LECORVÉ**, G. & DELAIS-ROUSSARIE, E. (2019). Synthèse vocale de l'arménien. Présenté à Digital Armenian

Mes activités en synthèse de la parole portent également sur le développement de systèmes pour des langues peu dotées, avec un enjeu fort relatif à la survie de ces langues dans le monde numérique d'aujourd'hui, à plus forte raison auprès des jeunes.

Mon travail le plus conséquent concerne le développement d'un système de synthèse pour le breton, tel que demandé par un marché public remporté par l'IRISA. Le moteur de synthèse étant indépendant de toute langue, les aspects travaillés sont la normalisation de texte, la conversion graphèmes-phonèmes, la prédiction prosodique à partir de séquences de phonèmes et la segmentation phonétique de signaux de parole. Le breton présente beaucoup de particularités pour chacune de ces étapes. Par exemple, le breton intègre un mécanisme de liaisons marqué par le voisement de certaines consonnes non voisées. Comme l'anglais, la notion d'accentuation tonique est également fondamentale. Le système de numération a aussi des spécificités. Ainsi, nous collaborons avec des specialistes de la langue bretonne pour repérer et traiter ces aspects. À cet égard, l'une des difficultés est que le breton se décline historiquement en plusieurs dialectes et que l'adoption du breton dit « standard » (par exemple, celui de la télévision) est parfois mal perçu car vu comme un breton hors-sol. C'est un aspect important pour la sélection des locuteurs dont la voix sera enregistrée pour produire le système final.

À une échelle nettement moindre, j'entretiens également une collaboration avec l'Inalco concernant l'arménien occidental. Il s'agit de produire un système de synthèse pour des mots isolés en vue de compléter un dictionnaire en ligne. Enfin, dans le cadre de l'édition 2015 du Blizzard Challenge, j'ai travaillé avec mes collègues sur 6 langues indiennes (bengali, hindi, malayalam, marathi, tamoul et télougou). Dans cette expérience, aucune expertise linguistique n'était disponible. Nos résultats, évalués en terme d'intelligibilité et de proximité acoustique avec les locuteurs d'origine, se sont avérés dans la moyenne des autres systèmes (ALAIN, CHEVELU et al. 2015).

Conclusion

Dans ce chapitre, nous avons passé en revue mes activités liées au soutien à la recherche et à la mise en application de connaissances scientifiques consolidées, principalement dans le domaine de la synthèse automatique de la parole. Bien que ces activités produisent peu en terme de publications scientifiques, elles font partie du cycle de la recherche. Elles sont d'ailleurs riches en enseignements quant au multiples tâches et compétences que requiert la conduite d'un projet depuis ses prémisses scientifiques jusqu'à son accès par le grand public.

Chapitre 8
Conclusion et perspectives

Ce manuscript d'habilitation à diriger des recherches a présenté une synthèse de mes activités de recherches sur les dix dernières années. Leur fil conducteur est l'analyse de la variabilité inhérente au langage naturel avec une prédominance sur les variations d'ordre stylistique. Les contributions sur le traitement de ces variations portent aussi bien sur le langage oral qu'écrit. En particulier, les chapitres 2 et 3 ont présenté des travaux sur la modélisation de variations phonétiques et sur le phénomène des disfluences dans le discours spontané. Ces deux activités sont appliquées au domaine de la synthèse de la parole. Nous avons également abordés mes travaux sur les registres familier, courant et soutenu (chapitre 4), ainsi que ceux portant sur le langage à destination des enfants (chapitre 5). Ceux-ci impliquent des interactions avec plusieurs champs de la linguistique. Ensuite, des travaux transversaux sur l'encodage de l'information linguistique *via* des techniques d'apprentissage neuronal ont été présentés (chapitre 6). Enfin, le chapitre 7 a mis en avant mes activités de développement logiciel et de diffusion à travers des projets appliqués.

Dans la suite, nous présentons des perspectives de recherche à court terme, dans la continuité des travaux présentés, puis à l'échelle plus longue d'une dizaine d'années sur des aspects plus fondamentaux.

8.1 Perspectives à court terme

J'encadre actuellement 4 thèses : celle d'Antoine Perquin (3e année) sur la synthèse de la parole par réseaux de neurones ; celle de Jade Mekki (2e année) sur la caractérisation automatique des registres de langue par l'extraction de motifs langagiers ; celle de Somayeh Jafaritazehjani (2e année) sur le transfert de style non supervisé ; et enfin celle d'Aline Étienne (1re année) sur la temporalité et les émotions dans des textes pour enfants. Par ailleurs, je supervise le travail de plusieurs chercheurs post-doctoraux et ingénieurs dans le cadre de projets. Mes perspectives à court terme s'inscrivent naturellement dans le prolongement de ces multiples travaux. Je les décline pour la modalité orale, puis écrite.

8.1.1 Modulation de la parole synthétique

Le domaine de la synthèse de la parole a connu une évolution majeure avec l'essor de l'apprentissage profond et les efforts de recherche des géants du numérique. Par exemple, l'introduction du vocodeur WaveNet par Google (van den Oord et al. 2006) a représenté un gain significatif en terme de rendu de la synthèse paramétrique (jusqu'alors assez médiocre) et a terminé de faire basculer l'essentiel des recherche vers cette famille, au détriment des approches par sélection d'unités. Ainsi, des architecture comme Tacotron 2 (J. Shen et al. 2018), Deep Voice 3 (Ping et al. 2018) ou FastSpeech (Ren et al. 2019) ont émergé. Ces approches tendent à poser le problème de la synthèse de parole neutre comme résolu et ouvre la recherche à l'intégration de facteurs de

variabilité, comme par exemple de multiples locuteurs, accents, émotions ou même langues. Ces approches reposent pour beaucoup sur la transformation de chacun de ces facteurs en un plongement qui vient conditionner des modèles acoustiques neuronaux.

— **Conditionnements multi-factoriels et linguistiques.** Dans ce contexte, mes perspectives se posent à différents niveaux. Tout d'abord, la combinaison des facteurs et l'objectif de disposer d'un système de synthèse universel ne sont pas encore des problèmes résolus. Ceci pose des questions quant à la conduite du protocole d'apprentissage car le nombre de paramètres du modèle devient important. Il est donc nécessaire de trouver des stratégies soit qui guident l'optimisation (par exemple, par des fonctions de perte dédiées), soit qui découpent l'apprentissage en plusieurs sous-apprentissages. Ensuite, les travaux actuels se concentrent essentiellement sur la dimension acoustique de la synthèse de la parole et peu d'entre eux travaillent sur sa dimension linguistique (au sens large) et sur l'étude du texte fourni en entrée des modèles. Pourtant, comme nous l'avons montré, la prononciation et les disfluences sont des éléments intéressants pour améliorer le pouvoir d'expressivité des systèmes. Ainsi, il serait intéressant de prolonger nos précédents travaux par l'utilisation de ces nouveaux modèles. Ceci semble notamment prometteur car la mise en cohérence de la prosodie avec ces deux phénomènes posaient jusqu'alors problème et l'une des forces de la synthèse neuronale est justement de bien traiter la prosodie.

— **Synthèse de la parole disfluente.** Sur la question spécifique de l'insertion de disfluences, nous avons souligné la difficulté à évaluer les résultats. La possibilité récente de recourir à des systèmes capables de produire un bon rendu des disfluences serait un point positif. Quand bien même cependant, l'emploi d'une telle synthèse ne serait pas exempt de biais favorisant des situations observées lors de l'apprentissage du système (c'est-à-dire généralement des situation de parole fluide). Ainsi, une autre piste est le développement de mesures objectives corrélées avec des indicateurs de perception, à l'image de ce que nous avons effectué dans (PERQUIN, **LECORVÉ** et al. 2019). La production d'un vaste ensemble d'annotations perceptuelles sur un volume important de variantes serait alors nécessaire pour rechercher et valider ces corrélations. Nos travaux sur les disfluences ont également montré les limites de l'insertion par un encodeur-décodeur neuronal standard. Je pense néanmoins qu'il s'agit d'une piste à poursuivre car les raisons de ces difficultés sont identifiées. D'une part, la séquence de sorties est très proche de la séquence d'entrée et, d'autre part, les éléments insérés doivent respecter une certaine structure. Ainsi, l'insertion automatique de disfluences se rapproche de la tâche de normalisation de textes, pour laquelle des propositions récentes seraient des pistes à explorer (H. ZHANG et al. 2019 ; J. ZHANG et al. 2020).

— **Démocratisation des approches.** Enfin, la qualité remarquable des signaux présentés dans beaucoup d'articles récents masque certaines limites. D'une part, beaucoup de travaux sont effectués sur l'anglais et sont difficilement transposables à d'autres langues car les corpus de parole nécessaires n'existent pas ou sont de faible taille, voire de mauvaise qualité. Ainsi, la constitution ou, plus pragmatiquement, la mutualisation de ressources est un objectif pour l'avenir, tout comme celui de développer des méthodes capables de travailler avec très peu de données (*zero-shot learning*, *few-shot learning*) ou sur des données bruitées. D'autre part, la gestion du temps et de la puissance de calcul, que ce soit à l'apprentissage ou à l'utilisation, est un élément à améliorer pour des raisons écologiques et de démocratisation de ces outils. À cet égard, les méthode d'adaptation de systèmes déjà appris (*fine tuning*) ou les techniques d'apprentissage par transfert semblent de bons points de départ pour ces études.

8.1.2 Modulation de textes

Les réseaux de neurones ont également apporté beaucoup de changements dans le domaine de la génération de textes, principalement sous l'impulsion des avancées en traduction auto-

matique (BAHDANAU et al. 2014). En particulier, le domaine de la génération automatique de paraphrases a connu des progrès significatifs (PRAKASH et al. 2016). À l'image de la section précédente, une part considérable des travaux actuels se portent désormais sur l'inclusion de diverses contraintes pour contrôler les textes générés, par exemple en fonction de la longueur souhaitée en sortie (KIKUCHI et al. 2016) ou d'une structure syntaxique particulière (IYYER et al. 2018).

— **Approches non supervisées.** En terme d'apprentissage automatique, ces méthodes se répartissent en deux familles. D'un côté, certains travaux font l'hypothèse de données textuelles alignées, c'est-à-dire que, pour chaque énoncé en entrée, le modèle peut s'appuyer sur une vérité terrain en sortie. À l'inverse, d'autres travaux visent une approche non supervisée où les corpus des styles source et cible ne sont pas alignés (mais généralement comparables, malgré tout). Ces approches s'appuient sur des architectures antagonistes où l'essence du style de chaque corpus est inférée automatiquement. L'approche non supervisée est plus difficile à mettre en place mais a l'avantage de pouvoir s'appliquer facilement à n'importe quelle situation, dès lors que l'on dispose de corpus spécifiques à chaque style. Comme évoqué dans le chapitre 6, la question qui se pose principalement est l'équilibre entre le transfert de style, le maintien du sens et la fluidité de la séquence de sortie. Le développement d'architectures qui garantissent mieux ces aspects est donc un enjeu. Pour cela, l'une des pistes est l'utilisation de fonction de pertes adéquates qui travaillent soient à partir des séquences produites en sorties, soit directement sur les plongements manipulés au sein du modèle. La difficulté est ici de définir ces fonctions. Par ailleurs, il est probable que la seule adoption de ces fonctions ne suffise pas car, là où style et sens peuvent s'envisager comme distincts dans l'espace des concepts, ils demeurent intriqués dans celui des mots. Ainsi, à mon sens, ces efforts sur les fonctions d'évaluation doivent être couplés avec l'étude de topologies qui proposent justement de distinguer (*distangle*) différentes dimensions d'un énoncé (BAO et al. 2019 ; JOHN et al. 2019).

— **Motifs linguistiques.** Une autre approche que je développe dans le cadre de la modulation de textes vise à rendre possible l'interaction avec l'humain. Ceci passe par (1) la possibilité de *caractériser* explicitement la distinction réciproque entre deux styles, c'est-à-dire produire des motifs discriminants compréhensibles par l'humain, et (2) celle de *reformuler* des textes sous la contrainte de ces mêmes motifs, en l'occurence afin d'en favoriser certains et d'en pénaliser d'autres. Dans cette approche, que nous appliquons aux registres de langue, des travaux doivent tout d'abord se poursuivre pour permettre l'extraction de motifs dans des conditions libres, à savoir sur la base d'un grand nombre de descripteurs et avec des distances potentiellement grandes entre itemsets des motifs. De même, les techniques actuelles retournent une grande quantité de motifs, souvent redondants. Le développement de méthodes de clustering de ces motifs serait une avancée pour simplifier leur analyse par des humains et leur éventuelle transmission à d'autres briques de TALN. Quant à la reformulation d'un texte, d'autres travaux sont à prévoir. Pour inclure des motifs comme des contraintes, une première piste, déjà en cours d'exploration, consiste à étudier la capacité des architectures encodeur-décodeur utilisées pour la production de paraphrases à être conditionnées de diverses manières (IYYER et al. 2018), c'est-à-dire à tenir compte de ces contraintes sans dégrader la qualité globale des paraphrases produites (fluidité, conservation du sens). Par ailleurs, un autre volet de travaux concerne la possibilité de conditionner ces architectures avec un nombre variable (et potentiellement important) de contraintes sans savoir à l'avance lesquelles seront appliquées pour tel ou tel texte. Ce type de problématique rejoint des propositions faites pour doter les modèles d'une mémoire augmentée (ZHAO et al. 2018).

— **Langage à destination des enfants.** Enfin, dans la perspective de mes travaux sur les textes à destination des enfants, d'autres étapes à venir s'annoncent. Nous allons, certes, nous porter sur l'étude de modèles de prédictions plus complexes et de caractéristiques linguistiques plus fines mais, au delà, il importe surtout de consolider le cadre méthodologique

de nos travaux afin d'offrir à la communauté la possibilité de s'y comparer et de les poursuivre. Cela passe par de nouvelles collaborations avec des psycholinguistes pour affiner la notion de compréhension et par des enquêtes auprès d'enfants de différentes tranches d'âge pour valider nos expérimentations. L'augmentation de notre quantité de données annotées est également primordiale. Nous avons largement entamé ce travail, notamment en complétant l'annotation en âge par une sous-catégorisation en genres journalistique, encyclopédique et fictionnel (romans, contes...). La question du partage de ces données est délicate car il est nécessaire de recueillir l'accord des auteurs ou des maisons d'éditions mais nous y travaillons également. Enfin, les problématiques abordées s'ouvrent à des tâches plus complexes de remédiations ou de reformulation. Alors que la tâche de reformulation a déjà été abordée plus haut, nous entendons par remédiation l'accompagnement d'auteurs pour leur faire prendre conscience de bonnes pratiques à adopter et d'erreurs à éviter, que ce soit dans l'absolu (indépendamment de tout texte) ou par l'analyse de textes en cours de rédaction. Dans ce dernier cas, ceci nécessite de savoir lier les prédictions des réseaux de neurones avec des interprétations linguistiques. Cette tâche s'inscrit dans le domaine grandissant de l'intelligence artificielle explicable (*Explainable Artificial Intelligence*, XAI) au sein duquel la justification de recommandations est une question déjà bien connue (Y. ZHANG, X. CHEN et al. 2020). Enfin, une autre application pour l'avenir des prédictions d'adéquation est l'inclusion de nos travaux dans un moteur de recherche dédié à un jeune public. En l'occurrence, dans le cadre du projet qui englobe ces activités, nous démarrons une collaboration avec le moteur Qwant Junior pour aller dans cette direction.

8.2 Perspectives à plus long terme

Si l'on fait un bilan de la décennie 2010-2020, la principale évolution du TALN tient dans le changement de son schéma-type pour résoudre une tâche. Historiquement, ce schéma s'appuyait successivement sur une phase d'analyse linguistique de la tâche, une autre de conception de descripteurs informatiques, puis une dernière d'assemblage des descripteurs, notamment *via* des techniques d'apprentissage automatique. Aujourd'hui, dans beaucoup de situations, c'est l'approche neuronale dite bout-en-bout (*end-to-end*) qui domine. Celle-ci se passe globalement de l'étape d'analyse linguistique et travaille directement sur des données brutes en laissant à l'algorithme d'optimisation la charge de déterminer les traitements nécessaires à leur appliquer (sélection, transformation...). Ce changement apparaît comme la simple poursuite du mouvement engagé par l'introduction des approches statistiques dans les années 1990 et de l'idée que les connaissances expertes sont totalement remplaçables par l'émanation récurrente des phénomènes latents présents dans les données.

Dans ce contexte, j'adresse ci-suit des perspectives plus générales, qui se détachent de la seule question du style et qui cristallisent des manques persistants dans les approches actuelles ou des enjeux que les tendances d'aujourd'hui laissent présager pour demain. Ces perspectives sont organisées autour de questions liées l'apprentissage automatique par réseaux de neurones, puis à la place des sciences humaines dans le TALN. Enfin, je conclus par un questionnement, plus politique, quant à l'organisation de la recherche en intelligence artificielle dans le futur.

8.2.1 Apprentissage neuronal

Avec l'adoption quasi générale des réseaux de neurones, la majorité des travaux récents reposent sur quelques principales architectures (encodeur-décodeur, feed-forward/CNN/RNN, mécanisme d'attention, GAN...). Cette évolution amène de nouvelles problématiques, plus fondamentales que celles listées précédemment et, à cet égard, également généralement partagées par d'autres domaines que le TALN. Je liste ci-dessous celles que je juge principales, accompagnée lorsque possible de premières pistes de réflexions.

— **Généricité et apprentissage multi-tâches.** L'essor du TALN amène une production régulière et rapide de nouveaux modèles. Néanmoins, leur utilisation dans un cas pratique nécessite souvent de les réapprendre de zéro ou de modifier quelques aspects des architectures. Et ce phénomène se répète à chaque innovation majeure en apprentissage automatique. Ce manque de généricité reflète à la fois une dépendance aux données d'apprentissage et à la tâche visée. En cela, je pense que les travaux en apprentissage par transfert (*transfer learning*) sont une piste importante à développer afin de capitaliser les apprentissages. En particulier, l'une des possibilités offertes par la littérature est d'apprendre des modèles multi-tâches, c'est-à-dire des modèles qui partagent une phase amont d'encodage de l'information linguistique, puis qui se décomposent en branches parallèles qui se spécialisent pour chaque tâche respective à traiter. Ces approches sont utiles pour produire des plongements génériques et pour construire des modèles indépendants d'une tâche particulière. L'une de ses limites fortes est que les corpus annotés pour de multiples tâches sont rares, de faible taille ou difficiles à constituer. Pour résoudre cela, une piste intéressante consiste en l'apprentissage par brassage des tâches. Dans cette approche, chaque tâche est associée à un corpus différents et, lors de l'apprentissage, il n'est demandé au modèle que de prédire un résultat que pour la tâche à laquelle l'exemple en entrée est associée.

— **Apprentissages légers.** Comme précédemment évoqué, la performance des approches actuelles tient en leur capacité à inférer statistiquement des connaissances latentes à partir de quantités massives d'exemples annotés. Face à l'objectif de pouvoir traiter toute tâche dans toute langue ou sous-langage d'une langue, il apparaît important de progresser sur la légèreté des apprentissages. Par légèreté, il est premièrement question de la quantité de données nécessaires. Il peut s'agir d'apprendre de zéro avec très peu de données ou de savoir rapidement adapter un modèle générique à partir de quelques exemples d'un domaine cible. Sous un autre angle, cette problématique rejoint en partie celle de savoir modéliser des signaux faibles, c'est-à-dire des signaux masqués par d'autres plus récurrents ou plus saillants en contexte, dans des grandes quantités de données. La légèreté concerne également la levée progressive de la supervision de l'apprentissage. Par exemple, dans les approches génératives, la levée de la nécessité d'aligner les entrées et sorties permet de traiter des tâches où cet alignement ne peut pas se faire. Il est remplacé par la confrontation de corpus comparables par des architectures antagonistes. À plus long terme, cela amène la question d'un apprentissage totalement non supervisée où le modèle identifie de lui-même les multiples dimensions d'un contenu et la manière dont celles-ci varient ou non d'un échantillon/corpus à un autre. Enfin, point également déjà abordé dans les perspectivement à court terme, il me semble important de souligner de nouveau l'objectif de réduire la consommation énergétique des modèles (apprentissage et utilisation) étant donné la crise climatique et la démocratisation galopante des outils d'intelligence artificielle.

— **Analyse et formalisation des espaces latents pour la conception de modèles complexes.** La tendance logique de la recherche consiste à étudier des tâches de plus en plus complexes, nécessitant des analyses à de multiples niveaux. En guise d'illustration, en synthèse de la parole neuronale, les travaux portaient initialement sur la synthèse mono-locuteur, puis ils se sont déplacés vers le cas multi-locuteurs, puis encore plus récemment vers la synthèse multi-accents ou multilingues, avec différents styles, etc. Ces changements impliquent à chaque fois d'ajouter de nouvelles briques d'encodage/reconnaissance de locuteurs, d'accents, de styles, etc. Ainsi, les modèles de synthèse deviennent de plus en plus complexes, tant en nombre de paramètres à estimer qu'en terme de sophistication dans la conception (choix des fonctions de coûts, hyper-paramètres, stratégies d'optimisation...). Et ce constat s'applique à de nombreuses tâches en TALN. Ainsi, pour maîtriser cette complexification, il me semble important de progresser sur les stratégies d'assemblage de modèles « élémentaires ». L'une des pistes pour cela est l'étude des propriétés des espaces de plongements. Par exemple, il est actuellement difficile de déterminer avec précision l'information que contiennent des plongements (à partir d'une donnée en entrée, quelles informations sont gardées ou sup-

primées ?). À plus forte raison, les propriétés qu'entretiennent ou que l'on souhaite voire entretenir entre eux des plongements sont, à mon sens, encore largement laissées de côté (orthogonalité, invariance, transformations géométriques). De même, la dimension intrinsèque d'un plongement est une propriété peu intégrée dans la conception de modèles, remplacée par une phase de réglage empirique d'hyper-paramètres. Des avancées sur ces aspects permettraient de mieux interfacer des plongements d'origine diverses. Par exemple, elles laisseraient imaginer la conception de modèles dont l'espace de plongement offre des garanties topologiques telles que l'adjonction de nouvelles informations ne présenterait ni risques d'altération des performances ni nécessité d'un ré-entrainement du modèle.

— **Données structurées.** Enfin, l'une des limites actuelles des travaux en TALN est de se limiter à des grains de données relativement restreints (celui de la phrase ou de la proposition). Or, le besoin de travailler à des échelles plus grandes est présent dans de nombreuses tâches. Par exemple, en résumé automatique, l'analyse globale d'un document permet de connaître la trame générale d'un discours et d'en produire une approximation qui maximise la conservation de l'information. Dans les systèmes de questions-réponses, l'interconnexion de phrases permet de répondre à des questions plus complexes. En synthèse de la parole, l'historique d'un paragraphe ou chapitre permet de situer certains éléments du discours (personnage, intrigue...) ou de proposer une prosodie moins répétitive lorsque les phrases s'enchaînent. Dans d'autres situations, c'est à l'échelle de la collection de documents que l'analyse se porte. À nouveau, la tâche de résumé automatique est concernée mais citons également la vérification de faits, la recherche d'informations, l'attribution d'auteurs... À cela peut s'ajouter le besoin de connaître l'organisation interne des documents ou collections. Cette organisation peut représenter une structuration en terme de proximités thématiques entre textes, de relations sociales ou hiérarchiques entre entités ou encore de logique argumentaire entre paragraphes. Les techniques actuelles d'apprentissage profond sont encore relativement limitées pour ces problèmes. Tout d'abord, en comparaison aux tâches habituelles, les entrées des modèles sont structurellement plus complexes (longues séquences, séquences de séquences, arbres, graphes...). Ensuite, étant donné la quantité d'information accrue, les besoins de mémorisation et de découverte d'interdépendances (mécanismes d'attentation) croissent également. Ceci a un impact direct sur la complexité des modèles et la difficulté à les apprendre. Enfin, paradoxalement, alors que chaque donnée est relativement grosse dans ce type de problème, le travail à une échelle plus large rend difficile la constitution de corpus avec un grand nombre de données. Ainsi, les tâches au delà de la phrase s'appuient souvent sur des systèmes où les réseaux de neurones sont combinés à d'autres techniques ou s'insèrent dans des algorithmes *ad hoc* d'aggrégation de l'information.

Pour conclure sur les perspectives liées à l'apprentissage automatique, il est important de noter que celles-ci sont souvent intriquées. Par exemple, les propriétés de l'espace de plongement d'un modèle ont un impact sur sa généricité. De même, l'étude de structures de données complexes nécessite de savoir mieux apprendre sur des récurrences diffuses. Ainsi, sur le plan organisationnel, ces perspectives devraient prendre place dans le cadre d'efforts coordonnés.

8.2.2 La place des sciences humaines

Une autre incidence du succès des approches neuronales est la tendance à l'agnostisme concernant les données manipulées. En cela, les perspectives de fertilisation croisées avec les sciences humaines semblent se réduire. Cette section partagent quelques réflexions qui, je l'espère, démontre au contraire l'avenir de ces interactions.

— **Modèle de communication et d'interaction.** Comme exposé en introduction de ce manuscrit, la réalisation d'un énoncé s'inscrit dans un schéma de communication où de multiples facteurs contextuels se combinent (énonciateur, mode, public...). La tendance générale des travaux en TALN consiste à étudier l'un de ces facteurs et à le corréler avec les obser-

vations linguistiques offertes par les énoncés (lexique, syntaxe, prosodie...). L'objectif peut être soit de diagnostiquer la présence d'un facteur parmi les causes principales de l'énoncé, soit, à l'inverse, de générer un énoncé crédible qui simule le facteur étudié. Néanmoins, à ma connaissance, la mise en lien plus généralisées de tous les facteurs d'influence reste peu étudiée en TALN. Ainsi, la simulation d'un schéma de communication complet reste un enjeu. Pour illustrer les problèmes à résoudre, nous pouvons prendre l'exemple d'un répondeur automatique intelligent qui répondrait à des appels téléphoniques ou messages textuels reçus quand le propriétaire du téléphone n'est pas disponible. Pour que l'assistant intelligent remplisse parfaitement sa tâche, il se devrait d'inférer le lien qui lie l'utilisateur et l'interlocuteur, le sujet dont il est question, le positionnement de chaque partie prenante par rapport à ce sujet, etc. Il s'agit d'un sujet difficile qui ne peut, à mon sens, se résoudre sans une approche multidisciplinaire. D'une part, la reconnaissance de certains facteurs contextuels dans un dialogue ou dans un historique de discussions appelle encore de nombreux travaux. Par exemple, la détection de sous-entendus ou d'ironie, l'estimation d'un niveau de connaissance sur un sujet ou le diagnostic fin d'une opinion sont des tâches utiles sur lesquelles des travaux sont à conduire. D'autre part, les relations entre facteurs ne sont évidentes. Par exemple, si l'interlocuteur est un ami, un style relâché semble adéquat mais il ne l'est sans doute plus si le sujet est sérieux. Les sciences humaines (psychologie, sociologie, linguistiques) doivent permettre d'apporter un éclairage sur ces questions. Dans un cadre plus large, ces questions ouvrent également vers la prise en compte d'autres modalités pour l'induction de ces éléments contextuels (expressions du visage, gestes, informations physiologiques...), par exemple pour analyser un niveau de stress ou encore l'éloquence d'un discours.

— **Évaluation de systèmes complexes.** À l'image des considérations de la section précédente, l'évolution du TALN, et plus généralement de l'intelligence artificielle, conduit la recherche vers des tâches de plus en plus complexes, dont l'étude requiert des méthodologies de plus en plus élaborées. Premièrement, les contraintes liées à la collecte de données augmentent. Il s'agit en effet de recueillir des informations plus riches, éventuellement organisées entre elles, dans des situations plus variées et, *a priori*, en plus grande quantité. Ensuite, les conditions de validation des systèmes doivent garantir l'absence de biais lors des évaluations (par exemple, domaine trop restreint, dimensions oubliées, métriques inadaptées, connaissances antérieures des testeurs, durée trop courte des tests...). Ceci est d'autant plus important que certains systèmes, comme ceux de production automatique de contenus, ne peuvent se passer de campagnes d'évaluation en situation factice ou réelle. Enfin, la question de la pérennité des systèmes développés me semble également trop peu étudiées. Par exemple, il s'agit de savoir si les performances d'un assistant intelligent qui s'adapte aux usages de son propriétaire ne vont pas se dégrader avec le temps car un décalage trop grand apparaîtra petit à petit entre les conditions initiales de développement du système et celles d'utilisation. Ce comportement est généralement difficile à appréhender car les systèmes sont constituées de multiples composants dont les effets de bord ne sont pas nécessairement bien connus. Par leurs travaux sur l'usage de la langue, son évolution ou encore les comportements humains, les sciences humaines offrent des pistes de solutions ou de garanties faces à ces multiples problématiques.

— **Sauvegarde des langues.** Enfin, bien que mes activités liées à des langues peu dotées soient relativement peu développées, il me semble important que le TALN se place comme outil de sauvegarde du patrimoine culturel de l'humanité et notamment de celle des langues. En effet, alors que des milliers de langues existent aujourd'hui, seule une petite minorité d'entre elles dominent et tirent partie des avancées de la recherche. Cette fracture participe au risque de disparation des autres langues car leur apprentissage ou leur pratique n'est pas favorisé, notamment auprès des nouvelles générations. Il s'agit d'un sujet où le TALN et les sciences humaines doivent pouvoir collaborer pour constituer et traiter des ressources linguistiques (textes, parole, connaissances...) ainsi que proposer des outils (informatiques

ou non) pour leur traitement (automatique ou non). En particulier, certaines initiatives visant à documenter des langues rares ou non écrites par des techniques computationelles me semblent prometteuses pour progresser sur ces enjeux. Ces travaux visent à construire automatiquement des éléments de linguistique formelle (lexiques, grammaires...) dans un principe d'échange avec les linguistes sur ces langues.

8.2.3 Quelle place pour quelle recherche ?

Enfin, je termine ce manuscrit par une mise en perspective des évolutions organisationnelles de ces dernières années en TALN et en intelligence artificielle. L'objectif est de s'interroger sur leur avenir, sans prise de position particulière.

— **L'omniprésence des géants du numérique.** Un premier constat est que beaucoup d'avancées majeures récentes sont le fait des entreprises géantes du numérique (Google, Baidu, Facebook, Apple, Microsoft...), fortes de leurs moyens financiers, matériels et humains. En outre, ces entreprises disposent naturellement par leurs activités de très grandes quantités de données pour alimenter l'apprentissage de leurs modèles (requêtes, messages, chat, photos, vidéos, graphes d'amitiés, de préférences...). Or, ces données ne peuvent généralement pas être diffusées en raison de la propriété intellectuelle. Face à cette situation, il est légitime de se demander si les acteurs institutionnels doivent chercher à concurrencer les géants du numériques. Dans le camp du « non », il est en effet raisonnable d'estimer que certaines tâches progresseront quoiqu'il en soit, même sans les efforts de la recherche publique car ces tâches sont au cœur des activités des acteurs industriels. Par ailleurs, par le rôle de formation des universités, il peut être risqué pour la réussite d'un doctorant de le lancer sur un sujet sur lequel ces acteurs travaillent. Ainsi, les efforts libérés pourraient servir d'autres sujets, moins prioritaires (notamment commercialement) pour ces acteurs et qui restent difficiles pour les méthodes actuelles (par exemple, les langues peu dotées, l'explicabilité des modèles...). Pour autant, dans le camp du « oui », il paraît illusoire que quelques acteurs de la recherche puissent être aussi créatifs que la communauté tout entière. De multiples avancées médiatisées sont en réalité le fruit de propositions initiales moins spectaculaires. Par ailleurs, le rôle de transmission des savoirs des acteurs publiques oblige à ne pas laisser les connaissances devenir l'exclusivité de quelques uns. En ce sens, la veille sur l'état de l'art peut être vue comme un devoir envers la société, bien qu'elle puisse parfois paraître vaine pour ceux qui la conduisent.

— **L'accélération de la recherche.** Un deuxième constat est que le rythme de la recherche s'est beaucoup accéléré. De nombreuses propositions sont ainsi publiées en dehors des conférences et journaux (par exemple *via* arXiv), si bien que d'autres propositions présentent au bout de quelques mois des extensions de travaux à peine publiés. Ensuite, il est devenu commun si ce n'est fortement recommendé de partager le code et les données qui ont permis l'expérimentation d'une propoposition. Sans remettre en question cette saine pratique, elle a comme effet de bord de soutenir l'accélération en facilitant la reprise des travaux par d'autres. Tout comme pour la concurrence des géants du numériques, certains sujets sont ainsi difficiles à suivre.

— **La nécessité du travail en équipe.** Enfin, paradoxalement, le coût du ticket d'entrée dans un domaine est parfois devenu élevé. En dépit des bonnes pratiques et des accès facilités aux connaissances, la conduite en autonomie de nouveaux travaux requiert un ensemble de connaissances pratiques de plus en plus important, notamment dans le domaine de l'apprentissage neuronal. À plus forte raison, certains travaux majeurs ne sont pas directement reproductibles car le code n'est pas public, certains hyper-paramètres ne sont pas spécifiés ou les données ne sont pas distribuées. Ainsi, la conduite de travaux requiert un travail d'équipe pour l'appropriation des savoirs-faire et la conduite des aspects expérimentaux (collecte, réglages, évaluation).

Bien que ce panorama puisse paraître sombre, il vise simplement à souligner, en conclusion de ce manuscrit, l'importance d'inciter à des réflexions collégiales sur ces constats, leurs conséquences (positives comme négatives) et à d'éventuelles actions, en particulier pour garantir la pluralité et la vitalité de la recherche de demain.

Partie III

Annexes

Annexe A

Liste des publications

Chapitre de livre international (1)

Huet, S., **Lecorvé**, G., Gravier, G. & Sébillot, P. (2008). Toward the Integration of Natural Language Processing and Automatic Speech Recognition : Using Morpho-Syntax and Pragmatics for Transcription. P. Maragos, A. Potamianos & P. Gros (Éd.), *Multimodal Processing and Interaction : Audio, Video, Text*. Springer US

Journaux internationaux (2)

Gravier, G., Guinaudeau, C., **Lecorvé**, G. & Sébillot, P. (2011). Exploiting Speech for Automatic TV Delinearization : From Streams to Cross-Media Semantic Navigation. *EURASIP Journal on Image and Video Processing, 2011*

Tahon, M., **Lecorvé**, G. & Lolive, D. (2018). Can we Generate Emotional Pronunciations for Expressive Speech Synthesis ? *IEEE Transactions on Affective Computing*

Conférences internationales (17)

Lecorvé, G., Gravier, G. & Sébillot, P. (2008a). An unsupervised web-based topic language model adaptation method. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE

Lecorvé, G., Gravier, G. & Sébillot, P. (2008b). On the Use of Web Resources and Natural Language Processing Techniques to Improve Automatic Speech Recognition Systems. *Proceedings of the Language Resources and Evaluation Conference (LREC)*

Lecorvé, G., Gravier, G. & Sébillot, P. (2009). Constraint selection for topic-based MDI adaptation of language models. *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*

Lecorvé, G., Gravier, G. & Sébillot, P. (2011). Automatically Finding Semantically Consistent N-grams to Add New Words in LVCSR Systems. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*

Lecorvé, G. & Motlicek, P. (2012). Conversion of recurrent neural network language models to weighted finite state transducers for automatic speech recognition. *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*

Lecorvé, G., Dines, J., Hain, T. & Motlicek, P. (2012b). Supervised and unsupervised Web-based language model domain adaptation. *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*

CHEVELU, J., **LECORVÉ**, G. & LOLIVE, D. (2014b). ROOTS : a toolkit for easy, fast and consistent processing of large sequential annotated data collections. *Proceedings of Language Resources and Evaluation Conference (LREC)*

LECORVÉ, G. & LOLIVE, D. (2015). Adaptive statistical utterance phonetization for French. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE

QADER, R., **LECORVÉ**, G., LOLIVE, D. & SÉBILLOT, P. (2015). Probabilistic Speaker Pronunciation Adaptation for Spontaneous Speech Synthesis Using Linguistic Features. *Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP)*

TAHON, M., QADER, R., **LECORVÉ**, G. & LOLIVE, D. (2016a). Improving TTS with corpus-specific pronunciation adaptation. *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*

TAHON, M., QADER, R., **LECORVÉ**, G. & LOLIVE, D. (2016b). Optimal feature set and minimal training size for pronunciation adaptation in TTS. *Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP)*. Springer

QADER, R., **LECORVÉ**, G., LOLIVE, D., TAHON, M. & SÉBILLOT, P. (2017). Statistical pronunciation adaptation for spontaneous speech synthesis. *Proceedings of the International Conference on Text, Speech, and Dialogue (TSD)*. Springer

TAHON, M., **LECORVÉ**, G., LOLIVE, D. & QADER, R. (2017). Perception of expressivity in TTS : linguistics, phonetics or prosody ? *Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP)*. Springer

QADER, R., **LECORVÉ**, G., LOLIVE, D. & SÉBILLOT, P. (2018). Disfluency Insertion for Spontaneous TTS : Formalization and Proof of Concept. *Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP)*. Springer

PERQUIN, A., **LECORVÉ**, G., LOLIVE, D. & AMSALEG, L. (2018). Phone-Level Embeddings for Unit Selection Speech Synthesis. *Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP)*. Springer

LECORVÉ, G., AYATS, H., FOURNIER, B., MEKKI, J., CHEVELU, J., BATTISTELLI, D. & BÉCHET, N. (2019). Towards the Automatic Processing of Language Registers : Semi-supervisedly Built Corpus and Classifier for French. *International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*

BLANDIN, A., **LECORVÉ**, G., BATTISTELLI, D. & ÉTIENNE, A. (2020a). Age recommendation for texts. *Proceedings of the Language Resources and Evaluation Conference (LREC)*

Ateliers internationaux (5)

IMSENG, D., BOURLARD, H., CAESAR, H., GARNER, P. N., **LECORVÉ**, G. & NANCHEN, A. (2012). MediaParl : Bilingual mixed language accented speech database. *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE

ALAIN, P., CHEVELU, J., GUENNEC, D., **LECORVÉ**, G. & LOLIVE, D. (2015). The IRISA Text-To-Speech System for the Blizzard Challenge 2015. *Proceedings of the Blizzard Challenge 2015 Workshop*

ALAIN, P., CHEVELU, J., GUENNEC, D., **LECORVÉ**, G. & LOLIVE, D. (2016). The IRISA Text-To-Speech System for the Blizzard Challenge 2016. *Proceedings of the Blizzard Challenge 2016 Workshop*

LOLIVE, D., ALAIN, P., BARBOT, N., CHEVELU, J., **LECORVÉ**, G., SIMON, C. & TAHON, M. (2017). The IRISA Text-To-Speech System for the Blizzard Challenge 2017. *Proceedings of the Blizzard Challenge Workshop*

ALAIN, P., **LECORVÉ**, G., LOLIVE, D. & PERQUIN, A. (2018). The IRISA Text-To-Speech System for the Blizzard Challenge 2018. *Proceedings of the Blizzard Challenge 2018 Workshop*

Conférences nationales (13)

Lecorvé, G., Gravier, G. & Sébillot, P. (2008c). Vers une adaptation thématique non supervisée de modèles de langage : utilisation d'Internet comme un corpus ouvert. *Journées d'Étude sur la Parole et Conférence sur le Traitement Automatique du Langage Naturel (JEP-TALN)*

Lecorvé, G., Gravier, G. & Sébillot, P. (2010). L'adaptation thématique d'un modèle de langue fait-elle apparaître des mots thématiques ? *Actes des Journées d'Étude sur la Parole (JEP)*

Lecorvé, G., Dines, J., Hain, T. & Motlicek, P. (2012a). Impact du degré de supervision sur l'adaptation à un domaine d'un modèle de langage à partir du Web. *Actes des Journées d'Études sur la Parole et de la conférences sur le Traitement Automatique du Langage Naturel (JEP-TALN)*

Chevelu, J., **Lecorvé**, G. & Lolive, D. (2014a). ROOTS : un outil pour manipuler facilement, efficacement et avec cohérence des corpus annotés de séquences. *Journées d'Etude sur la Parole (JEP)*

Lecorvé, G. & Lolive, D. (2016). Phonétisation statistique adaptable d'énoncés pour le français. *Actes des Journées d'Études sur la Parole (JEP)*

Qader, R., **Lecorvé**, G., Lolive, D. & Sébillot, P. (2016). Adaptation de la prononciation pour la synthèse de la parole spontanée en utilisant des informations linguistiques. *Actes des Journées d'Études sur la Parole (JEP)*

Qader, R., **Lecorvé**, G., Lolive, D. & Sébillot, P. (2017). Ajout automatique de disfluences pour la synthèse de la parole spontanée : formalisation et preuve de concept. *Traitement automatique du langage naturel (TALN)*. **Prix du meilleur article**

Lecorvé, G., Ayats, H., Fournier, B., Mekki, J., Chevelu, J., Battistelli, D. & Béchet, N. (2018). Construction conjointe d'un corpus et d'un classifieur pour les registres de langue en français. *Traitement automatique du langage naturel (TALN)*

Mekki, J., Battistelli, D., **Lecorvé**, G. & Béchet, N. (2018). Identification de descripteurs pour la caractérisation de registres. *Actes des Rencontres Jeunes Chercheurs (RJC) de la conférence CORIA-TALN*

Perquin, A., **Lecorvé**, G., Lolive, D. & Amsaleg, L. (2019). Évaluation objective de plongements pour la synthèse de parole guidée par réseaux de neurones. *Actes de la conférences sur le Traitement automatique du langage naturel (TALN)*

Étienne, A., Battistelli, D. & **Lecorvé**, G. (2020a). Apports de la linguistique et du TAL à l'analyse des émotions dans les textes pour enfants. *Actes de colloque Langage et éMOTions*

Étienne, A., Battistelli, D. & **Lecorvé**, G. (2020b). L'expression des émotions dans les textes pour enfants : constitution d'un corpus annoté. *Actes de la Conférences sur le Traitement Automatique du Langage Naturel (TALN)*

Blandin, A., **Lecorvé**, G., Battistelli, D. & Étienne, A. (2020b). Recommendation d'âge pour des textes. *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (JEP-TALN)*

Fayet, C., Blond, A., Coulombel, G., Simon, C., Lolive, D., **Lecorvé**, G., ... Le Maguer, S. (2020). FlexEval, création de sites web légers pour des campagnes de tests perceptifs multimédias. *Actes de la conférence conjointe sur le Traitement Automatique du Langage Naturel et des Journées d'Étude sur le Parole (JEP-TALN) – session démo.*

Mekki, J., Béchet, N., Battistelli, D. & **Lecorvé**, G. (2020). Caractérisation de registres de langue par extraction de motifs séquentiels émergents. *Actes des Journées Internationales d'Analyse statistique des Données Textuelles (JADT)*

Communication sans actes (1)

Toparlak, T., Donabédian, A., Lolive, D., **Lecorvé**, G. & Delais-Roussarie, E. (2019). Synthèse vocale de l'arménien. Présenté à Digital Armenian

Mémoires personnels (2)

LECORVÉ, G. (2007). *Adaptation thématique d'un système de transcription automatique de la parole* (mém. de mast., INSA Rennes)

LECORVÉ, G. (2008). *Adaptation thématique non supervisée d'un système de reconnaissance automatique de la parole* (thèse de doct.). **Prix de thèse de l'association francophone de la communication parlée**

Mémoires encadrés (6)

QADER, R. (2017). *Pronunciation and disfluency modeling for expressive speech synthesis* (thèse de doct., University of Rennes 1)

PERQUIN, A. (2017). *Big deep voice : indexation de données massives de parole grâce à des réseaux de neurones profonds* (mém. de mast., University of Rennes 1)

LASSELIN, H. (2018). *Make text look like speech : disfluency generation using sequence-to-sequence neural networks* (Rapport de recherche, Rapport de recherche. IRISA)

BLANDIN, A. (2019). *Prédiction de recommandations d'âge pour l'accès à des enfants à des textes* (mém. de mast., Université de Rennes 1)

ÉTIENNE, A. (2019). *Compréhension de textes par les enfants et émotions : point(s) de vue psycholinguistique(s) et leur mise en œuvre en TAL* (mém. de mast., Université de Paris-Nanterre)

DABBEBI, I. (2015). *Emerging Pattern Mining to Characterize Language Registers in French* (mém. de mast., Université de Tunis)

Rapports de recherche (3)

QADER, R., **LECORVÉ**, G., LOLIVE, D. & SÉBILLOT, P. (2014). *Phonology Modelling for Expressive Speech Synthesis : a Review* (Rapport de recherche N° PI-2020). Rapport de recherche. IRISA

MEKKI, J., BATTISTELLI, D., BÉCHET, N. & **LECORVÉ**, G. (2017). ≪ *Nous nous arrachâmes promptement avec ma caisse* ≫ *: quels descripteurs linguistiques caractérisent les registres de langue ?* Rapport de recherche. IRISA, MoDyCo

BOURGOIN, C., BATTISTELLI, D. & **LECORVÉ**, G. (2018). *Les notions temporelles dans la mise en récit d'événements dans le discours journalistique enfantin*. Rapport de recherche. IRISA, MoDyCo

Annexe B

Encadrements

Stagiaires de recherche (10)

— **Hugo Ayats :** élève-ingénieur en informatique (stage de 6 semaines), juin 2017 – juillet 2017, « Collecte, préparation et annotation de textes familiers, neutres et soutenus », ANR TREMoLo, 50 % d'encadrement avec Jonathan Chevelu (50 %).

— **Alexis Blandin :** étudiant de master 2 recherche en informatique, octobre 2018 – juillet 2019, « Prédiction d'une recommandations d'âge pour l'accès par des enfants à des textes », 50 % d'encadrement avec Delphine Battistelli (50 %).

— **Charlotte Bourgoin :** étudiante de master 2 recherche en linguistique anglaise à l'université Paris-Diderot (stage), février 2018 – juillet 2018, « Aide rédactionnelle pour le récit d'événements à des enfants », ANR TREMoLo, 40 % d'encadrement avec Delphine Battistelli (40 %) et Elsa Maudet (20 %) de Libération.

— **Inès Dabbebi :** élève de mastère recherche en sciences et techniques de l'informatique décisionnelle, de mars à juillet 2015, « Fouille de motifs séquentiels émergents afin de caractériser des registres de langue », 50 % d'encadrement avec Nicolas Béchet (50 %).

— **Aline Étienne :** étudiante de master 2 en traitement automatique des langues, mars 2019 – juillet 2019, « Compréhension de textes par les enfants et émotions : point(s) de vue psycholinguistique(s) et leur mise en œuvre en TAL », ANR TREMoLo, 50 % d'encadrement avec Delphine Battistelli (50 %).

— **Benoît Fournier :** élève-ingénieur en informatique (stage de 6 semaines), juin 2017 – juillet 2017, « Collecte, préparation et annotation de textes familiers, neutres et soutenus », ANR TREMoLo, 50 % d'encadrement avec Jonathan Chevelu (50 %).

— **Henri Lasselin :** étudiant de master 2 science informatique (étude bibliographique + stage), novembre 2017 – juillet 2018, « Make text look like speech : disfluency generation using sequence-to-sequence neural networks », fonds propres, 100 % d'encadrement.

— **Jade Mekki :** étudiante de master 1 en traitement automatique des langues à l'université Paris-Nanterre (stage), avril 2017 – juillet 2017, « Nous nous arrachâmes promptement avec ma caisse » : quels descripteurs linguistiques caractérisent les registres de langue ?, ANR TREMoLo, 25 % d'encadrement avec Delphine Battistelli (50 %) et Nicolas Béchet (25 %).

— **Antoine Perquin :** étudiant de master 2 recherche en informatique (étude bibliographique + stage), octobre 2016 – juillet 2017, « Big deep voice : indexation de données massives de parole grâce à des réseaux de neurones profonds », ANR SynPaFlex, 50 % d'encadrement avec Damien Lolive (50 %).

— **Xuyang Zhang :** élève-ingénieure en informatique, de septembre 2014 à mars 2015, « Utilisation de Wiktionary pour construire un service de phonétisation multilingue », non rémunéré (interne à la formation), 100 % d'encadrement.

Doctorants (1 thèse soutenue, 4 en cours)

— **Aline Étienne :** traitement automatique des langues, depuis décembre 2019, « Temporalité et émotions dans la compréhension de textes par des enfants : une problématique pour le TAL », ANR TextToKids, 33 % d'encadrement, avec Delphine Battistelli (Univ. Paris-Nanterre, 67 %).
— **Somaye Jafaritazehjani :** informatique, depuis novembre 2018, « Deep neural natural language style transfer », CD UR1 et Technical University of Dublin (Irlande), 33 % d'encadrement, avec Damien Lolive (IRISA, Lannion, 33 %) et John D. Kelleher (TU Dublin, Irlande, 34 %).
— **Jade Mekki :** informatique, depuis septembre 2018, « Caractérisation de registres de langue par extraction de motifs séquentiels », ANR TREMoLo, co-directeur, 25 % d'encadrement, avec Nicolas Béchet (IRISA, Vannes, 50 %) et Delphine Battistelli (MoDyCo, Nanterre, 25 %).
— **Antoine Perquin :** informatique, depuis octobre 2017, « Universal speech synthesis through embeddings of massive heterogeneous data », CD INSA Rennes, 37,5 % d'encadrement, avec Damien Lolive (IRISA, Lannion, 37,5 %) et Laurent Amsaleg (IRISA, Rennes, 25 %).
— **Raheel Qader :** informatique, allocation ministérielle, de janvier 2014 à mars 2017, « Pronunciation and disfluency modeling for expressive speech synthesis », co-directeur, 50 % d'encadrement avec Pascale Sébillot (20 %) et Damien Lolive (30 %).

Chercheurs post-doctoraux (7, dont 4 en cours)

— **Nazanin Dehghani :** depuis août 2019, « Paraphrase generation for language register transfer », ANR TREMoLo, 50 % d'encadrement avec Jonathan Chevelu (50 %).
— **Cédric Fayet :** d'octobre 2019 à décembre 2019, « Multilingual speech processing », H2020 NADINE, 50 % d'encadrement avec Damien Lolive (50 %).
— **David Guennec :** depuis mai 2020, « Deep neural network text-to-speech », P2IA, 50 % d'encadrement avec Damien Lolive (50 %).
— **Md Rashedur Rahman :** depuis mai 2020, « Natural language processing for kids », ANR TextToKids, 50 % d'encadrement avec Nicolas Béchet (25 %) et Jonathan Chevelu (25 %).
— **Waseem Safi :** de novembre 2018 à août 2019, « Multilingual speech processing », H2020 NADINE, 50 % d'encadrement avec Damien Lolive (50 %).
— **Aghilas Sini :** depuis mars 2020, « Multilingual speech processing », H2020 NADINE, 33 % d'encadrement avec Arnaud Delhay-Lorrain (33 %) et Damien Lolive (34 %).
— **Marie Tahon :** de décembre 2015 à août 2017, « Pronunciation variant modelling for speech synthesis », ANR SynPaFlex, 50 % d'encadrement avec Damien Lolive (50 %).

Ingénieurs et experts (5 en cours)

— **Quentin Di-Fant :** ingénieur en informatique, projet Kaligo DYS, depuis novembre 2018, « Portage sur environnement mobile d'un système de synthèse de la parole », co-encadrement (50 %) avec Damien Lolive (50 %).
— **Simon Giddings :** ingénieur en informatique, projet SPAM, puis P2IA, depuis janvier 2019, « Portage sur environnement mobile d'un système de synthèse de la parole », co-encadrement (50 %) avec Damien Lolive (50 %).
— **Hassan Hajipoor :** ingénieur en informatique, projet Synthèse du breton, Kaligo et Synthèse du breton, depuis avril 2019, « Acquisition et traitement de ressources sonores », co-encadrement (50 %) avec Damien Lolive (50 %).
— **Pascal Lintanf :** expert linguiste, projet Synthèse du breton, depuis février 2020, « Production et analyse de ressources linguistiques du breton standard », co-encadrement (50 %) avec Damien Lolive (50 %).
— **Gaëlle Vidal :** ingénieure en acoustique, projet NADINE, Kaligo et Synthèse du breton, depuis avril 2019, « Acquisition et traitement de ressources sonores », co-encadrement (50 %) avec Damien Lolive (50 %).

ANNEXE C

Implications dans des projets

Kaligo DYS

— **Titre :** Aide à l'apprentissage de l'écriture pour des enfants souffrant de troubles DYS
— **Financement :** Région Bretagne, FEDER, Pôle Images & Réseaux
— **Dates :** 2018-2020
— **Rôle :** responsable local
— **Partenaires :** Learn&Go (coordinateur), IRISA, LOUSTIC, KARDI, HOALI
— **Budget :** 100K€

NADINE

— **Titre :** *Digital integrated system for the social support of migrants and refugees*
— **Financement :** H2020
— **Dates :** 2018-2021
— **Rôle :** participant
— **Partenaires :** Script&Go/Learn&Go (coordinateur), IRISA, ASPIRE-IGEN (Royaume-Uni), Caritas Hellas (Grèce), CERTH (Grèce), Cibervoluntarios Foundation (Espagne), IntraSoft International (Luxembourg), ISON Psychometrica (Grèce), Pluriversum (Italie), Odyssea (Grèce), VVA Group (Belgique)
— **Budget :** 4M€

P2IA (Projet d'Investissement Intelligence Artificielle)

— **Titre :** Aide à l'apprentissage de l'écriture en cycle 2
— **Financement :** Ministère de l'Éducation nationale
— **Dates :** 2019-2021
— **Rôle :** responsable local
— **Partenaires :** Learn&Go (coordinateur), Académie de Caen, Académie de Rennes, HOALI, INSPÉ Bretagne, IRISA, KARDI, LOUSTIC
— **Budget :** 1.5M€

SynPaFlex

— **Titre :** Synthèse de la parole flexible
— **Financement :** ANR

— **Dates :** 2015-2019
— **Rôle :** responsable de tâche
— **Partenaires :** IRISA (coordinateur), LLF, ATILF
— **Budget :** 250K€

Synthèse du breton

— **Titre :** Développement d'outils et de données pour la synthèse du breton
— **Financement :** Office public de la langue bretonne
— **Dates :** 2019-2020
— **Rôle :** coordinateur adjoint
— **Partenaires :** IRISA (coordinateur), Skol Vreizh
— **Budget :** 200K€

TAO-CSR

— **Titre :** *Task adaptation and optimisation for conversational speech recognition*
— **Financement :** Commission pour la Technologie et l'Innovation (CTI, Suisse)
— **Dates :** 2011-2012
— **Rôle :** participant
— **Partenaires :** Idiap Research Institute (Suisse), Koemei SA (Suisse)
— **Budget :** 100K€

TextToKids (ANR)

— **Titre :** Accès au contenu informationnel de textes par les enfants
— **Financement :** ANR
— **Dates :** 2019-2023
— **Rôle :** coordinateur adjoint, responsable local
— **Partenaires :** MoDyCo (coordinateur), IRISA, Qwant, Synapse Développement, Le P'tit Libé
— **Budget :** 650K€

TextToKids (CNRS)

— **Titre :** Aide rédactionnelle pour le récit d'événements à des enfants
— **Financement :** CNRS (PEPS S2IH INS2I)
— **Dates :** 2018
— **Rôle :** coordinateur
— **Partenaires :** IRISA (coordinateur), MoDyCo, LLing
— **Budget :** 10K€

TREMoLo

— **Titre :** Transformation de registres de langue par extraction de motifs langagiers
— **Financement :** ANR
— **Dates :** 2017-2021
— **Rôle :** coordinateur
— **Partenaires :** IRISA (coordinateur), MoDyCo
— **Budget :** 250K€

Bibliographie

ADELL, J., BONAFONTE, A. & ESCUDERO, D. (2007). Filled pauses in speech synthesis : towards conversational speech. In *Proceedings of Text, Speech and Dialogue (TSD)*.

ADELL, J., BONAFONTE, A. & MANCEBO, D. E. (2008). On the generation of synthetic disfluent speech : local prosodic modifications caused by the insertion of editing terms. In *Proceedings of Annual Conference of the International Speech Communication Association (Interspeech)*.

ADELL, J., ESCUDERO, D. & BONAFONTE, A. (2012). Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. *Speech Communication*, *54*.

AGRAWAL, R., SRIKANT, R. et al. (1995). Mining sequential patterns. In *Proceedings of the International Conference on Data Engineering* (T. 95).

AGUERT, M., BERNICOT, J. & LAVAL, V. (2009). Prosodie et compréhension des énoncés chez les enfants de 5 à 9 ans. *Enfance*, *2009*.

ALAIN, P., CHEVELU, J., GUENNEC, D., **LECORVÉ**, G. & LOLIVE, D. (2015). The IRISA Text-To-Speech System for the Blizzard Challenge 2015. In *Proceedings of the Blizzard Challenge 2015 Workshop*.

ALAIN, P., CHEVELU, J., GUENNEC, D., **LECORVÉ**, G. & LOLIVE, D. (2016). The IRISA Text-To-Speech System for the Blizzard Challenge 2016. In *Proceedings of the Blizzard Challenge 2016 Workshop*.

ALAIN, P., **LECORVÉ**, G., LOLIVE, D. & PERQUIN, A. (2018). The IRISA Text-To-Speech System for the Blizzard Challenge 2018. In *Proceedings of the Blizzard Challenge 2018 Workshop*.

ARGAMON, S. E. (2019). Register in computational language research. *Register Studies*, *1*(1).

BAHDANAU, D., CHO, K. & BENGIO, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*.

BAO, Y., ZHOU, H., HUANG, S., LI, L., MOU, L., VECHTOMOVA, O., ... CHEN, J. (2019). Generating Sentences from Disentangled Syntactic and Semantic Spaces. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

BATES, R. & OSTENDORF, M. (2002). Modeling pronunciation variation in conversational speech using prosody. In *Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*.

BÉCHET, F. (2001). LIA_PHON : un système complet de phonétisation de textes. *Traitement Automatique des Langues (TAL)*, *42*(1).

BÉCHET, N., CELLIER, P., CHARNOIS, T. & CRÉMILLEUX, B. (2015). Sequence mining under multiple constraints. In *Proceedings of the Annual ACM Symposium on Applied Computing*. ACM.

BELL, A. (1984). Language style as audience design. *Language in society*, *13*(2).

BETZ, S., WAGNER, P. & SCHLANGEN, D. (2015). Micro-Structure of Disfluencies : Basics for Conversational Speech Synthesis. *Proceedings of Annual Conference of the International Speech Communication Association (Interspeech)*.

BIBER, D. (1991). *Variation across speech and writing*. Cambridge University Press.

BIBER, D. & FINEGAN, E. (1994). *Sociolinguistic perspectives on register*. Oxford University Press on Demand.

BLANC, N. (2010). La compréhension des contes entre 5 et 7 ans : Quelle représentation des informations émotionnelles ? *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *64*(4).

BLANDIN, A. (2019). *Prédiction de recommandations d'âge pour l'accès à des enfants à des textes* (mém. de mast., Université de Rennes 1).

BLANDIN, A., **LECORVÉ**, G., BATTISTELLI, D. & ÉTIENNE, A. (2020a). Age recommendation for texts. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

BLANDIN, A., **LECORVÉ**, G., BATTISTELLI, D. & ÉTIENNE, A. (2020b). Recommendation d'âge pour des textes. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (JEP-TALN)*.

BORZEIX, A. & FRAENKEL, B. (2005). *Langage et travail (communication, cognition, action)*. CNRS éd.

BOULA DE MAREÜIL, P., HABERT, B., BÉNARD, F., ADDA-DECKER, M., BARRAS, C., ADDA, G. & PAROUBEK, P. (2005). A quantitative study of disfluencies in French broadcast interviews. In *Proceedings of the Disfluency in Spontaneous Speech Workshop*.

BOURGOIN, C., BATTISTELLI, D. & **LECORVÉ**, G. (2018). *Les notions temporelles dans la mise en récit d'événements dans le discours journalistique enfantin*. Rapport de recherche. IRISA, MoDyCo.

BRONCKART, J.-P. & BOURDIN, B. (1993). L'ACQUISITION DES VALEURS DES TEMPS DES VERBES : Etude comparative de l'allemand, du basque, du catalan, du français et de l'italien. *Langue française*, (97).

CALHOUN, S., CARLETTA, J., BRENIER, J. M., MAYO, N., JURAFSKY, D., STEEDMAN, M. & BEAVER, D. (2010). The NXT-format Switchboard Corpus : a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, *44*(4).

CARLETTA, J., EVERT, S., HEID, U. & KILGOUR, J. (2005). The NITE XML Toolkit : Data Model and Query Language. *Language Resources and Evaluation*, *39*(4).

CHARAUDEAU, P. (1997). *Le discours d'information médiatique : la construction du miroir social*. Nathan.

CHEN, K. & HASEGAWA-JOHNSON, M. (2004). Modeling pronunciation variation using artificial neural networks for English spontaneous speech. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*.

CHEVELU, J., **LECORVÉ**, G. & LOLIVE, D. (2014a). ROOTS : un outil pour manipuler facilement, efficacement et avec cohérence des corpus annotés de séquences. In *Journées d'Etude sur la Parole (JEP)*.

CHEVELU, J., **LECORVÉ**, G. & LOLIVE, D. (2014b). ROOTS : a toolkit for easy, fast and consistent processing of large sequential annotated data collections. In *Proceedings of Language Resources and Evaluation Conference (LREC)*.

CLARK, H. H. (1996). *Using Language*. Cambridge University Press.

CLARK, H. H. (2002). Speaking in time. *Speech Communication*, *36*.

CREISSEN, S. & BLANC, N. (2017). Quelle représentation des différentes facettes de la dimension émotionnelle d'une histoire entre l'âge de 6 et 10 ans ? Apports d'une étude multimédia. *Psychologie Française*.

CRESSOT, M. & GALLO, L. (1969). *Le style et ses techniques : précis d'analyse stylistique*. FeniXX.

CROWHURST, M. (1987). Cohesion in argument and narration at three grade levels. *Research in the Teaching of English*.

CUNNINGHAM, H., MAYNARD, D. & BONTCHEVA, V., K.and Tablan. (2002). GATE : an architecture for development of robust HLT applications.

DABBEBI, I. (2015). *Emerging Pattern Mining to Characterize Language Registers in French* (mém. de mast., Université de Tunis).

DALE, E. & CHALL, J. S. (1948). A formula for predicting readability : Instructions. *Educational research bulletin*.

DALL, R., TOMALIN, M., WESTER, M., BYRNE, W. J. & KING, S. (2014). Investigating automatic & human filled pause insertion for speech synthesis. In *Proceedings of Annual Conference of the International Speech Communication Association (Interspeech)*.

DAVIDSON, D. (2006). The role of basic, self-conscious and self-conscious evaluative emotions in children's memory and understanding of emotion. *Motivation and Emotion*, *30*(3).

DE BELDER, J. & MOENS, M.-F. (2010). Text simplification for children. In *Proceedings of the SIGIR workshop on accessible search systems*.

DONG, G. & PEI, J. (2007). *Sequence data mining*. Springer Science & Business Media.

DROIT-VOLET, S. (2000). L'estimation du temps : perspective développementale. *L'Année psychologique*, *100*(3).

ECKERT, P. & RICKFORD, J. R. (2001). *Style and sociolinguistic variation*. Cambridge University Press.

ÉTIENNE, A. (2019). *Compréhension de textes par les enfants et émotions : point(s) de vue psycholinguistique(s) et leur mise en œuvre en TAL* (mém. de mast., Université de Paris-Nanterre).

ÉTIENNE, A., BATTISTELLI, D. & **LECORVÉ**, G. (2020a). Apports de la linguistique et du TAL à l'analyse des émotions dans les textes pour enfants. In *Actes de colloque Langage et éMOTions*.

ÉTIENNE, A., BATTISTELLI, D. & **LECORVÉ**, G. (2020b). L'expression des émotions dans les textes pour enfants : constitution d'un corpus annoté. In *Actes de la Conférences sur le Traitement Automatique du Langage Naturel (TALN)*.

FAVART, M. (2005). Les marques de cohésion : leur rôle fonctionnel dans l'acquisition de la production écrite de texte. *Psychologie française*, *50*(3).

FAYET, C., BLOND, A., COULOMBEL, G., SIMON, C., LOLIVE, D., **LECORVÉ**, G., ... LE MAGUER, S. (2020). FlexEval, création de sites web légers pour des campagnes de tests perceptifs multimédias. In *Actes de la conférence conjointe sur le Traitement Automatique du Langage Naturel et des Journées d'Étude sur la Parole (JEP-TALN) – session démo*.

FERGUSON, C. A. (1982). Simplified registers and linguistic theory. *Exceptional language and linguistics*.

FERRUCCI, D. & LALLY, A. (2004). UIMA : an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, *10*(3-4).

FERRUCCI, D., LALLY, A., GRUHL, D., EPSTEIN, E., SCHOR, M., MURDOCK, J. W., ... DOGANATA, Y. et al. (2006). Towards an interoperability standard for text and multi-modal analytics. *IBM Research Report*.

FLESCH, R. (1948). A new readability yardstick. *Journal of applied psychology*, *32*(3).

FOSLER-LUSSIER, E. & MORGAN, N. (1999). Effects of speaking rate and word frequency on pronunciations in convertional speech. *Speech Communication*, *29*.

FOWLER, C. A. & HOUSUM, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, *26*.

FRANÇOIS, T. (2015). When readability meets computational linguistics : a new paradigm in readability. *Revue française de linguistique appliquée*, *20*(2).

FRANÇOIS, T. & FAIRON, C. (2012). An "AI readability" Formula for French as a Foreign Language. In *Proceedings of the joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

FRITH, U. (1985). Beneath the surface of developmental dyslexia. *In K. E. Patterson, J. C. Marshall, & M. Colthear (Eds.),Surface Dyslexia : Neuropsychological and Cognitive Studies of Phonological Reading*.

GADET, F. (1996a). Niveaux de langue et variation intrinsèque. *Palimpsestes*, *10*.

GADET, F. (1996b). Variabilité, variation, variété : le français d'Europe. *Journal of French Language Studies*, *6*(1).

GALA, N., FRANCOIS, T., JAVOUREY-DREVET, L. & ZIEGLER, J. C. (2018). Text simplification, a tool for learning to read. *Langue française*, (199).

GATHERCOLE, S. (1999). Cognitive approaches to the development of short-term memory. *Trends in cognitive sciences*, *3*.

GEYER, L. H. (1977). Recognition and confusion of the lowercase alphabet. *Perception & Psychophysics*, *22*(5).

GRAVIER, G., GUINAUDEAU, C., **LECORVÉ**, G. & SÉBILLOT, P. (2011). Exploiting Speech for Automatic TV Delinearization : From Streams to Cross-Media Semantic Navigation. *EURASIP Journal on Image and Video Processing*, *2011*.

GROMER, D. & WEISS, M. (1990). *Lire, tome 1 : apprendre à lire*. Armand Colin.

HASSAN, H., SCHWARTZ, L., HAKKANI-TÜR, D. & TÜR, G. (2014). Segmentation and disfluency removal for conversational speech translation. In *Proceedings of Annual Conference of the International Speech Communication Association (Interspeech)*.

HICKMANN, M. (2012). Diversité des langues et acquisition du langage : espace et temporalité chez l'enfant. *Langages*, (4).

HU, Z., YANG, Z., LIANG, X., SALAKHUTDINOV, R. & XING, E. P. (2017). Controllable Text Generation. *CoRR*, *abs/1703.00955*. arXiv : 1703.00955

HUET, S., **LECORVÉ**, G., GRAVIER, G. & SÉBILLOT, P. (2008). Toward the Integration of Natural Language Processing and Automatic Speech Recognition : Using Morpho-Syntax and Pragmatics for Transcription. In P. MARAGOS, A. POTAMIANOS & P. GROS (Éd.), *Multimodal Processing and Interaction : Audio, Video, Text*. Springer US.

HUNT, A. J. & BLACK, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (T. 1).

ILLINA, I., FOHR, D. & JOUVET, D. (2011). Grapheme-to-Phoneme Conversion using Conditional Random Fields. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*.

ILMOLA, M. (2012). Les registres familier, populaire et vulgaire dans Le Canard enchaîné et Charlie Hebdo : étude comparative.

IMSENG, D., BOURLARD, H., CAESAR, H., GARNER, P. N., **LECORVÉ**, G. & NANCHEN, A. (2012). MediaParl : Bilingual mixed language accented speech database. In *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE.

IYYER, M., WIETING, J., GIMPEL, K. & ZETTLEMOYER, L. (2018). Adversarial Example Generation with Syntactically Controlled Paraphrase Networks. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT)*.

JIAMPOJAMARN, S., KONDRAK, G. & SHERIF, T. (2007). Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion. In *Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT)*.

JOHN, V., MOU, L., BAHULEYAN, H. & VECHTOMOVA, O. (2018). Disentangled representation learning for text style transfer. *arXiv preprint arXiv :1808.04339*.

JOHN, V., MOU, L., BAHULEYAN, H. & VECHTOMOVA, O. (2019). Disentangled Representation Learning for Non-Parallel Text Style Transfer. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

JOHNSTONE, B. (2009). Stance, style, and the linguistic individual. *Stance : sociolinguistic perspectives*.

KAUSHIK, M., TRINKLE, M. & HASHEMI-SAKHTSARI, A. (2010). Automatic detection and removal of disfluencies from spontaneous speech. In *Proceedings of the Australasian International Conference on Speech Science and Technology (SST)*.

KIKUCHI, Y., NEUBIG, G., SASANO, R., TAKAMURA, H. & OKUMURA, M. (2016). Controlling Output Length in Neural Encoder-Decoders. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

LASSELIN, H. (2018). *Make text look like speech : disfluency generation using sequence-to-sequence neural networks* (Rapport de recherche, Rapport de recherche. IRISA).

LECORVÉ, G. (2007). *Adaptation thématique d'un système de transcription automatique de la parole* (mém. de mast., INSA Rennes).

LECORVÉ, G. (2008). *Adaptation thématique non supervisée d'un système de reconnaissance automatique de la parole* (thèse de doct.).

LECORVÉ, G., AYATS, H., FOURNIER, B., MEKKI, J., CHEVELU, J., BATTISTELLI, D. & BÉCHET, N. (2018). Construction conjointe d'un corpus et d'un classifieur pour les registres de langue en français. In *Traitement automatique du langage naturel (TALN)*.

LECORVÉ, G., AYATS, H., FOURNIER, B., MEKKI, J., CHEVELU, J., BATTISTELLI, D. & BÉCHET, N. (2019). Towards the Automatic Processing of Language Registers : Semi-supervisedly Built Corpus and Classifier for French. In *International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*.

LECORVÉ, G., DINES, J., HAIN, T. & MOTLICEK, P. (2012a). Impact du degré de supervision sur l'adaptation à un domaine d'un modèle de langage à partir du Web. In *Actes des Journées d'Études sur la Parole et de la conférences sur le Traitement Automatique du Langage Naturel (JEP-TALN)*.

LECORVÉ, G., DINES, J., HAIN, T. & MOTLICEK, P. (2012b). Supervised and unsupervised Web-based language model domain adaptation. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*.

LECORVÉ, G., GRAVIER, G. & SÉBILLOT, P. (2008a). An unsupervised web-based topic language model adaptation method. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

LECORVÉ, G., GRAVIER, G. & SÉBILLOT, P. (2008b). On the Use of Web Resources and Natural Language Processing Techniques to Improve Automatic Speech Recognition Systems. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

LECORVÉ, G., GRAVIER, G. & SÉBILLOT, P. (2008c). Vers une adaptation thématique non supervisée de modèles de langage : utilisation d'Internet comme un corpus ouvert. In *Journées d'Étude sur la Parole et Conférence sur le Traitement Automatique du Langage Naturel (JEP-TALN)*.

LECORVÉ, G., GRAVIER, G. & SÉBILLOT, P. (2009). Constraint selection for topic-based MDI adaptation of language models. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*.

LECORVÉ, G., GRAVIER, G. & SÉBILLOT, P. (2010). L'adaptation thématique d'un modèle de langue fait-elle apparaître des mots thématiques ? In *Actes des Journées d'Étude sur la Parole (JEP)*.

LECORVÉ, G., GRAVIER, G. & SÉBILLOT, P. (2011). Automatically Finding Semantically Consistent N-grams to Add New Words in LVCSR Systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

LECORVÉ, G. & LOLIVE, D. (2015). Adaptive statistical utterance phonetization for French. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

LECORVÉ, G. & LOLIVE, D. (2016). Phonétisation statistique adaptable d'énoncés pour le français. In *Actes des Journées d'Études sur la Parole (JEP)*.

LECORVÉ, G. & MOTLICEK, P. (2012). Conversion of recurrent neural network language models to weighted finite state transducers for automatic speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*.

LEEFTINK, W. & SPANAKIS, G. (2019). Towards controlled transformation of sentiment in sentences. *arXiv preprint arXiv :1901.11467*.

LEVELT, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, *14*.

LI, J., JIA, R., HE, H. & LIANG, P. (2018). Delete, Retrieve, Generate : A Simple Approach to Sentiment and Style Transfer. *CoRR*, *abs/1804.06437*. arXiv : 1804.06437

LIEBERMAN, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, *6*.

LIU, Y., SHRIBERG, E., STOLCKE, A., HILLARD, D., OSTENDORF, M. & HARPER, M. (2006). Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*.

LOLIVE, D., ALAIN, P., BARBOT, N., CHEVELU, J., **LECORVÉ**, G., SIMON, C. & TAHON, M. (2017). The IRISA Text-To-Speech System for the Blizzard Challenge 2017. In *Proceedings of the Blizzard Challenge Workshop*.

MEKKI, J., BATTISTELLI, D., BÉCHET, N. & **LECORVÉ**, G. (2017). « *Nous nous arrachâmes promptement avec ma caisse* » *: quels descripteurs linguistiques caractérisent les registres de langue ?* Rapport de recherche. IRISA, MoDyCo.

MEKKI, J., BATTISTELLI, D., **LECORVÉ**, G. & BÉCHET, N. (2018). Identification de descripteurs pour la caractérisation de registres. In *Actes des Rencontres Jeunes Chercheurs (RJC) de la conférence CORIA-TALN*.

MEKKI, J., BÉCHET, N., BATTISTELLI, D. & **LECORVÉ**, G. (2020). Caractérisation de registres de langue par extraction de motifs séquentiels émergents. In *Actes des Journées Internationales d'Analyse statistique des Données Textuelles (JADT)*.

MERRITT, T., CLARK, R. A., WU, Z., YAMAGISHI, J. & KING, S. (2016). Deep neural network-guided unit selection synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

MICHELI, R. (2014). *Les émotions dans les discours : modèle d'analyse et perspectives empiriques*. De Boeck.

MIKOLOV, T., KARAFIAT, M., BURGET, L., ČERNOCK, J. & KHUDANPUR, S. (2010). Recurrent Neural Network Based Language Model. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*.

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S. & DEAN, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. BURGES, L. BOTTOU, M. WELLING, Z. GHAHRAMANI & K. Q. WEINBERGER (Éd.), *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*. Curran Associates, Inc.

MOHRI, M., PEREIRA, F. C. & RILEY, M. (2008). Speech Recognition with Weighted Finite-State Transducers. *Springer Handbook of Speech Processing*.

MOIRAND, S. (2007). *Les discours de la presse quotidienne. Observer, analyser, comprendre*. Puf.

MOUW, J. M., VAN LEIJENHORST, L., SAAB, N., DANEL, M. S. & van den BROEK, P. (2019). Contributions of emotion understanding to narrative comprehension in children and adults. *European Journal of Developmental Psychology*, *16*(1).

NEY, H. & ORTMANNS, S. (1999). Dynamic programming search for continuous speech recognition. *IEEE Signal Processing Magazine*.

PÉRENNOU, G. & DE CALMES, M. (2000). MHATLex : Lexical Resources for Modelling the French Pronunciation. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

PERQUIN, A. (2017). *Big deep voice : indexation de données massives de parole grâce à des réseaux de neurones profonds* (mém. de mast., University of Rennes 1).

PERQUIN, A., **LECORVÉ**, G., LOLIVE, D. & AMSALEG, L. (2018). Phone-Level Embeddings for Unit Selection Speech Synthesis. In *Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP)*. Springer.

PERQUIN, A., **LECORVÉ**, G., LOLIVE, D. & AMSALEG, L. (2019). Évaluation objective de plongements pour la synthèse de parole guidée par réseaux de neurones. In *Actes de la conférences sur le Traitement automatique du langage naturel (TALN)*.

PETERS, M. E., NEUMANN, M., IYYER, M., GARDNER, M., CLARK, C., LEE, K. & ZETTLEMOYER, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv :1802.05365*.

PING, W., PENG, K., GIBIANSKY, A., ARIK, S. O., KANNAN, A., NARANG, S., ... MILLER, J. (2018). Deep Voice 3 : 2000-Speaker Neural Text-to-Speech. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

PIOLAT, A. & BANNOUR, R. (2009). An example of text analysis software (EMOTAIX-Tropes) use : The influence of anxiety on expressive writing. *Current psychology letters. Behaviour, brain & cognition*, *25*(2, 2009).

PITT, M. A., JOHNSON, K., HUME, E., KIESLING, S. & RAYMOND, W. (2005). The Buckeye corpus of conversational speech : labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*.

PRAKASH, A., HASAN, S. A., LEE, K., DATLA, V., QADIR, A., LIU, J. & FARRI, O. (2016). Neural Paraphrase Generation with Stacked Residual LSTM Networks. In *Proceedings of the International Conference on Computational Linguistics (COLING) : Technical Papers*.

QADER, R. (2017). *Pronunciation and disfluency modeling for expressive speech synthesis* (thèse de doct., University of Rennes 1).

QADER, R., **LECORVÉ**, G., LOLIVE, D. & SÉBILLOT, P. (2014). *Phonology Modelling for Expressive Speech Synthesis : a Review* (Rapport de recherche N° PI-2020). Rapport de recherche. IRISA.

QADER, R., **LECORVÉ**, G., LOLIVE, D. & SÉBILLOT, P. (2015). Probabilistic Speaker Pronunciation Adaptation for Spontaneous Speech Synthesis Using Linguistic Features. In *Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP)*.

QADER, R., **LECORVÉ**, G., LOLIVE, D. & SÉBILLOT, P. (2016). Adaptation de la prononciation pour la synthèse de la parole spontanée en utilisant des informations linguistiques. In *Actes des Journées d'Études sur la Parole (JEP)*.

QADER, R., **LECORVÉ**, G., LOLIVE, D. & SÉBILLOT, P. (2017). Ajout automatique de disfluences pour la synthèse de la parole spontanée : formalisation et preuve de concept. In *Traitement automatique du langage naturel (TALN)*. **Prix du meilleur article**.

QADER, R., **LECORVÉ**, G., LOLIVE, D. & SÉBILLOT, P. (2018). Disfluency Insertion for Spontaneous TTS : Formalization and Proof of Concept. In *Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP)*. Springer.

QADER, R., **LECORVÉ**, G., LOLIVE, D., TAHON, M. & SÉBILLOT, P. (2017). Statistical pronunciation adaptation for spontaneous speech synthesis. In *Proceedings of the International Conference on Text, Speech, and Dialogue (TSD)*. Springer.

QUARTIER, V. (2008). Le développement de la temporalité : théorie et instrument de mesure du temps notionnel chez l'enfant. *Approche Neuropsychologique des Apprentissages chez l'Enfant*, *20*(100).

RAÏSSI, C. & PONCELET, P. (2007). Sampling for sequential pattern mining : From static databases to data streams. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. IEEE.

REN, Y., RUAN, Y., TAN, X., QIN, T., ZHAO, S., ZHAO, Z. & LIU, T.-Y. (2019). Fastspeech : Fast, robust and controllable text to speech. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

RIFFATERRE, M. (1961). Vers la définition linguistique du style. *Word*, *17*(2).

ROSE, R. L. (1998). *The communicative value of filled pauses in spontaneous speech* (thèse de doct., University of Birmingham).

SANDERS, C. (1993). *Sociosituational variation*. Cambridge : Cambridge University Press.

SEBEOK, T. A. (1960). Style in language.

SHEN, J., PANG, R., WEISS, R. J., SCHUSTER, M., JAITLY, N., YANG, Z., ... SKERRV-RYAN, R. et al. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

SHEN, T., LEI, T., BARZILAY, R. & JAAKKOLA, T. (2017). Style Transfer from Non-Parallel Text by Cross-Alignment. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT (Éd.), *Proceedings of Advances in Neural Information Processing Systems (NIPS)*. Curran Associates, Inc.

SHRIBERG, E. E. (1994). *Preliminaries to a theory of speech disfluencies* (thèse de doct., University of California).

SHRIBERG, E. E. (1999). *Phonetic consequences of speech disfluency*. DTIC Document.

STOLCKE, A. & SHRIBERG, E. (1996). Statistical language modeling for speech disfluencies. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

STOLCKE, A., SHRIBERG, E., BATES, R. A., OSTENDORF, M., HAKKANI, D., PLAUCHE, M., ... LU, Y. (1998). Automatic detection of sentence boundaries and disfluencies based on recognized words. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*.

SUNDARAM, S. & NARAYANAN, S. (2003). An empirical text transformation method for spontaneous speech synthesizers. In *Proceedings of Annual Conference of the International Speech Communication Association (Interspeech)*.

SUTSKEVER, I., VINYALS, O. & LE, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of Advances in neural information processing systems (NIPS)*.

TAHON, M., **LECORVÉ**, G. & LOLIVE, D. (2018). Can we Generate Emotional Pronunciations for Expressive Speech Synthesis ? *IEEE Transactions on Affective Computing*.

TAHON, M., **LECORVÉ**, G., LOLIVE, D. & QADER, R. (2017). Perception of expressivity in TTS : linguistics, phonetics or prosody ? In *Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP)* (T. 10583). Springer.

TAHON, M., QADER, R., **LECORVÉ**, G. & LOLIVE, D. (2016a). Improving TTS with corpus-specific pronunciation adaptation. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*.

TAHON, M., QADER, R., **LECORVÉ**, G. & LOLIVE, D. (2016b). Optimal feature set and minimal training size for pronunciation adaptation in TTS. In *Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP)*. Springer.

TARTAS, V. (2010). Le développement de notions temporelles par l'enfant. *Développements*, *4*.

TIKHONOV, A. & YAMSHCHIKOV, I. P. (2018). What is wrong with style transfer for texts ? *arXiv preprint arXiv :1808.04365*.

TOPARLAK, T., DONABÉDIAN, A., LOLIVE, D., **LECORVÉ**, G. & DELAIS-ROUSSARIE, E. (2019). Synthèse vocale de l'arménien. Présenté à Digital Armenian.

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*.

TREE, J. E. F. (2001). Listeners' uses ofum anduh in speech comprehension. *Memory & cognition*, *29*.

TSENG, S.-C. (1999). Grammar, prosody and speech disfluencies in spoken dialogues. *Unpublished doctoral dissertation. University of Bielefeld*.

URE, J. (1982). Introduction : approaches to the study of register range. *International Journal of the Sociology of Language*, *1982*(35).

van den OORD, A., DIELEMAN, S., ZEN, H., SIMONYAN, K., VINYALS, O., GRAVES, A., ... KAVUKCUOGLU, K. (2006). WaveNet : A Generative Model for Raw Audio. In *Proceedings of the ISCA Speech Synthesis Workshop (SSW)*.

VION, M. & COLAS, A. (1999). L'emploi des connecteurs en français : contraintes cognitives et développement des compétences narratives (le cas de la narration de séquences arbitraires d'événements). In *Proceedings of the Conference of the International Association for the Study of Child Language*.

WAN, V., AGIOMYRGIANNAKIS, Y., SILEN, H. & VIT, J. (2017). Google's next-generation real-time unit-selection synthesizer using sequence-to-sequence lstm-based autoencoders. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*.

WANG, D. & KING, S. (2011). Letter-to-sound pronunciation prediction using conditional random fields. *IEEE Signal Processing Letters*, *18*(2).

WIDLÖCHER, A. & MATHET, Y. (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. In *Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN'09), session posters*.

WIGHTMAN, C. W. & OSTENDORF, M. (1994). Automatic labeling of prosodic patterns. *Transactions on Speech and Audio Processing*, *2*.

WU, Z. & KING, S. (2016). Improving trajectory modelling for dnn-based speech synthesis by using stacked bottleneck features and minimum generation error training. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, *24*(7).

YAN, X. & AL. (2003). CloSpan : Mining : Closed sequential patterns in large datasets. In *Proceedings of the 2003 SIAM*. SIAM.

ZHANG, H., SPROAT, R., NG, A. H., STAHLBERG, F., PENG, X., GORMAN, K. & ROARK, B. (2019). Neural models of text normalization for speech applications. *Computational Linguistics*, *45*(2).

ZHANG, J., PAN, J., YIN, X., LI, C., LIU, S., ZHANG, Y., ... MA, Z. (2020). A hybrid text normalization system using multi-head self-attention for mandarin. In *Proceedings of ICASSP*.

ZHANG, Y., CHEN, X. et al. (2020). Explainable Recommendation : A Survey and New Perspectives. *Foundations and Trends® in Information Retrieval*, *14*(1).

ZHAO, S., MENG, R., HE, D., ANDI, S. & BAMBANG, P. (2018). Integrating transformer and paraphrase rules for sentence simplification. In *Proceedings of the ACL*.

Résumé

Historiquement, le traitement automatique du langage naturel (TALN) s'est majoritairement concentré sur l'accès au sens des énoncés (écrits ou oraux). Pourtant, de multiples autres dimensions liées au style, c'est-à-dire à la manière dont ces énoncés sont réalisés, apportent des informations complémentaires pour comprendre leur contexte de communication (public visé, relation entre 2 interlocuteurs, état émotionnel ou niveau socio-culturel d'un auteur/locuteur, modalité...). Le traitement du style est donc un enjeu important pour rendre les applications d'intelligence artificielle plus "humaines", notamment lorsque ces applications doivent produire des énoncés pour des utilisateurs.
Cette habilitation à diriger des recherches est une synthèse de mes activités de recherche au cours de ces dix dernières années, essentiellement dans le cadre de l'équipe Expression, sur les questions de variabilité et de style en TALN. Nous y traitons la notion de style comme le recours cohérent au sein d'un énoncé à différents traits linguistiques spécifiques (choix de certains mots, certaines structures syntaxiques, prononciations...). Nous cherchons principalement à caractériser ces traits, puis à savoir les introduire dans des énoncés afin de leur conférer un style voulu. Ces problématiques sont déclinées sur des instances particulières de styles (parole spontanée, formalité, langage pour les enfants) tout comme dans le cadre d'une conception plus générique de cette notion. Au delà des différentes contributions, ces deux angles d'approche nous amènent à questionner les places respectives de l'apprentissage automatique et des sciences humaines dans la perspective d'un meilleur traitement automatique du style.

Abstract

Historically, natural language processing (NLP) has mainly focused on accessing the meaning of written or spoken contents. However, multiple other dimensions related to style, i.e. the way in which statements are realized, provide additional information to understand their underlying communication context (target audience, relationship between two interlocutors, emotional state or socio-cultural level of an author/speaker, modality...). Style processing is therefore an important issue to make artificial intelligence applications more "human", especially when these applications must generate linguistic contents for end-users.
This HDR summarizes my research activities over the last ten years, mostly within the Expression team, on the issues of variability and style in NLP. We consider the notion of style as the consistent use in texts or speech of various specific linguistic features (particular words, syntactic structures, pronunciations, etc.). Mainly, we seek to characterize these features, and then to introduce them into new sentences or utterances in order to mimic a desired style. These problems are declined on several instances of style (spontaneous speech, formality, language for children) as well as through a more generic conception of it. Beyond the different contributions, these two approaches lead us to question the respective roles of machine learning and human sciences towards a better processing of style in the future.