Secteur des sciences humaines
Faculté de Philosophie, Arts et Lettres

EFFECTS OF COGNITIVE LOAD ON SPEECH PRODUCTION AND PERCEPTION

GEORGE CHRISTODOULIDES

Thèse soutenue en vue de l'obtention
du grade de Docteur en Langues et Lettres

Membres du jury
Prof. Michel Francard (Université catholique de Louvain), président
Prof. Anne Catherine Simon (Université catholique de Louvain), promotrice
M.E.R. Antoine Auchlin (Université de Genève)
Dr Céline De Looze (Trinity College Dublin)
Prof. Piet Mertens (Katholieke Universiteit Leuven)
Prof. Arnaud Szmalec (Université catholique de Louvain)

September 2016

George Christodoulides: *Effects of Cognitive Load on Speech Production and Perception,* PhD Thesis, Université catholique de Louvain, September 2016

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ANN	Artificial Neural Network
AP	Accentual Phrase
ASR	Automatic Speech Recognition
BDU	Basic Discourse Unit
BIC	Bayesian Information Criterion
CART	Classification And Regression Tree
CL	Cognitive Load
CoNLL	Conference on Natural Language Learning
ConTRe	Continuous Tracking and Reaction
CPP	Cepstral Peak Prominence
CRF	Conditional Random Field
DM	Discourse Marker
DU	Dependency Unit
ECG	Electrocardiography
EEG	Electroencephalography
EGG	Electroglottography
f_0	Fundamental Frequency
F1, F2, F3	First, second, third Formant
FM	Frequency Modulation
GMM	Gaussian Mixture Model
GSR	Galvanic Skin Response
HCI	Human-Computer Interaction
HIPS	Human Information Processing Stages
HMM	Hidden Markov Models
ILP	Integer Linear Programming
IP, IU	Intonation Phrase, Intonation Unit
L1	First Language (mother tongue)
L2	Second Language (acquired)
LOCAS-F	Louvain Corpus of Annotated Speech (French)
LTM	Long-Term Memory

MFCC	Mel-Frequency Cepstral Coefficients
MLLR	Maximum Likelihood Linear Regression
NASA-TLX	NASA Task Load Index
PB	Prosodic Boundary
PFC	Phonologie du Français contemporain
POS	Part-of-Speech
PPB	Perceived Prosodic Boundary
RF	Random Forest
SI	Simultaneous Interpreting
SQL	Structured Query Language
STM	Short-Term Memory
STSS	Short-Term Sensory Store
SVM	Support Vector Machine
SUS	Syntactically Unpredictable Sentences
TERP	Task-Evoked Pupillary Response
WEAVER++	Word Encoding by Activation and VERification
WM	Working Memory

ABSTRACT

The objective of this thesis is to study the effects of cognitive load on the production and perception of speech, and especially the prosodic characteristics of French speech produced under high levels of cognitive load. Cognitive load reflects the mental demand placed by a task on the person performing it, and is derived from the limited capacity of cognitive systems, such as working memory and attention. Speech production (conceptualisation of a message, formulation and articulation) and speech perception and comprehension are processes that engage cognitive resources, including working memory, to a different degree. It is expected that situations of high cognitive load will cause detectable effect in the prosodic structuring of speech.

The main hypothesis is that the effects of cognitive load on prosody will be detected in the temporal organisation of speech and the segmentation (phrasing) of utterances. It is hypothesised that limitations of working memory will affect speech planning, leading to a marked increase in mismatches between prosodic and syntactic boundaries. We expect to find an increase in the number of occurrences of major prosodic boundaries inside minor syntactic units (chunks), i.e. in positions where there are normally not expected. On this basis, with respect to temporal prosodic measures, we expect to find an increase in the frequency of long pauses, and a more variable speech rate.

Four experimental studies were conducted. In Study 1, participants performed a Stroop naming task under dual-task and time pressure, and a Reading Span task. Speech recordings and electroglottographic (EGG) data were collected, in order to analyse voice features. Study 1 is a replication of the CLSE English corpus (Yap, 2012) for French. In Study 2, monologue speech was recorded and subjects were asked to memorise information of increasing complexity and answer reading comprehension questions. Study 3 focuses on simultaneous interpreting, a task that is cognitively demanding and involves both language comprehension and production. The prosodic characteristics of two versions of the same text were compared: the original output of a professional conference interpreter, and a rehearsed reading of it by the same person. In Study 4, cognitive load was induced through the Continuous Tracking and Reaction task in a driving simulator, while pairs of participants engaged in a series of tasks (memorisation and summarising, dialogue to exchange information, debate, repeating syntactically unpredictable sentences and a simple game).

Results indicate that under increased cognitive load, speakers produce more numerous and longer in duration silent pauses; there is also an increase in the variability of articulation rate (more accelerations and decelerations). We confirm our hypothesis that cognitive load incurs an increase in the num-

ber of occurrences of major prosodic boundaries inside minor syntactic units (chunks), and that silent and filled pauses are placed incongruently with the syntactic structure. However, while filled pause ratio (proportion of speech time spent on filled pauses) increases, some speakers do not produce more filled pauses but rather produce longer filled pauses or drawls (hesitation-related lengthening). There was no systematic change in the mean pitch of speakers under cognitive load; in some of the studies it was observed that high cognitive load tasks result in a decrease of mean pitch variance and pitch movements. EGG results indicate that the Closed Quotient measure increases as cognitive load increases.

Further contributions of this thesis include the development of software tools for the automatic annotation of French spoken corpora (DisMo: morpho-syntactic tagging, disfluency detection and annotation; Promise: syllabic prosodic prominence and prosodic boundary detection; a tool for extracting temporal measures from annotated dialogues), the development of a new tool for working with spoken corpora (Praaline), and the development of 3 corpora of French speech produced under cognitive load (approximately 50h in total).

ACKNOWLEDGMENTS

First and foremost I would like to express my gratitude to my doctoral supervisor, Anne Catherine Simon, for her continuous support and encouragement. I am indebted to her for her willingness and ability to share knowledge, and for the enthusiasm she has always shown whenever a new idea sprang up.

I would like to thank my thesis committee: Antoine Auchlin, Céline De Looze, Piet Mertens and Arnaud Szmalec. I am grateful for the advice you gave me at the turning points of this project over the past three years, and I would like to thank you for your detailed comments to the manuscript.

A special thank you goes to four colleagues with whom we shared many common projects (and drinks!): Mathieu Avanzi, Cédric Lenglet, Jean-Philippe Goldman and Giulia Barreca. Thank you for listening, for speaking out, and sometimes doing both at the same time!

I would also like to thank those who helped me in preparing parts of this thesis: Liesbeth Degand for her help in the analysis of Study 3, Philippe Boula de Mareüil for his help with the SUS corpus, and Alexandre Zénon for introducing me to eye-tracking and kindly lending me a Pupil device. I would also like to thank the members of the PFC project for our collaboration. Many thanks to Zeinab Traoré, Alice Panier, Agathe Pierson (and Giulia!) for their help in transcribing the corpus of Study 4.

I am grateful to the students of the course "Phonologie et Prosodie" during the academic years 2014-2015 and 2015-2016 who participated in the experimental studies and who convinced me that teaching is fun. And, of course, to the colleagues at Valibel, ILC and Cental (my Belgian-university family) for all the shared moments.

And finally, a big thank you to my family and friends who lived through this PhD!

1

INTRODUCTION

The objective of this thesis is to study the effects of cognitive load on the production and perception of speech, with a special emphasis on the prosodic characteristics of speakers under high levels of cognitive load. Cognitive load reflects the mental demand placed by a task on the person performing it, and is derived from the limited capacity of cognitive systems, especially working memory and attention. Speech production, consisting of the conceptualisation of a message, its formulation and articulation, and speech perception and comprehension are all processes that engage cognitive resources, including working memory, to different degrees. We can therefore expect that in situations of high cognitive load, there will be detectable effects on speech. The present thesis focuses on the effects of cognitive load in the prosody of French speech, a relatively unexplored area of research.

1.1 BACKGROUND AND MOTIVATION

Studying the effects of cognitive load on speech has interesting applications. There are several scenarios where people engage in a cognitively demanding activity and simultaneously produce and/or perceive speech. Typical examples include the use of voice-controlled interfaces in vehicles; the automatic detection of problematic interactions in call centres; the communication between air traffic control and pilots, and between crew members in aviation; the interaction in high-risk or high-stress situations; and tasks that are inherently cognitively demanding and require the production and perception of speech, such as simultaneous interpreting. The research findings related to speech under cognitive load can be applied in these scenarios. For example, a human-computer interface may be designed to be *adaptive*, i.e. modulate its output quantity and quality of information to avoid overly taxing the user. In high-risk or high-stress situations, such as aviation communication, a possible application is to detect when performers of tasks find themselves under excessive cognitive load, and try to alleviate the problem through training or by changing procedures. In simultaneous interpreting, speech markers of excessive cognitive load can be used to improve training material and techniques.

Furthermore, the study of speech production and perception under high cognitive load can advance our knowledge about speech production and perception in general. Taking a complex system (such as the human cognitive architecture that enables us to use language) to its limits and studying how it adapts or fails can help us construct better models of the system. In this respect, research on the prosody of speech under cognitive load is linked to the investigation of emotional speech, speech under stress, pathological speech, and speaker fluency. In this thesis, our work on speech under cognitive load has lead us to more general work, in order to improve the tools available for the automatic annotation and analysis of French spoken corpora.

1.2 MAIN OBJECTIVES AND HYPOTHESES

In order to attain our goal of studying speech production and perception under cognitive load in French, we set out the following main objectives:

- Review the literature on cognitive load, working memory and prosody, as well as previous studies performed with English-speaking participants, to identify possible effects and the relationship between cognitive load and speech prosody.
- Collect data and high-quality recordings of speech produced under cognitive load, in controlled experiments. There is a lack of publicly-available French corpora containing recordings of speech produced while the speaker is engaging in controlled tasks inducing cognitive load. In the context of this thesis, we constructed such corpora.
- Perform an exploratory analysis of the data collected, to identify effects of cognitive load on speech prosody. We mainly focus on the temporal organisation of speech (pauses, speech rate), pitch variability, disfluencies, and the interface between syntax and prosody.
- Interpret these findings in light of other research on the prosody of French.

The main hypothesis of this thesis is that cognitive load, conceptualised as the demands placed on the working memory and attentional resources of the speaker, will affect the prosodic characteristics of their speech, especially with respect to the temporal organisation of speech and the segmentation (phrasing) of utterances. The limitations of working memory will affect prosodic planning, and this will surface as an increase in the mismatches between the mapping of prosodic and syntactic units. We refine this general hypothesis, in a series of more specific hypothesis relating to aspects of prosody, after presenting the literature review, in Chapter 10 (Global Hypotheses).

1.3 HOW THIS THESIS IS ORGANISED

This thesis is organised in three parts: Part I – Theoretical Background, which presents a survey of the literature related to the thesis; Part II – Methodology, Tools and Baseline Studies, in which we outline the choices made in analysing spoken language data, as well as our work on the development of software tools to automate such analyses; and Part III – Experimental Studies, in which we present four studies on speech produced under cognitive load.

The THEORETICAL BACKGROUND part is organised as follows:

- In Chapter 2, we explore the literature on cognitive load in the field of psychology and ergonomics: Perceptual Load Theory (section 2.1), Cognitive Load Theory (section 2.2), Rasmussen's Levels of Behaviour (section 2.3) and the Human Information Processing Stages model (section 2.4). We operationalise the concept of cognitive load as a measure of the demands created by a task on the attentional and working memory resources of the performer.

- Chapter 3, therefore, presents models of attention and working memory from the field of cognitive psychology. From the first studies that led to the definition of the concept of working memory (section 3.1), to three influential models: Baddeley's WM model (section 3.2), Ericsson & Kintsch's long-term WM model (section 3.3) and Cowan's model of activated long-term memory (section 3.4). In Chapters 5 and 6 we discuss the relationship between working memory and the production and perception of language.

- Chapter 4 introduces the second main theme of the present thesis, prosody. We present the basic dimensions of speech prosody and the functions of prosody in communication. This chapter gives an overview of the subject, while Part II contains more details on the specific domains of prosody explored in this thesis.

- Chaptter 5 reviews models of language production (section 5.1) and explores the role of working memory in language production (section 5.2), as well as the related concept of automaticity. We particularly focus on studies on self-monitoring and repair strategies during production (5.3) and the role of prosody in signalling discourse structure (5.4), as these two aspects of production are explored in the experimental studies.

- Chapter 6 on speech perception and language comprehension has a structure parallel to the previous chapter: after reviewing the relationship between auditory cognition and speech perception (section 6.1), we present findings related to the basic mechanisms of language comprehension (section 6.2). We then explore the role of working memory

and automaticity in language comprehension (section 6.3), and the role of prosody in signalling segmentation and structure (section 6.4).

- Chapter 7 reviews methods proposed in the literature for assessing and quantifying cognitive load. The methods fall in three main categories: analytical and subjective methods (section 7.1); performance-based and behavioural methods (section 7.2); and physio-psychological methods (section 7.3), including cognitive pupillometry (section 7.4). In the experimental studies we have collected and analysed subjective assessment data; we have also collected (but not analysed in the present thesis) performance-based and pupillometric data.

- Chapter 8 discusses the methodological challenges inherent to designing and executing speech elicitation experiments (section 8.1), and the potential and limitations of (re-)using spoken language corpora (section 8.2). The arguments advanced in this chapter justify our choice to collect spoken material and create new corpora of French speech produced under cognitive load.

- Chapter 9 presents a review of previous studies on the production (section 9.1) and perception (section 9.2) of speech under cognitive load. We present different methods used to induce cognitive load, to elicit speech and to interpret the findings.

- Finally, Chapter 10, synthesises the information drawn from all chapters of Part I, in order to formulate our global hypotheses on the effects of cognitive load on speech production and perception. These hypotheses are the blueprint for designing the experimental studies, as outlined in the first section of Chapter 11.

The Methodology, Tools and Baseline Studies part sets the methodological and technical foundations for the thesis, and is organised as follows:

- Chapter 11 is an introduction to the second part of the thesis. It starts by outlining the choices made to conduct experimental studies and build new corpora of French speech produced under cognitive load (section 11.1). Section 11.2 presents the aspects of spoken language that will be analysed; each of these aspects is developed in more detail in the subsequent chapters of Part II. In this work, we used several pre-existing spoken corpora, in order to develop and test new tools, and in order to obtain baseline measurements of prosodic characteristics of speech (in a wide range of communicative situations, not necessarily linked to high cognitive load). We describe all pre-existing corpora used in section 11.3 to facilitate the reader.

- Chapter 12 presents our work to improve the automatic morpho-syntactic annotation of spoken language corpora, the first step in a cascade of

automatic annotations presented in the next three chapters. We present related previous work (section 12.1), the annotation scheme we propose and apply on the data collected in the experimental studies (section 12.2), a comparison with existing annotation schemes (section 12.3), the design and architecture of the automatic tool DisMo (section 12.4) and an evaluation of its performance (section 12.5).

- Chapter 13 is closely linked to the previous chapter and outlines our work on the detailed annotation of disfluencies in spoken language corpora. We present related previous work (section 13.1), the annotation scheme we propose and apply to a part of the data collected in the experimental studies (section 13.2). We present a corpus-based study to establish a baseline for the occurrence of disfluencies in French spontaneous speech (section 13.3). We then present our efforts to automate the detection and annotation of disfluencies (section 13.4) and an evaluation of the performance of this automated tool (section 13.5).

- Chapter 14 presents work to improve the automatic detection of prosodically prominent syllables in transcribed and aligned French spoken language corpora. As explained in Chapter 4, prosodic prominence is crucial in prosodic phrasing. We present related previous work (section 14.1), the automatic tool Promise that we developed (section 14.2) and an evaluation of its performance (section 14.3).

- Chapter 15 presents work on prosodic phrasing in French. We briefly present related previous work (section 15.1), followed by a corpus study on the acoustic and syntactic correlates of prosodic boundaries based on the expert annotation of a 4,5 hour corpus (section 15.2), as well as an on-line perception experiment that validates this annotation with respect to major prosodic boundaries (section 15.3). On the basis of these findings, we propose to automate the annotation of prosodic boundaries (section 15.4).

- Chapter 16 focuses on the main measures related to the temporal organisation of speech: speech pauses (section 16.1) and the methodological challenges in describing the statistical distribution of silent pause length (section 16.2), as well as the measurement of speech rate (section 16.3). The chapter concludes with a corpus-based study to establish baseline data on the distribution of silent pause length in four different speaking styles (section 16.4); the same method is applied to analyse the data of Experimental Study 4.

- Finally, Chapter 17 presents our work to develop Praaline, a new tool for working with spoken language corpora. We cover its features in corpus management (section 17.1), annotation (section 17.2), visualisation (section 17.3), querying and concordancing (section 17.4), as well as statistics and extensibility (section 17.5).

The EXPERIMENTAL STUDIES part consists of four studies:

- Chapter 18 presents Study 1, in which we elicited speech from participants performing a series of simple tasks frequently used in cognitive psychology, namely a Stroop naming task and a Reading Span task. The experimental design was inspired from the English Cognitive Load Speech and EGG corpus (CSLE, Yap, 2012) and our study is essentially a replication for French. Along with the speech recordings, we collected electroglottographic (EGG) data and eye-tracker recordings. We analyse the EGG data to explore the effect of CL induced by simple tasks on low-level, voice quality (glottal source) features.

- Chapter 19 presents Study 2, in which we collected monologue speech produced under increasing levels of cognitive load, induced by asking subjects to perform tasks that necessitated memorisation of recently presented information. The experimental design elicited both controlled and spontaneous speech, and utterances long enough to permit prosodic analysis. We focus on the automatic analysis of basic prosodic features.

- Chapter 20 presents Study 3, on simultaneous interpreting (SI), performed in collaboration with Cédric Lenglet at the University of Mons (FTI-EII). Simultaneous interpreting is a cognitively demanding task, as the interpreter must both comprehend and produce speech, under time pressure and working memory load; this study allowed the collection of data with greater ecological validity. We compare the rendition of the same text by the same subject, a professional interpreter, who first performed the SI task under real-world conditions, and then read out the resulting rendition, after rehearsal. We focus on speech prosody, and on the prosody-syntax interface. Furthermore, a perception experiment explored how these characteristics affect the perception of fluency and interpreting accuracy, by non-expert listeners.

- Chapter 21 presents Study 4, in which we collected both monologue and dialogue speech, produced under conditions that will tax the attentional resources of the speaker. We created continuous attentional load using a dual-task paradigm: participants interacted in pairs, and one of the participants was simultaneously performing a tracking and reaction task in a driving simulator. We sought to create a realistic communicative situation that would encourage participants to produce long stretches of speech. Our analysis focuses on the temporal organisation of speech.

Finally, the Conclusion (Chapter 22) summarises the main findings of our studies on French speech produced under cognitive load, and discusses them in relation to the findings in the literature (on English speech produced under

cognitive load). It furthermore lists the main contributions of the thesis, its limitations and perspectives for further research. The materials used in the studies are reproduced in Appendix A.

The reader will notice that we have chosen to organise the material presented in the thesis so that separate chapters can be read in a relatively autonomous manner. For this reason, Part I includes only the general theoretical background that is relevant to the entire thesis. Individual chapters in Part II begin with a presentation of background information and a literature review specific to the topic of each chapter. Similarly, chapters presenting the experimental studies contain background and theoretical information specifically related to each study. For example, a discussion about different EGG measures is to be found in the beginning of Chapter 18, and a review of literature related to simultaneous interpreting can be found in the beginning of Chapter 20. Cross-references have been included throughout the thesis.

Part I

THEORETICAL BACKGROUND

2

WHAT IS COGNITIVE LOAD?

Some tasks are perceived to be more difficult and demanding than others; it is also well established that performance drops when a person is required to engage simultaneously in two or more attention-demanding tasks. In this chapter, we explore definitions of cognitive load and models of human performance. Cognitive psychology has focused on such questions since the 1950s, and current theories link the perceived difficulty and limits of performance with limits in working memory; the next chapter therefore focuses on models of working memory.

In the following, we present theoretical frameworks that have been put forward to describe and quantify cognitive load. We start with Lavie's perceptual load theory, followed by a discussion on Sweller's cognitive load theory (which is mainly focused on the design of instructional material). We then examine Rasmussen's model of human behaviour, which offers a graded view of skilled-based vs. knowledge-based actions. And finally, we present Wickens & Hollands's model of Human Information Processing Stages, which explicitly describes a role for working memory in the execution of complex tasks.

2.1 PERCEPTUAL LOAD THEORY

Several theories have been put forward in an attempt to explain the relationship between attention, task demands and the limited capacity of perceptual and cognitive resources. Load Theory (Lavie, Hirst, de Fockert, & Viding, 2004; Lavie, 1995, 2000) distinguishes between two types of demand (load): *perceptual load* and *cognitive load*. Perceptual load is higher when several task-irrelevant items (distractors) are present; cognitive load is linked to the demands of processing information (using working memory). The distinction between low and high load are relative rather than absolute. Load theory postulates that perceptual and cognitive load have opposite effects on the allocation of attentional resources. It predicts that under low load conditions, the remaining capacity will spill over to task-irrelevant stimuli, automatically and without voluntary control. In high load conditions, there is no remaining capacity and therefore, according to Load Theory, interference from distract-

ors will only occur under low-load conditions. Although this hypothesis is confirmed on an aggregate basis, a number of studies (e.g Fitousi & Wenger, 2011) indicate that, in reality, the situation is more complex: the perceptual load effect may be eliminated or reversed under certain conditions, and individual differences are significant. More importantly, it is difficult to distinguish between perceptual and cognitive load because the manipulations performed to increase perceptual load (e.g. increasing the number of available items in a search task) also affect demands on working memory and thus cognitive load.

2.2 COGNITIVE LOAD THEORY

In the field of educational psychology, John Sweller (1988) proposed Cognitive Load Theory, as a model of problem-solving behaviour, and in order to improve the design of instructional and educational material (Paas, Renkl, & Sweller, 2004). Cognitive load theory distinguishes between three types of load: intrinsic load, which is the inherent difficulty of a task; extraneous load, which is generated by the method of presentation of the task (e.g. the instructions); and germane load, which is generated when cognitive resources are devoted to the creation and automation of mental schemata. Cognitive Load Theory postulates that, while it is impossible to modify intrinsic load, reducing extraneous load will allow for more resources to be devoted to germane load, which leads to optimal learning. Research based on Cognitive Load Theory has focused on redesigning instructional material and methodology to that effect.

2.3 RASMUSSEN'S LEVELS OF BEHAVIOUR

In the field of ergonomics, Rasmussen (Rasmussen, 1980, 1987) proposed a model of human behaviour and performance, in an attempt to categorise different types of errors that arise mainly in industrial or time-critical activities. Rasmussen's model is a general framework that postulates that there are three levels of cognitive control in the performance of tasks, depending on the familiarity with the task and the environment: skill-based behaviour, rule-based behaviour and knowledge-based behaviour. The model is hierarchical, in the sense that more cognitive control is required for rule-based behaviour compared to skill-based behaviour, and for knowledge-based behaviour compared to rule-based behaviour.

According to this model, skill-based behaviour "represents sensorimotor performance during acts or activities that, after a statement of intention, take place without conscious control as smooth, automated, and highly-integrated patterns of behaviour" (Rasmussen, 1986, p. 100, cited in Reason, 1990). These activities are basically physical in nature and have been highly practiced. The second level of behaviour is rule-based behaviour, defined as a "composition

of a sequence of subroutines" which is "controlled by a stored rule or procedure". A familiar situation (e.g. a task that has been encountered many times before) triggers this stored rule or procedure. The performance of a task under rule-based behaviour is goal-oriented, but the goal is often not explicitly formulated (Rasmussen, 1987). The third level of behaviour is knowledge-based behaviour. In an unfamiliar situation, there are no stored procedures formed by previous experience, nor rules to follow. Therefore the person will explicitly formulate goals, and seek to attain them using their knowledge, reasoning and an internal representation of external reality (a mental model).

In Rasmussen's model, the greatest levels of cognitive load will be experienced while performing tasks using knowledge-based behaviour, less cognitive resources will be needed for rule-based behaviour, and skill-based behaviour will be fluid and automatic. During training in a particular task, control will move from knowledge-based behaviour to skill-based behaviour. This will allow for more cognitive resources to be freed, and potentially be allocated to other, simultaneous tasks. A typical example is driving (see Patten, 2007 for a study of cognitive workload in driving based on Rasmussen's model, as well as Wickens & Hollands' model), where novice drivers have trouble combining driving the car with other activities, such as changing a station on the radio or even talking (an observation that served as an inspiration for one of the experimental studies in the present thesis).

2.4 THE HUMAN INFORMATION PROCESSING STAGES MODEL

Wickens (1984) and Wickens, Hollands, Banbury, and Parasuraman (2012) propose a theoretical framework for analysing the different psychological processes in play when a person carries out a (potentially complex) task. The framework is called Human Information Processing Stages (HIPS) and is presented in diagram form in Figure 2.1.

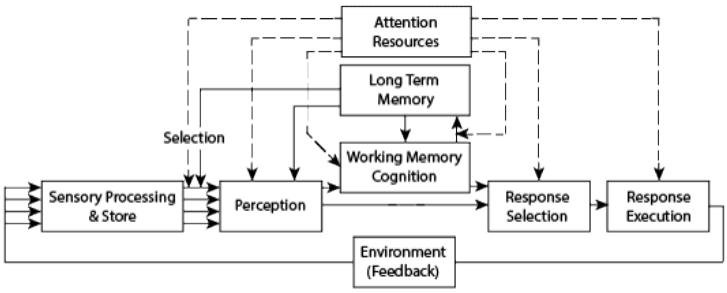


FIGURE 2.1: Human Information Processing Stages model (From Wickens, Hollands, Banbury, and Parasuraman, 2012)

In this model, information processes are shown as a series of stages, each of which build upon the input of the previous stage and transform it before proceeding to the next stage. Information is thus selectively filtered, based on the focus of attention. The process is continuous and the HIPS model also includes a feedback loop, i.e. postulates that the person's activities are constantly monitored and possibly altered based on the effects they had on the physical environment. The HIPS model comprises of the following stages: sensory processing (through short-term sensory stores), perception, working memory and cognition, response selection and execution; it also postulates that there is constant interaction between the working memory and long-term memory (cf. Cowan's model of working memory), and that the entire process is regulated by attentional resources (internal) and feedback from the environment (external). It must be noted that while HIPS may be a useful tool in modelling human performance, it may be an oversimplification of the true processes taking place in real environments: for example, the HIPS model does not explicitly attempt to explain or accommodate for simultaneous or parallel processing (cf. Patten, 2007, p. 5).

During the stage of sensory processing, the human brain receives information from the environment in the form of stimuli (auditory, visual, olfactory, tactile etc.). These stimuli may convey information, or may require a response. During the stage of sensory processing, there are large amounts of unordered information received by the sensory organs; each is equipped with a short-term sensory store (STSS) that acts as a buffer. Experimental evidence indicates that the STSS for visual stimuli retains them for approximately 0.5 sec, while the auditory STSS is much longer. Auditory stimuli may be perceived even 2-4 seconds after their completion (Wickens et al., 2012), which means that information (including speech) received through the auditory modality may be perceived even if the person was distracted during the actual transmission.

During the following perception stage, the unprocessed sensory information is filtered, interpreted and meaning is assigned to it. This processing stage generally requires little attention, and is considered as automatic. However, cognitive processing (such as reasoning, planning, analysis of speech input and integration of incoming information into a meaningful representation etc.), on the other hand requires time, attention and resources. The HIPS model postulates that such mental operations that transform or retain information make use of the working memory. Working memory is volatile, has a limited capacity and is vulnerable to interruptions or distraction of attentional resources.

The HIPS framework predicts that there must be constant interaction between the working memory and long-term memory (a long-term store of memorised information, where it is much less vulnerable to decay and attrition). The response selection and response execution stages appear as separate stages in the model to underline the fact that both of these stages require atten-

tional resources. Selecting an appropriate course of action will involve working memory, and in the case of non-automatic responses, will also involve access to the long-term memory (to retrieve rules or knowledge). According to the HIPS model, even the most automated, skill-based actions require a minimal amount of attentional resources. As a consequence, high cognitive load may disrupt skill-based, automatic actions as well.

Finally, attention is defined as the allocation of limited cognitive resources in order to focus on some aspects of the environment while ignoring others. Selective attention i.e. ability to focus on a task or an aspect of the environment is essential, and failures of selective attention occur when the limited capacity of perceptual and cognitive resources is exceeded (Shriffin & Schneider, 1977; Schneider & Shriffin, 1977).

The models and theories presented above were essentially designed to provide a conceptual framework for human performance, to explain empirical findings in the operation of complex systems by people (e.g. flying an airplane, driving a car, controlling machines in a factory, etc.), and make predictions about possible sources of error when attentional or cognitive resources are overtaxed. While they do not make explicit predictions about the use of language, they highlight the importance of working memory in explaining and quantifying cognitive load. In the following chapter, we present several models of working memory, followed by a discussion about the relationship between working memory and language. In this thesis we adopt a definition of cognitive load based on working memory. A task is placing high cognitive load demands on the performer when it significantly engages the central executive resources of their working memory.

3

MODELS OF WORKING MEMORY AND ATTENTION

The human memory system is composed of three interconnected stores: sensory memory, working memory (or short-term memory) and long-term memory. Sensory memory stores the information received by the sensory organs (sight, hearing, taste, smell, touch) just long enough to be transferred to short-term memory. Guided by attention, some information is deemed as important enough to be perceived, i.e. interpreted and transferred to working memory. Working memory stores information currently processed by the brain and it has a limited capacity. Long-term memory is a permanent store where information can remain indefinitely. The evidence to support the distinction between short-term and long-term memory comes from findings from brain-damaged patients and from differences in forgetting rate (Eysenck, 2006, pp. 143, 160).

Working memory "refers to a small amount of information held in the mind, readily accessible for a short time to help an individual comprehend language and solve problems" (N. Cowan, 2011, p. 75). It is therefore central to any cognitive activity and indeed to consciousness. Several models have been successively proposed to explain memory, attention and cognitive processing. New experimental findings have often led to adjustments or reviewing of theories and models. In this chapter, we will present the most influential models of working memory and attention. They relate to our working definition of cognitive load, as being the result of limited working memory and attentional resources.

3.1 FROM SHORT-TERM MEMORY TO WORKING MEMORY

The difference between large long-term memory in humans and a much smaller portion of memory that holds information for a brief period of time was pointed out in a seminal paper by Miller (1956); he observed that he was "being persecuted" by a number, the approximately seven (plus or minus two) items he was able to recall from a list after immediate presentation. More importantly, he described a mechanism of chunking, i.e. of grouping information into meaningful groups, or chunks (e.g. grouping letters into syllables or words, digits into numbers, using acronyms or mnemonics etc.) that allowed for better recall. Later observations would establish that the limit of

working memory is approximately 3-5 chunks (N. Cowan, 2001, cited in N. Cowan, 2011, p. 76). Miller's article sparked a long strand of research into working memory, and helped define the field of cognitive psychology.

Atkinson and Shiffrin (1968) proposed a model of human memory that was comprised of multiple stores: the sensory register through which sensory information enters the memory system; the short-term store, which can receive and hold information both from the sensory register and the long-term store; and the long-term store where information can be held indefinitely. The interactions between these three components are shown in Figure 3.1. In Atkinson & Shiffrin's model, the information that enters through sensory memory is quickly forgotten by a process of decay (iconic sensory memory decays after 0.5–1 second; auditory sensory memory decays after 1.5–5 seconds). Part of the information that enters through sensory memory is attended to and therefore transferred to short-term memory. The model also foresees a process of decay that will eliminate the information in short-term memory after 20 to 30 seconds (depending on its modality), but information can be kept longer by rehearsal. Rehearsal of information also permits its transfer to long-term memory. Information from long-term memory is retrieved into short-term memory. The capacity of short-term memory is 7 ± 2 items.

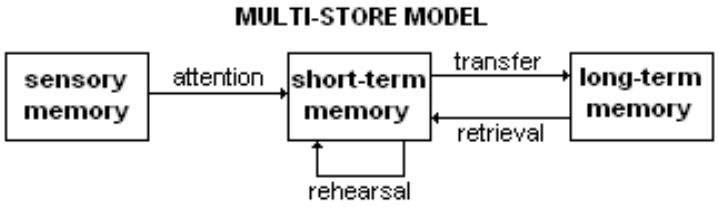


FIGURE 3.1: Atkinson & Shiffrin's (1968) multi-store memory model

Baddeley (1976) questioned the single short-term memory store in Atkinson and Shiffrin's model, based on experiments in which verbal and pictorial material was presented for immediate recall. Their results suggested that verbal material interfered with other verbal material, and that the source of interference was anchored in the sound system of language, rather than on other features such as meaning. Sound-based interference occurred even when letters were visually presented. However, there was little interference between a spatial layout and printed or spoken letters. Based on these findings, Baddeley and Hitch argued for a multi-component structure of working memory (see next section). Furthermore, Atkinson & Shiffrin's model has been criticised for suggesting that rehearsal is the only mechanism that allows transfer of information from short-term to long-term memory; and for postulating a single store for long-term memory, regardless of the nature of information (motor tasks, vocabulary, autobiographic information etc.).

3.2 BADDELEY'S MULTI-COMPONENT MODEL

In the mid-1970s, Alan Baddeley and Graham Hitch proposed model of working memory (Baddeley & Hitch, 1974) that remains influential. They argued that a large body of experimental findings is better explained by replacing the concept of a single memory store (the short-term memory as it was described until then) with a system of multiple processing and storage components. In this model, the working memory system is comprised of four components: the phonological loop, the visuospatial sketchpad, the central executive and the episodic buffer (this last component was added to the model later; Baddeley, 2000). The latest description of the model, informed by more recent findings, can be found in Baddeley, Eysenck, and Anderson (2015).

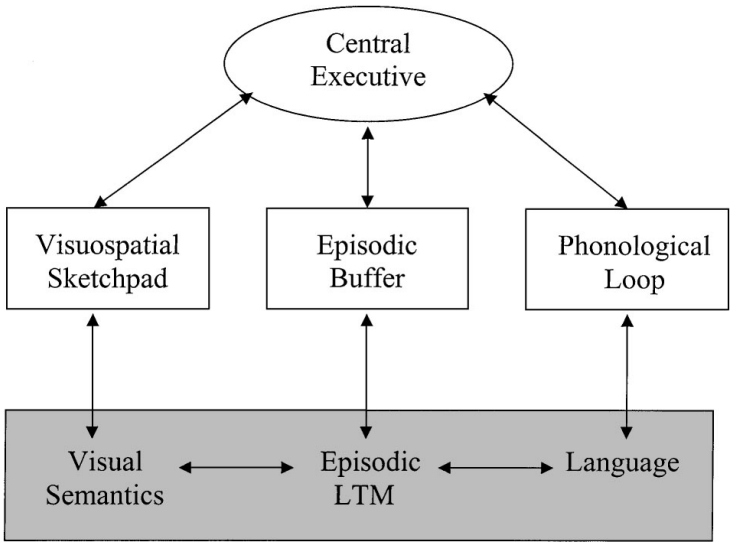


FIGURE 3.2: Baddeley's working memory model and relationship with long-term memory (LTM) (From Baddeley, Eysenck, & Anderson, 2015)

There are two domain-specific slave systems, which temporarily store information depending on its modality of presentation. The phonological loop is a temporary storage system holding verbal information in a phonological (speech-based) form. It is involved when people have to recall a set of words in the correct order immediately after hearing them, as evidenced by the phonological similarity effect (the finding that immediate recall of word lists in the correct order is impaired when the words sound similar to each other) and the word-length effect (fewer long words than short words can be recalled immediately after presentation in order). The phonological loop is further subdivided into a phonological store and an articulatory control process. Material presented verbally can be immediately stored in the phonological

store, while information presented visually (e.g. reading a text) needs to be sub-vocally articulated, through the articulatory control process, in order to be stored into the phonological store. Empirical findings indicate that the capacity of the phonological store is approximately 2 seconds (a person will be able to remember a list of as many words as they can pronounce in 2 seconds). The visuospatial sketchpad is the second domain-specific, temporary storage system, holding spatial and visual information. It is now generally accepted that it also consists of two separate components (the visual cache, which stores information about form and colour, and the inner scribe, which encodes spatial and movement information).

The two slave systems are dependent on a supervisory component of working memory, the central executive which controls and co-ordinates mental operations. While early research had focused on establishing the capacity limitations of the phonological loop and the visuospatial sketchpad, Baddeley (1996) argued that the central executive is perhaps the most important component in this working memory model, suggesting four lines of further inquiry: (1) "the capacity to coordinate performance on two separate tasks", (2) the "capacity to switch retrieval strategies as reflected in random generation"; (3) the "capacity to attend selectively to one stimulus and inhibit the disrupting effect of others"; and (4) "the capacity to hold and manipulate information in long-term memory, as reflected in measures of working memory span". A fundamental question is whether the central executive is best modelled as a unified system with multiple functions, or as the result of many interdepending and co-ordinated control processes.

More recent findings led Baddeley (2000) to add an additional slave system, called the episodic buffer. The episodic buffer is general storage component which can hold information from the phonological loop, the visuospatial sketchpad and long-term memory. This addition solves two deficiencies of the original model, namely the question of how working memory interacts with long-term memory and the empirical findings of "prose recall" (i.e. better recall of strings of text that form coherent sentences, compared to lists of individual words).

A further distinction is made in Baddeley's model between fluid systems, which are supposed to be fairly stable and unaffected by learning, and crystallised systems which are to a large extent the result of learning. The relationship between different working memory components is shown in Figure 3.2: the central executive interacts with the three slave systems (two specialised ones and the general-purpose episodic buffer); fluid systems are shown with a white background, while corresponding crystallised systems are shown in a grey background.

3.3 ERICSSON & KINTSCH LONG-TERM WORKING MEMORY MODEL

Ericsson and Kintsch (1995) focused on the relationship between working memory and skilled performance. This skilled performance includes everyday activities that have been practiced extensively, such as reading and comprehension, to specialised activities where few experts exhibit excellent performance, such as playing chess. They argued that models of working memory fail to accommodate for expert performance in the following ways: a) experts appear to greatly exceed the typical limits of working memory in skilled performance; b) if experts are interrupted by another attention-demanding task, they are able to resume their skilled activity with little disruption to their memory; c) experts are accurate in recall, even when not expecting it; d) it appears that experts are able to expand the capacity limits of short-term memory (Eysenck & Keane, 2015, p. 460).

Ericsson and Kintsch (1995) proposed the notion of long-term working memory, that postulates that experts learn how to store relevant information in long-term memory in such a way that it can be readily accessed through the retrieval cues in working memory (Eysenck & Keane, 2015, p. 460). This advantage is domain-specific and does not generalise to other activities or to working memory capacity in general. Crucially, the Ericsson and Kintsch (1995) model does not attempt to replace existing theories of working memory, but to complement them with an explanation of skilled behaviour, acquired after long and sustained practice.

3.4 COWAN'S MODEL OF ACTIVATED LONG-TERM MEMORY

Cowan's model of working memory (1988, 1999) focuses more on the relationship between attention and working memory. Cowan argues that "the differential interference results cannot be denied but could be accommodated for with the general principle that stimuli sharing features of various sorts are more likely to interfere with one another" (N. Cowan, 2011, p. 82). Instead of separate phonological and visuospatial stores, it may be the case that storage differs based on the sensory modality of the input. N. Cowan (1988, p. 171) therefore suggested that stimuli activate elements of long-term memory; working memory is the activated elements of the long-term memory (see Figure 3.3). Previous models of cognition (the information processing models) suggested that information is selectively forwarded from long-term memory to short-term memory. These models did not provide a more detailed explanation on how this mechanism works. Cowan's model stresses that not all information in working memory has the same status (information that was recently the focus of attention is more readily available) and that some stimuli activate features of (long-term) memory automatically (N. Cowan, 2011, p. 82).

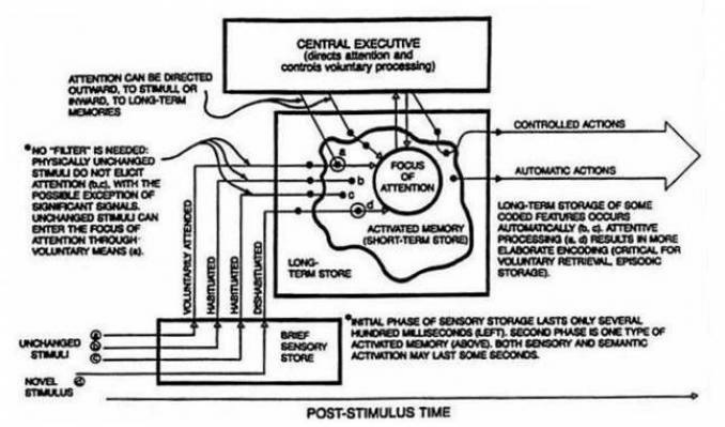


FIGURE 3.3: The Activated Long-Term Memory Model (From N. Cowan, 1988)

There are two different working memory mechanisms in Cowan's model, which lose information in different ways. The temporarily activated elements of long-term memory lose activation through decay (approximately 10–30 seconds). The subset of these activated elements which are the focus of attention resist decay and is capacity-limited: N. Cowan (2001) mentions approximately 4 chunks of information. The focus of attention is voluntarily controlled by the central executive. Working memory consists not only of a collection of activated features of long-term memory, but also includes the new links between such features.

3.5 SUMMARY

In the previous sections we described some of the most influential models of working memory. Further information on the evolution of working memory models can be found in Miyake and Shah (1999). Working memory models can be compared across several axes:

- Structural vs. functional models: Baddeley & Hitch's model suggest that there are structural differences between the components of working memory, while Cowan's and Ericsson & Kintsch define their models in terms of processes.
- Individual differences in working memory capacity are explained on the basis of the total amount of mental resources available (e.g. Just & Carpenter, 1992), or as differences in skills and knowledge encoded in long-term memory (e.g. Ericsson & Kintsch, 1995).
- Control and regulation of working memory: in Baddeley & Hitch and in Cowan's models the central executive is an essential component of the

model, and is the expression of controlled attention; Ericsson & Kintsch do not explicitly include a central executive component in their model.

Further differences include the mechanisms of representation of information in working memory, the nature of working memory limitations (limits of processing speed, decay, inhibition, interference, etc.), the relationship of working memory with long-term memory, the relationship of working memory and attention and its role in complex cognitive activities (Miyake & Shah, 1999, pp. 5-6).

In this thesis, we focus on the relationship of working memory with the production and perception of speech, and more specifically with the prosodic aspects of speech. Since cognitive load is linked with the amount of mental effort in working memory, we have selected tasks for our study that either tax the storage component of working memory, or tax the processing resources of participants. In order to formulate our hypotheses on the effects of such cognitive load on speech, we now turn our attention to models of language production and perception, the relationship of working memory with language production and perception and the role of prosody in speech production and perception: these are the subjects of the next two chapters.

4

PROSODY

Werner and Keller (1994, p. 23) define prosody as "the speech features whose domain is not a single phonetic segment, but larger units of more than one segment, possible whole sentences or even longer utterances". Prosody focuses on supra-segmental phenomena in speech, i.e. the phenomena manifesting themselves beyond the level of the individual phone. In section 4.1 we present a way to organise the study of prosody around four levels: physical, articulatory, perceptive and linguistic. In section 4.2 we list certain linguistic, para-linguistic and extra-linguistic functions of speech prosody. Key aspects of prosody that will be examined in this thesis are briefly introduced in section 4.3: prosodic prominence, stress and intonation, phrasing and boundaries (and the function of prosody in structuring discourse), as well as the temporal organisation of speech. These subjects will be developed in separate chapters, to which we provide cross-references.

4.1 LEVELS OF PROSODY

Dutoit (1997, p. 130) lists three interconnected levels of prosody: the acoustic (physical) features that are directly measurable on the speech signal; the perceptual level that refers to the way these acoustic features are perceived; and the linguistic level that refers to the functions of the perceived prosodic features. We could also add the dimension of articulation (producing speech). The study of speech prosody entails an understanding of how these four levels interact with each other: Table 4.1 shows the relationship between them.

In the following three subsections, we examine: the physical quantities that can be measured with the help of instruments and on the basis of recordings (subsection 4.1.1), the articulatory processes involved in speech production (subsection 4.1.2), and the corresponding perceptual effects and the linguistic interpretation by the listener (subsection 4.1.3).

Acoustic property	Articulation	Perception	Linguistic property
Fundamental frequency (fo)	Vocal fold vibration frequency	Pitch Pitch register Aspects of prominence	Tone and Intonation Stress, Phrasing
Signal amplitude, energy and intensity Amplitude dynamics	Egressive air-stream pressure	Loudness, Strength Aspects of prominence	Stress, Phrasing
Duration of segments	Articulation rate	Length of segments, timing Aspects of prominence	Rhythm (rate, tempo) Stress, Phrasing
Pause (in the signal)	Pause, Breathing	Pause (perceived)	Pause (interpreted)
Spectrum, Harmonics	Voicing, vocal tract configuration, vocal effort	Voice quality, timbre	Paralinguistic functions

TABLE 4.1: Levels of prosody: acoustic, articulatory, perceptual and linguistic

4.1.1 *The Acoustic Features Level*

At the level of acoustic features, information consists of objective measurements of the physical properties of sound. *Sound* is an oscillation of pressure that propagates through some medium (a mechanical wave) with a velocity that depends on features of the medium (material, temperature, pressure). A *pure tone* is a sinusoidal waveform whose mathematical expression is

$$f(t) = Asin(2\pi ft + \phi)$$

where A is its *amplitude*, f its *frequency* and ϕ its *phase*. This is a periodic function with a period of $T = 1/f$. When the period is measured in seconds, the frequency is measured in hertz (Hz). The amplitude expresses the magnitude of pressure change. The phase denotes the time difference of the sinusoidal wave with respect to a reference point and is expressed in angle units (e.g. radians) or equivalent time units.

The *amplitude* is proportional to sound pressure (measured in pascal, Pa) which is the local pressure deviation from the ambient atmospheric pressure. The pressure variation created by sound waves varies from 20 μPa to 20 Pa. The root-mean-square pressure (average over one complete cycle) is directly related to the energy carried by the sound wave. Sound intensity is defined

as the sound power (energy per time unit) per unit area. Therefore, the intensity is proportional to the square of the amplitude, and ranges from 10–12 W/m^2 to 1 W/m^2. Since the human ear is sensitive to such a wide range of sound pressures (amplitudes) and intensities, we use a logarithmic scale in our measurements. The Sound Pressure Level (SPL) is defined as

$$L = 20log_{10}\frac{p}{p_{ref}}$$

where p is the root-mean-square sound pressure and p_{ref} is a reference sound pressure, set to 20 μPa at 1 kHz (in air). The SPL is expressed in decibels. The auditory threshold (the faintest perceivable sound) is 0 dB at 1 kHz and the upper limit before hearing damage occurs is approximately 120 dB.

The principle of Fourier analysis is that any periodical signal can be expressed as the sum of sinusoids each scaled and shifted by appropriate time constants. The only sinusoids required are those whose frequency is an integer multiple of the fundamental frequency of the periodic sequence, i.e. its harmonics. Speech is a non-periodic signal, but by taking a small window of analysis and assuming the excerpt is repeated ad infinitum, we obtain the Fourier analysis of this excerpt and arrive at the speech signal's spectrum (energy or power per frequency band).

The peaks of the speech signal's energy spectrum correspond to the resonators of the vocal tract. These peaks are called *formants* and are used to distinguish individual phoneme realisations (phones) on the spectrogram. The *fundamental frequency* (f_0) given by the Fourier analysis of the speech signal correlates with the vibration frequency of the vocal folds, and with the perceived pitch of the human voice.

4.1.2 *The Articulation Level*

Three systems need to be co-ordinated for speech production: the respiratory, phonatory and articulatory system. Most of speech sounds are produced by an egressive pulmonic airstream, i.e. an outgoing stream of air passing through the larynx and along the vocal tract, a tube of complex shape formed by the mouth and nose; the vocal tract changes shape in order to articulate. In normal breathing, the inspiration and expiration phases have roughly the same duration. When speaking the inspiration phase gets considerably shorter (producing audible breath pauses) and speech is produced during the longer expiration phase (P. Martin, 2008, p. 54).

The vocal folds vibrate rapidly when the airstream is allowed to pass between them, in which case the sound produced is voiced. The frequency of the vibration of the vocal folds defines the perceived pitch of the speaker (see below). In voiceless sounds, the vocal folds are held apart and the airstream passes freely. A glottal stop is produced when the airstream is momentarily

blocked; a creak is a succession of glottal stops. Creaky voice is the combination of creak with voicing. In whisper the vocal folds are brought together but without vibrating; breathy voice is the combination of whisper with voicing. The airstream is modulated passing through the pharyngeal, oral and nasal cavity. These cavities act as resonators, amplifying some of the harmonics of the vibration produced by the vocal cords and dampening others. Their configuration is modified by the articulators: lips, teeth, alveolar ridge, hard and soft palate, uvula and tongue. *Voice quality* is defined by voicing and the articulation configuration. Speech sounds are organised in a way specific to each language: a distinction of place and manner of articulation can be pertinent (as demonstrated by the existence of a minimal pair) in which case the two sounds are considered different phonemes. Phonemes are classified into vowels and consonants.

4.1.3 *The Perceptual Level*

At the perceptual level of prosody, "one tries to establish to what extent acoustic prosodic events are perceived by the average listener and in what way" (Mertens, 2014, p. 18). With respect to the main measures we have seen so far:

- *Pitch* is the perceptual correlate of f_0. Vocal fold vibration frequency, fundamental frequency f_0 and pitch are correlated but not identical. Accurate measurements of vocal fold vibration patterns can be obtained by using the technique of electroglottography (used in Study 1, presented in Chapter 18). Pitch tracking algorithms estimate the f_0 by performing a Fourier analysis of the recorded signal. A perceptual measure of pitch (e.g. the semitone scale used in this thesis) takes into account that human perception of pitch is logarithmic and not linear.

- *Loudness* is the perceptual correlate of intensity. It is a subjective measure and is measured in phones, as defined by the equal-loudness contours (reflecting the fact that the human ear is more sensitive to certain frequency bands and less sensitive to others).

- *Timing* is the perceptual correlate of duration. Several properties of the signal interact to affect the perception of duration (the actual duration, the position of a segment relative to other segments, f_0, etc.).

- *Voice quality* is the perceptual correlate of the spectral characteristics of the speech signal.

Note, however, that there is no one-to-one relationship between objectively-measured physical properties and their perceptual correlates. It is often the case that the perception of one prosodic feature interacts with several acoustic measures.

4.2 FUNCTIONS OF PROSODY

A phonological description of prosody "posits a minimal set of distinctive forms, which are assumed to have a function in speech communication. The representations for these levels of observation differ, while being related: an abstract phonological form may correspond to several forms of the perceptual level, the shapes of which differ to some extent, while being related" (Mertens, 2014, p. 18).

Prosody fulfils several functions at the linguistic, paralinguistic and extralinguistic level:

- At the linguistic level, prosodic features facilitate lexical access by aiding listeners perceive word boundaries (Cutler, 1997) and indicate prosodic phrasing (cf. Frazier, Carlson, and Clifton, 2006; Delais-Roussarie, 2000). Prosodic features function as the "punctuation marks" of speech: prosodic boundaries segment utterances, facilitate turn-taking and signal reported speech segments. Prosody denotes information structure, such as focus and saliency, topic, parentheses etc. (for a review, see Cole, 2014). Prosodic features also denote modality, i.e. whether the utterance is a statement, a question, an imperative etc. (e.g. Delattre, 1966; Delais-Roussarie et al., 2015). Prosodic cues can be used to manage interaction (e.g. dialogue turns Cutler & Pearson, 1985).
- At the para-linguistic level, prosody fulfils functions that are of interest to pragmatics (meaning in context). It is used to express the attitude of the speaker towards the content (e.g. irony, Kade, 1963, p. 19), or to demonstrate the emotional state of the speaker (e.g. Ohala, 1996).
- At the extra-linguistic level, prosodic features indicate sociolinguistic variation. Prosodic features may indicate geographical variation (e.g. Simon, 2012; Avanzi, 2014) and social variation (e.g. Weinreich, Labov, & Herzog, 1968; Labov, 1972). Prosodic features are related to a speaker's gender and age and are an expression of a speaker's idiolect, i.e. the linguistic and stylistic habits that differentiate an individual from others. Prosody is also an indication of and influenced by situational factors and constraints (Koch & Oesterreicher, 2001; Simon, Auchlin, Avanzi, & Goldman, 2010; Goldman, Prsir, Christodoulides, & Auchlin, 2014).

4.3 ASPECTS OF PROSODY EXAMINED IN THIS THESIS

Prosodic prominence is defined by Terken (1991) as follows: "we say that a linguistic entity is prosodically prominent when it stands out from its environment by virtue of it prosodic characteristics". Prosodic prominence is related to multiple functionally different phenomena, including phonological lexical stress, accentuation and prosodic phrasing. Its perceptual correlates include several acoustic features (f_0 movement, duration, intensity etc.).

In variable-stress (lexical-stress) languages, such as English or Dutch, the term "stress" is a lexical property of a specific syllable in a word; in these languages syllabic prosodic prominence is a distinguishing property of words in the lexicon, while "pitch accents" (correlating with f_0 movements) are used to signal the information status of a linguistic unit. In French, the prosodic prominence of a syllable mainly contributes in the demarcation of (minor) prosodic boundaries. Most models of French prosody admit at least three degrees of prosodic boundaries and a hierarchy of three levels of units (Mertens, 1993; Rossi, 1999; Di Cristo, 1999).

A detailed description of the following aspects of prosody will be presented in the following chapters of the thesis:

- The role of prosody in discourse structuring, from a production and a perception point of view (sections 5.4 and 6.4 respectively).
- The issue of prosodic prominence, and more specifically syllabic prosodic prominence in French (Chapter 14).
- The demarcation of prosodic boundaries in French, their perceptual correlates and their relationship with syntactic structure (Chapter 15)
- The temporal organisation of speech, including speech pauses (section 16.1), speech rate (section 16.3) and disfluencies (Chapter 13).

5

LANGUAGE PRODUCTION

Language production is divided in three major stages: conceptualisation, i.e. deciding on the intended message, formulation, i.e. selecting a way to express this message and articulation, the actual production of speech using the vocal tract and the articulators. Despite a large body of research, there is yet not a single, unifying framework to describe speech production, taking into account all factors known to influence it.

Language production processes fundamentally differ from language comprehension processes in many respects. Speaking starts with the intention to produce language. Producing a word may take up to five times longer than recognising a word. Listeners direct their gaze towards the referent of a noun before the speaker finished articulating the noun, while speakers typically take 900 ms to start articulating in a picture naming task (all examples from Griffin and Ferreira, 2006, p. 21). While the relationship between working memory and comprehension has been studied thoroughly, less literature is available on the relationship between working memory and production.

In the following, we review the evolution of models of language production (section 5.1). Many of the improvements in models of language production have resulted from the systematic study of speech errors, i.e. situation in which there is a mismatch between the intended message and the utterance actually spoken. We note that, while it is known that many factors such as emotion, situational characteristics, individual characteristics of the speaker etc. affect spoken language production, mainstream psycholinguistic models do not integrate such factors in their description of mental processes. However, researchers in the field are gradually recognising that it would be beneficial to do so (e.g. O'Connell & Kowal, 2008).

In order to link the models of language production discussed in this chapter with the models of cognitive load and working memory described in Chapters 2 and 3, we review the literature on the relationship between working memory and the production of language (section 5.2). In section 5.3 we focus on the self-monitoring and repair, a feedback mechanism which is constantly operating during speaking and may lead to the production of disfluencies. These phenomena can be informative of the underlying processes of production or the cognitive state of the speaker, and are of particular interest in the study of

speech produced under high levels of cognitive load. Finally, we review the literature on the role and function of prosody in spoken language production (section 5.4), and especially with respect to the prosodic marking of discourse segmentation.

5.1 MODELS OF LANGUAGE PRODUCTION

Language production models attempt to describe the process under which linguistic units are retrieved from long-term memory and assembled to form utterances. There is consensus on two main aspects: that language is represented in a hierarchy of levels (articulation features, phonemes, syllables, morphemes, words and multi-word expressions, syntax and semantics), and that language production is sequential and incremental. Linguistic units are retrieved and assembled in stages, with each stage building upon the results of the previous stage. All language production models postulate that in the process of encoding a message, semantic information precedes syntax and grammatical encoding, and is followed by phonological planning and articulation. Language production models differ in the number of stages and the interactions between them, as well as in the strictly sequential or parallel nature of the process. The study of speech errors has helped refine language production models (Bock, 1991); we examine this subject in further detail in section 5.3.

In serial processing models, earlier stages define larger linguistic units (e.g. propositions) and later stages specify their smaller constituent parts (e.g. morphemes). The stages are strictly sequential and information flows only in one direction, from the early stages to the late stages. These were the earliest models to be proposed in the literature.

Fromkin (1971) proposed a Five Stages Model of language production, with the following steps / modules: (1) generation of the intended meaning (conceptualisation); (2) formulation of a syntactic structure; (3) selection of an intonation contour (for the entire utterance) and placement of primary stress; (4) word selection, that comprises of two sub-stages: first, content words are inserted into the syntactic frame, and second, function words and affixes are added; (5) access to phonemic representations and application of phonological rules, leading to articulation. This model's main features is the strict serial order of processing, the fact that a syntactic frame, followed by a prosodic frame (according to Fromkin, syntax drives prosody), are available early on and before the selection of content words. The model successfully explains many speech errors (e.g. that function words accommodate for word exchange) and two main observations: that word exchanges occur primarily between words of the same grammatical category; and that, when such substitutions occur, the intonation contour of the utterance remains unchanged.

Garrett (1975, 1980) also proposed a model of serial processing in speech production (again, the sequence of stages runs form semantic information

to syntactic structure to phonological rules). In Garrett's model, there are three levels of representation: the *message level* (conceptualisation, or generation of the intended message); the *sentence level* (formulation of sentence structure); and the *articulatory level* (the execution of motor plans by the articulators). The sentence level is subdivided into two separate stages: the functional level, which is comprised of lexical selection or lexicalisation, where the speaker selects the appropriate content words for the message, and the functional assignment, where syntactic functions are assigned to content words and grammatical rules are applied; and the positional level, where the order of words defines their output sound (co-articulation). Garrett justified the separate stages in his model based on speech error. He argued that the lexicalisation and the functional assignment stages were separate in order to accommodate for the fact that substitutions may be meaning-related (substitution of a word with another one having the same grammatical function) or function-related (substitution of a morpheme with another one, thereby changing the grammatical function of the word articulated). He also argued that the positional level is separate, based on phonological accommodation in speech errors, i.e. the observation that phonological rules are correctly applied in cases of word substitutions (e.g. choosing between the article "an" or "a" in English, or performing a liaison in French). Garrett's two stage model is represented in graphical form in Figure 5.1.

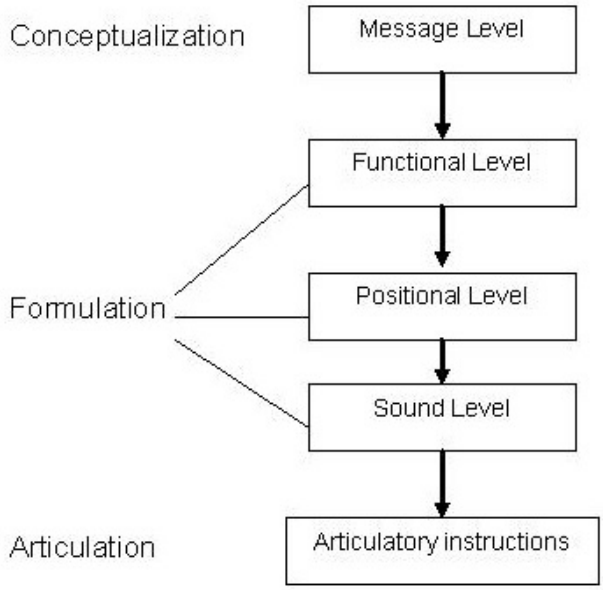


FIGURE 5.1: Garrett's two-stage model of language production

While Fromkin's and Garrett's models explain several speech errors, they do not account for phrase blends (mixing of two semantically related phrases), word substitutions between two phonologically similar words, and Freudian slips (which indicate competition at the message encoding level).

A more recent and widely-used serial model of speech production is the one proposed by Bock and Levelt (Bock, 1991, 1982; Bock & Levelt, 1994; Levelt, 1989). It consists of four levels of processing: the *message* level, the *functional processing* level, the *positional processing* level, and the *phonological encoding* level (see Figure 5.2).

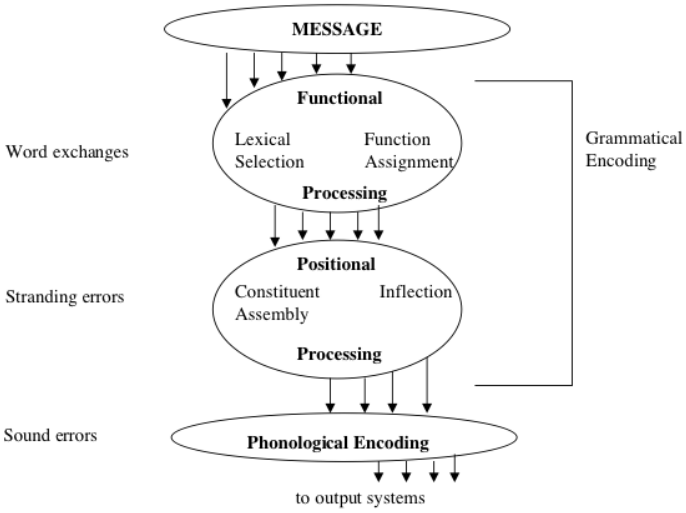


FIGURE 5.2: Bock & Levelt model of language production (From Levelt, 1989)

On the message level, the speaker generates an idea or message. The second level is the functional level, which is composed of two stages: lexical selection and functional assignment. During lexical selection, the conceptual representation of the message is converted into a choice of words; these are lemmas, i.e. lexical categories bearing possible syntactical roles. Concrete syntactical roles are assigned to lemmas during the functional assignment sub-stage. At the positional level, the speaker creates an ordered set of morphological slots and proceeds with grammatical encoding (inflection, agreement etc.). At the fourth level, the phonological encoding, these slots are encoded into lexemes (combining morphological and phonological information). In Bock & Levelt's model, the prosodic properties of the utterance (e.g. its intonation) are selected at the fourth level of production. The output of this process drives the articulation motor system.

Bock & Levelt's model can explain most speech errors observed in corpora. Levelt (1989) completed the model with a feedback (monitoring and repair) component, which we examine in section 5.3. While Bock & Levelt's model al-

lows for bidirectional flow of information between the processing stages, it is still a serial processing model. Parallel processing models, on the other hand, are based on the premise that parallel paths are simultaneously activated during production, and activation can spread in any direction. More specifically, the conceptualisation level may receive feedback from the formulation level and the articulation level and vice versa. Choices made in the beginning of an utterance (e.g. lexical selection) constrain subsequent choices.

The most notable parallel-processing model of production is Dell's model (Dell, Chang, & Griffin, 1999). In this model, speech is produced by the interaction of a number of connected nodes. These nodes represent units of language at any level: concepts for semantic features, words including their morphological and syntactical properties and phonemes. A node may be connected in multiple directions: for example, the representation of a /d/ phoneme would be connected with all morphemes containing this sound (e.g. dog), as well as with the articulatory features of the sound; similarly the word "dog" would be semantically connected with the word "cat" (Figure 5.3, right). Because of this fundamental property, Dell's model is also referred to as a *connectionist* model of speech production.

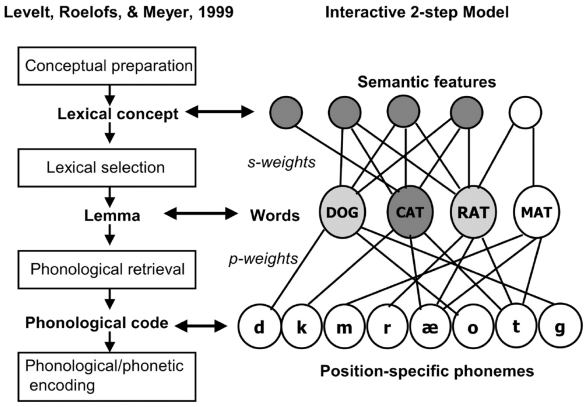


FIGURE 5.3: Relationship between Levelt's model (left) and Dell's model (right)

During lexical selection, all nodes related to a specific word are activated and, through competition, the most highly activated nodes define the actual speech output produced. Dell's model, therefore, explains speech errors as resulting from the activation of the wrong node, at any level. It can thus explain a variety of speech errors, including word blends, phrase blends, phonemic slips and cognitive intrusions. The relationship between Levelt's model and Dell's model is presented in Figure 5.3.

Finally, one of the most recent computational models of speech production is WEAVER++ (Word Encoding by Activation and VERification). It seeks to explain the role of attention in the process, and stresses that production is in-

cremental. According to this model, planning is the process of incrementally extending verbal goals: lemmas as selected form concepts, morphemes are selected for lemmas, segments for morphemes, and syllable motor programmes for syllabified segments (the syllabification of segments is also incremental). WEAVER++ has been successful in explaining and predicting the results of controlled naming experiments.

Having reviewed the main models of language production, we now turn our attention to the potential role of working memory in each of the stages of production (in the organisation of the following sections, we essentially adopt Levelt's division of stages).

5.2 WORKING MEMORY AND AUTOMATICITY IN LANGUAGE PRODUCTION

In this section we consider the potential role of working memory in different stages of speech production. Reviewing studies on this subject contributed to formulating our hypotheses about the expected effects of cognitive load on language production, i.e. what we would expect to happen if a speaker is engaging in activities that place significant demands on his working memory (this load may be caused by linguistic processing itself, or from engaging in simultaneous non-linguistic tasks).

According to Meyer, Wheeldon, and Krott (2007, p. xii), "producing and understanding utterances necessarily involves holding various types of information in working memory. For instance, speakers must remember whom they are talking to; keep track of what they have said already; and, according to many theories, temporarily buffer utterance fragments before producing them. Similarly, listeners must remember what they have been told already, and keep utterance fragments in a buffer, minimally until semantic and syntactic processing is possible. Therefore, an important issue is how the role of working memory should be conceptualised in models of speech processing".

These questions are closely linked with the degree of automaticity in language production: to what extent are cognitive and attentional resources needed for each stage of production? Following Garrod and Pickering (2007) we argue for a graded definition of automaticity (see below). Examining the different stages of speech production outlined in section 5.1 is not sufficient: it is also necessary to take into account the role of the self-monitoring and repair mechanisms, and to what extent these mechanisms impose demands on cognitive and attentional resources (section 5.2), as well as the role of prosody in speech planning and production (section 5.4).

The studies presented here can be grouped into two main categories. On the one hand, correlational studies that seek a relationship between scores obtained in a working memory test and performance in language production. These studies have used many different types of working memory tests, and different definitions of "performance" or "quality" of linguistic production.

On the other hand, studies have focussed on the effects of concurrent working memory load on language production; the latter are more relevant to the present thesis.

5.2.1 *Correlation of Working Memory Capacity with Different Measures of Fluency*

Early studies on the role of working memory in language production tried to correlate working memory capacity, as measured by one of the several available tests, with different measures of fluency (R. C. Martin & Slevc, 2014, p. 439). For example Daneman (1991) measured working memory capacity using the speaking span test (Daneman & Green, 1986) and found a positive correlation with the number of words produced in a picture description task (calling this measure "discourse fluency"), and a negative correlation with reading speed (speech rate) and the number of errors in reading a passage.

Fortkamp (2003) investigated whether there is a relationship between working memory capacity and L2 (English) speech production, using the speaking span test to measure WM capacity and speech elicited through a picture description task. The results show a positive correlation between WM capacity on the one hand, and fluency (measured on the basis of speech rate), accuracy and complexity on the other hand; a negative correlation of WM capacity with weighted lexical density; and interactions between the four dependent measures. They explain the results as showing that "grammatical encoding is a complex subtask of L2 speech production that requires the control and regulation of attention".

Fehringer and Fry (2007) studied "the use of hesitation phenomena, specifically filled pauses (ums and ers), automatisms (sort of, at the end of the day), repetitions and reformulations, in both the mother tongue (L1) and second language (L2) of highly proficient adult bilingual speakers (English and German)", with a purpose "to ascertain whether speakers who are highly proficient in L2 produce an approximately similar amount of hesitation phenomena in both languages; and whether the production of such elements (in both languages) is linked to working memory capacity". They found that "despite high proficiency, speakers produced a higher overall rate of hesitation phenomena in their L2, indicating that there was an additional cognitive load imposed by working in L2; and in each language there was an underlying negative relationship between memory capacity and the production of hesitation phenomena, implying that speakers with lower memory ability rely more heavily on such time-buying devices. Furthermore, it was shown that the individual types of hesitation phenomena produced by speakers in their L1 were carried over into their L2, which suggests that a speaker's planning behaviour is mirrored in both languages".

Note that these studies correlate a global measure of working memory capacity with measures that conflate all stages of speech production and levels

of linguistic representation. There are many possible definitions of the notion of fluency; in any case, fluency measures "depend on success at all levels of production, from message generation to articulation" (R. C. Martin & Slevc, 2014, p. 440).

5.2.2 *Working Memory in Message Encoding*

A speaker needs to maintain the central theme of a monologue or dialogue, and to keep track of preceding discourse, such as changes in topic and dialogue acts (e.g. a question requiring an answer). Successful communication requires organisation of discourse as summarised by Grice's maxims: the maxim of quantity, whereby the speaker tries to be as informative as possible, and no more; the maxim of quality, where one tries to be truthful; the maxim of relation, according to which the speaker will strive to make contributions relevant to the topic discussed; and the maxim of manner, stating that the speaker will avoid ambiguous and obscure propositions. These maxims can be deliberately violated, e.g. when the speaker wants to make an implicit statement. Obviously Grice's maxims describe a fully collaborative dialogue between two ideal interlocutors, but they indicate that for communication to be successful, speakers need to have a mental representation of previously-held discourse.

R. C. Martin and Slevc (2014, p. 440) suggest that "a more appropriate measure of fluency at the message level is probably discourse coherence, or how well utterances are linked with what has come before". Kemper, Herman, and Lian (2003) studied young and older adults, who answered questions while walking, finger tapping, and ignoring speech or noise. They scored the samples in three dimensions: fluency, complexity and content. They found that "older adults' speech was less fluent and less complex than young adults' speech. Young adults adopted a different strategy in response to the dual-task demands than older adults: they reduced sentence length and grammatical complexity. In contrast, older adults shifted to a reduced speech rate in the dual-task conditions". R. C. Martin and Slevc (2014, p. 441) summarise the role of working memory in discourse cohesion as follows: "[there is] evidence for a role for working memory in discourse fluency, and for a role of simple short-term memory span in discourse coherence, but the relative contributions of working memory and short-term memory have not been directly assessed". They also note that "these findings might be well-explained by a retrieval model of working memory" (e.g. Cowan's model of working memory as the activated part of long-term memory, see section 3.4).

Another aspect of message encoding potentially related to working memory constraints is that of audience design, i.e. the speaker's ability to adapt their utterances to their listener's expectations. The success of this effort depends on the speaker's understanding on the common ground they share with their listeners. Keeping track of common ground is a cognitively demanding task

and apparently is part of the monitoring function of production. William S. Horton and Keysar (1996) tested whether common ground is involved in initial utterance planning, or whether it only plays a role in monitoring, by adding time pressure to a referential communication task (a task where it is necessary to have some kind of information is exchanged between two speakers, e.g. a description of an object by one participant so that the other can identify it among a number of similar ones). They found that common ground was used under no time constraints and was not used under time pressure, concluding that speakers "do not engage in audience design in the initial planning of utterances; instead, they monitor those plans for violations of common ground". This view is also supported by Clark and Krych (2004) who conducted an experiment establishing that speakers not only monitor their own speech for errors, but also "monitor addressees for understanding and, when necessary, alter their utterances in progress". However, an alternative view is that audience design is not the result of special, cognitively demanding monitoring mechanisms, but "can be understood as emergent features of ordinary memory processes" (W. S. Horton & Gerrig, 2005). C. Rossnagel (2000), C. S. Rossnagel (2004) found that speakers adapt less to listeners' expectations under greater levels of memory load, arguing in favour of a model of speech production where "the controlled processes of monitoring and adjustment operate on the output of a predominantly automatic stage of planning. Cognitive load impairs monitoring and adjustment, and leads to 'standard' utterances that are not adapted to the addressee's perspective".

Regarding the use of reference (anaphors), J. E. Arnold (2010) reviews a series of studies showing that accessibility is not only a function of discourse status but is also influenced by pressures on the working memory (R. C. Martin & Slevc, 2014, p. 441), and finds that "non-linguistic processing constraints increase the use of explicit forms".

5.2.3 *Working Memory in Grammatical Encoding*

The role of working memory in sentence production is less controversial, compared to the effects at the message encoding level, and it has been shown that speakers produce less syntactically complex speech under cognitive load (e.g. Kemper, Herman, & Lian, 2003). Kemper, Schmalzried, Herman, Leedahl, and Mohankumar (2009) used a dual-task paradigm (a concurrent digital pursuit rotor task) in young and older adults. After training on the pursuit rotor, participants were asked to track the moving target while speaking. They found that "young adults experienced greater dual task costs to tracking, fluency, and grammatical complexity than older adults. Older adults were able to preserve their tracking performance by speaking more slowly. Individual differences in working memory, processing speed, and Stroop interference affected vulnerability to dual task costs". They conclude that the use of such a

tracking task is a valid method for inducing dual-task demands to speakers and that their study "confirms prior findings that young and older adults use different strategies to accommodate to dual task demands".

Producing syntactically complex sentences may be cognitively demanding in itself, or may reflect the cognitive demands of accessing rarely produced syntactical structures. The first view is upheld for example by Scontras, Badecker, Shank, Lim, and Fedorenko (2015) who, in a speech elicitation experiment, found that "there is a cost associated with planning and uttering the more syntactically complex, object-extracted structures, and that this cost manifests in the form of longer durations and disfluencies". Conversely, Gennari and Macdonald (2009) used a combination of corpus methods (frequency count of specific structures) and comprehension studies to argue that "that the way in which the verb roles are typically mapped onto syntactic arguments in production plays a role in comprehension" and that less frequently occurring structures were more difficult to comprehend. While this seems like a chicken-and-egg problem, we may conclude that under higher levels of cognitive load, speakers will have the tendency to produce less syntactically complex utterances.

Furthermore, speakers tend to select structures that allow for later production of information that is relatively complex, and allow for earlier production of information that is more accessible (J. E. Arnold, Losongco, Wasow, & Ginstrom, 2000). Slevc (2011) conducted three experiments to study these effects. In the first one, it was shown that "speakers produced accessible information early less often when under a verbal working memory load than when under no load". The second experiment found the same pattern "when accessibility was manipulated by making information given". And the third one found that the "speakers' tendency to produce sentences respecting given-new ordering was reduced more by a verbal than by a spatial working memory load". They conclude that "accessibility effects do in fact reflect accessibility in verbal working memory, and that representations in sentence production are vulnerable to interference from other information in memory".

Finally, with respect to the scope of planning (i.e. how far ahead speakers plan and choose lexical representations), R. C. Martin and Slevc (2014, p. 445) indicate that there is "evidence that the scope of planning during grammatical encoding is at the phrasal level, and thus several words may need to be maintained at the functional assembly stage during the planning of a phrase". This is compatible with the prediction that speakers under working memory load will select less complex structures, and therefore a syntactic frame necessitating a smaller number of lexemes to be activated.

5.2.4 *Working Memory in Phonological Encoding*

At the phonological level, one could expect that the phonological loop hypothesised in Baddeley's model of working memory would be crucially in-

volved in the ordered production of phonological forms. While there is scarce research attempting to isolate this stage of production in non-pathological speech, R. C. Martin and Slevc (2014) conclude that "there is little evidence on this proposal and it may be that the phonological capacity required for the output is quite small, such that individual differences in this capacity would be minimal". It appears that a very high level of memory load would be needed (or indeed a pathological condition) to induce errors in articulatory motor programmes.

However, we note that studies focusing on load effects at the message planning and the grammatical encoding stages of production do not sufficiently address the question of the prosody-discourse and the prosody-syntax interface. Will speakers have the tendency to select the simplest available prosodic structure corresponding to the intended message and a given syntactic structure? If that is the case, the implication would be that prosodic phrasing is performed during the very last stages of utterance production, following syntactic planning. Furthermore, will the effects of cognitive load be limited to hesitation phenomena, such as the insertion of silent pauses and filled pauses, or should we also expect effects at the segmental duration level? We revisit these questions in section 5.4.

5.2.5 *Automaticity in Language Production*

In cognitive psychology, automatic processes are those that are involuntary, do not draw on general cognitive resources such as attention, and resist interference from other activities (both attention-demanding and automatic). Conversely, controlled processes are voluntary and attention-demanding, interfere with other attention-demanding activities and are subject to interference. The view that processes either strictly controlled or strictly automatic has been recently challenged (Garrod & Pickering, 2007, p. 3). Bargh (1994) argues that complex processes are made up of both automatic and controlled components and identified the following four criteria (he calls the "four horsemen of automaticity"):

- Awareness: automatic processes are those of which the subject is not aware. (E.g. subliminal priming)
- Intentionality: does the subject need to voluntarily instigate the process? (E.g. Stroop effects are automatic because they occur independently of the subject's will).
- Efficiency: automatic processes are faster and more efficient than controlled processes, and require less or no focal attention.
- Interruptibility: automatic processes are those that the subject cannot, or finds it very difficult, to stop or modify once they have started.

These criteria combine to create the notion of graded automaticity. Garrod and Pickering (2007) propose that language production exhibits graded automaticity; that it is a complex activity consisting of both automatic and controlled processes. They point out that "most aspects of language production involve some degree of choice between alternatives. It may be that the degree of automaticity is related to the extent to which the speaker has to make such choices because choice relates to intentionality and strength of processing". They then proceed to examine the degree of automaticity of the stages of production in Levelt's (1989) model.

The message conceptualisation stage is controlled with respect to all four criteria. The speaker is aware of his intention to convey a message and the process is intentional. It is not efficient "in the sense that people can put considerable effort into deciding on what to talk about next" (p. 4) and it is interruptible. There is however a small automatic component, in the sense that unintended associations between ideas occur. Furthermore, "there is indirect evidence from models of Stroop interference results that establishing the linguistic concept is a controlled process" (Roelofs, 2003, cited in Garrod and Pickering, 2007, p. 5). In the Stroop test, reading words is automatic and not subject to interference, while naming colours (establishing the correct linguistic concept) is a controlled process and thus subject to interference.

Lexical selection is the most controlled: "to the extent that speakers are always presented with the problem of choosing the appropriate level of lexical specification, the process cannot be completely automatic. In terms of the four horsemen, speakers can become aware of lexical choice, before, during and after uttering a word". This observation applies to content words; Garrod and Pickering (2007) argue that speakers are not explicitly aware of the choice of function words, and that "such words are selected on the basis of the compatibility with other words, and do not require prior activation of a concept". Lexical access is not efficient, in the sense that cognitive load affects it; and it is normally interruptible, as evidenced by false starts and immediate lexical substitutions.

Grammatical encoding is both automatic and controlled. Speakers need not be aware of the grammatical form they have chosen, and they are not aware of the factors that led them to choose one construction over another one. "This seems to point to an awareness of the output of the production process, but, crucially, not to awareness about the process itself" (p. 7). Sometimes a construction may be chosen intentionally (e.g. a passive construction) but often this is not the case. However grammatical encoding seems to draw upon cognitive resources (see discussion above) and is normally interruptible. Grammatical encoding "shows some features that point to its being a controlled process, such as being partly open to awareness and competing for central resources. But it also seems more automatic than some earlier processes such as identifying the concept and the lemma" (pp. 7-8). We note here that some of the differential effects observed in L2 production compared to

L1 production (e.g. Fehringer & Fry, 2007) may be due to the fact that grammatical encoding in an acquired language is less automatic than in the L1.

Garrod and Pickering (2007) point out that there are aspects of phonological encoding that must be controlled, such as the use of contrastive stress (implementing prosodic prominence at the lexical level to show information status) or the choice of intonation (e.g. to indicate a question). Levelt (1989) proposes a "prosody generation module" under executive control, responsible for the discourse-prosody interface. On the other hand, the process of selecting phonemes and syllabification is highly automatic. Even at the level of articulation, though, comparative studies on the realisation of shandi phenomena by L1 and L2 speakers (e.g. Barreca, 2015, studying liaison in French) show that there must be a level of control by the speaker. The speaker can also choose to modulate his articulation rate, in order to convey a specific message.

The monitoring mechanism "may be automatic, but there may be a controlled process of interrupting production, to allow an aspect of speech to be corrected or abandoned" (Garrod & Pickering, 2007, p. 8). We will examine the extent to which self-monitoring and repair is automatic in the next section.

In summary, language production involves both automatic and controlled processes. Higher-level stages are more controlled, whereas lower-level stages are more automatic; however there is a degree of automaticity even in the higher-level processes, and a degree of control even in the lower-level processes. Under increasing levels of cognitive load, we expect that controlled processes will be affected. This hypothesis, along with the discussion on the role of working memory in each stage of language production, drives our global hypotheses on the effects of cognitive load on speech production.

5.3 SELF-MONITORING AND REPAIR

Speakers often interrupt themselves and correct something they have just said. A self-repair is defined as a "correction of errors without external prompting, frequently within a short span of time from the moment of error occurrence" (Postma, 2000, p. 98). This means that there must be a cognitive mechanism that monitors one's own language production processes, and intervenes when necessary (Hartsuiker, 2014, p. 417). We can break down the structure of a self-correction utterance into three contiguous regions, following Shriberg's (1994) annotation scheme for structured disfluencies:

```
(reparandum) * interruption point (interregnum and optional editing
terms) (repair)
```

The reparandum is the part of the utterance that is repeated or that will be corrected, edited, or deleted. The interruption point is the point between the

reparandum and the interregnum; this instance in time does not necessarily coincide with the moment the speaker detected the trouble or his intention to alter the utterance. The interregnum is the part between the reparandum and the repair (or reparans). It may optionally include explicit editing terms, i.e. words or phrases used by the speaker to signal the correction (e.g. a discourse marker like "enfin"). The repair is the continuation of the message that follows the disfluency, so that if the first two regions are removed the remainder is lexically fluent, grammatically correct, and appropriate based on the situational constraints and communicative intentions of the speaker.

The interruption point may be right after the reparandum (the word or words that will be replaced). In 20% of the cases in the corpus studied by Levelt (1983), the interruption took place before completing the word replaced (the last item in the reparandum was an unfinished articulation). In other cases, the speaker may continue with one or more words after those that will be replaced. The repair may start with the word to be replaced in the reparandum and Levelt (1989) calls these cases immediate repairs. In other cases, the repair starts one or more words before the corrected/replaced word, in which case the speaker performs an anticipatory retracting in Levelt's terminology.

What triggers a self-correction? The speaker may choose of more or less informative term (a hypernym or a hyponym), a synonym, or a completely unrelated lexical item or items. Levelt (1989) distinguishes between error repairs, where the speaker has violated a grammatical rule, has produced a different word form from the one intended, or has mispronounced the word; and appropriateness repairs, where the reparandum would be grammatically correct, but was deemed not appropriate in context by the speaker.

A further distinction can be made between overt and covert repairs. Repetitions of function words could be interpreted as signalling the intention of the speaker to utter a word form that was replaced before it was actually uttered (e.g. the repetition of the indefinite article in the utterance "there is *a a* blue line" could be interpreted as indicating that the speaker had planned to say "there is a red line" but replaced the "a red" with "a blue" before articulating the word "blue"). According to Postma and Kolk (1993), "self-repairing of speech errors demonstrates that speakers possess a monitoring device with which they verify the correctness of the speech flow. There is substantial evidence that this speech monitor not only comprises an auditory component (i.e., hearing one's own speech), but also an internal part: inspection of the speech program prior to its motoric execution. Errors thus may be detected before they are actually articulated".

They advanced the Covert Repair Hypothesis, according to which "disfluencies reflect the interfering side-effects of covert, pre-articulatory repairing of speech programming errors on the ongoing speech. Internally detecting and correcting an error obstructs the concurrent articulation in such manner that a disfluent speech event will result. Further, it is shown how, by combin-

ing a small number of typical overt self-repair features such as interrupting after error detection, retracing in an utterance, and marking the correction with editing terms, one can parsimoniously account for the specific forms disfluencies are known to take" (Postma & Kolk, 1993, p. 472).

An alternative view is presented in Clark and Wasow (1998), who analysed repetitions (possible combined with filled and silent pauses) of articles and pronouns in two large corpora of spontaneous speech. An utterance like "I uh I wouldn't be surprised at that" is analysed as having four regions: "an initial commitment to the constituent (with 'I'); the suspension of speech; a hiatus in speaking (filled with 'uh'); and a restart of the constituent ('I wouldn't...')". They find that "speakers are more likely to make a premature commitment, immediately suspending their speech, as both the local constituent and the constituent containing it become more complex. They plan some of these suspensions from the start as preliminary commitments to what they are about to say. And they are more likely to restart a constituent the more their stopping has disrupted its delivery", arguing that these observations are general principles applied to the reparandum, interregnum and reparans of any self-correction of this type and not only limited to repeats.

Beyond these efforts to explain the structure of self-corrections with a small set of rules or principles, there is ample evidence that speakers are able to monitor their own speech before it is articulated. The alternative hypothesis, i.e. that speakers only monitor their own speech after they have articulated it and interrupt themselves when they perceive and error, would be incompatible with the finding that oftentimes the reparandum and the repair are contiguous (there should have been a delay introduced by speech perception). Furthermore, studies have shown that speakers are able to detect errors in their own speech even when it is masked by loud noise (Lackner and Tuller, 1979; Oomen and Postma, 2001 cited in Hartsuiker, 2014, p. 418).

Based on these observations, Hartsuiker and Kolk (2001) proposed the Perceptual Loop Model, which builds upon Levelt's (1983, 1989) model of speech production and perceptual loop theory. There are three components in this model: detection of trouble in speech, interruption and repair.

According to the perceptual loop model, speakers possess a cognitive mechanism that uses two channels to detect trouble: the external or post-articulatory channel that listens to overt speech and the internal or pre-articulatory channel that monitors the speech planned but not yet articulated ("inner" speech). When the cognitive process of detection deems that a self-correction is needed, two processes are started: the process of interrupting and the process of repair. Based on Logan & Cowan's (1984) theory of inhibition, the model postulates that interrupting is an inhibition process, and therefore takes time by itself. According to Logan and Cowan (1984) "a control signal, such as an external stop signal or an error during performance, starts a stopping process that races against the processes underlying ongoing thought and action. If the stopping process wins, thought and action are inhibited; if the ongo-

ing process wins, thought and action run on to completion". This explains why sometimes speakers continue articulating more words (and/or a word fragment) than those that will be replaced. The process of interruption and the process of repair are assumed to start simultaneously and take place in parallel, which explains why sometimes the repair takes place immediately after the interruption without any delay (silent or filled pause). It is further hypothesised that when there is an overlap in meaning or form between the reparandum and the repair, this will speed up the repair process Hartsuiker (2014, p. 419).

In the following three subsections we present relevant literature findings for each of the three components of self-repair (detection of trouble in speech, interruption and repair).

5.3.1 *Detection*

While Nooteboom (1980) notes that in spontaneous speech 50% of lexical errors are not repaired, it has been suggested (e.g. Levelt, 1983) that all aspects of language production are subject to self-monitoring and repair. Speakers may correct themselves at the message level: when a speaker changes his mind, this will typically result in an interruption and fresh start; the speaker may respond to changes in the external environment while referring to an object that has changed; or the speaker may choose to make the message more informative or more specific. Self-repairs affect all levels of language: they may correct lexical, morphosyntactic, prosodic, or phonological errors. Hartsuiker (2014, p. 420) points out that "repairs in the speech rate or to the quality of articulation are usually triggered by interlocutors", but the cues by the interlocutor may be non-linguistic (e.g. gestural). The Lombard effect (the involuntary tendency of speakers to increase vocal effort when speaking in noisy environments) shows that the speakers are monitoring their loudness and intelligibility. In general, feedback from an interlocutor is an important factor in the self-monitoring system: we can hypothesise that speakers will engage in more self-repairs in a communicative situation than in a non-communicative one.

Postma (2000) lists 11 monitoring channels, at different stages and levels of language production. Figure 5.4 shows how these channels operate and intervene at the various levels in Levelt's model of speech production. It is not claimed, however, that there are eleven distinct cognitive mechanisms; these are all the different channels that have been proposed in the literature. At the conceptualiser level, monitoring concerns the appropriateness and whether the preverbal message corresponds with the speaker's intention. At the grammatical encoding stage, there is monitoring at the lexical selection level and at the syntactical structure level. Postma (2000) predicts the existence of an articulatory buffer to absorb the effects of asynchrony between speech planning

and execution. Two further channels (the auditory loop and the knowledge of results) are linked to the perception of one's own overt speech.

Different theories of speech monitoring exist, that can be categorised depending on the role and function they attribute to each of the monitoring channels. One categorisation is between editor theories, that assume that the output of production components is fed through an extrinsic device, such as speech perception, and connectionist theories, which assume that the monitoring is an intrinsic property of the production system (Hartsuiker, 2014, p. 422). Theories can also be distinguished according to the role they attribute to the perception system. Perception monitoring theories assume that detection uses the speech perception system, and therefore is restricted by attentional focus and capacity. The Perceptual Loop Monitor theory is a perception-based theory, making the (controversial) claim that both post-articulatory and pre-articulatory monitoring mechanisms depend on speech perception. Production-based theories (e.g. van Wijk & Kempen, 1987) postulate that the pre-articulatory mechanisms are based on information channels internal to the production system.

There is evidence that self-monitoring deteriorates when the speaker is under cognitive load or when the speaker cannot hear himself. Oomen and Postma (2001) examined the effects of divided attention on the production of filled pauses and repetitions. Based on the dual-task paradigm, their subjects performed a picture story-telling task, with and without simultaneously performing a tactile form recognition task. The number of filled pauses and repetitions increased in a situation of divided attention. They conclude that the production of the filled pauses and repetitions themselves "is governed by processes that operate relatively independently of the available attentional resources", but that these disfluencies might be "automatic reactions to the increased planning difficulties induced by the concurrent task".

5.3.2 *Interruption*

When do speakers decide to interrupt their speech flow? Nooteboom (1980) proposed the "main interruption rule", according to which speakers try to interrupt themselves as soon as they can. We have seen above that the Perceptual Loop Model suggests a "stop signal" in the sense of Logan and Cowan (1984). Xue, Aron, and Poldrack (2008) have shown that stopping verbal and manual actions activates the same areas of the brain. A corpus analysis by Levelt (1983) supported the main interruption rule: the corpus analysed contained approximately 20% of word-internal interruptions. However, Levelt (1989) points out that these words were part of the reparandum, whereas words that were not errors (in the case of words following the error and in the case of appropriateness repairs) were not interrupted. A modified version of the main interruption rule would be that "speakers interrupt as soon as they can unless this interrupts a correct word" (Hartsuiker, 2014, p. 425).

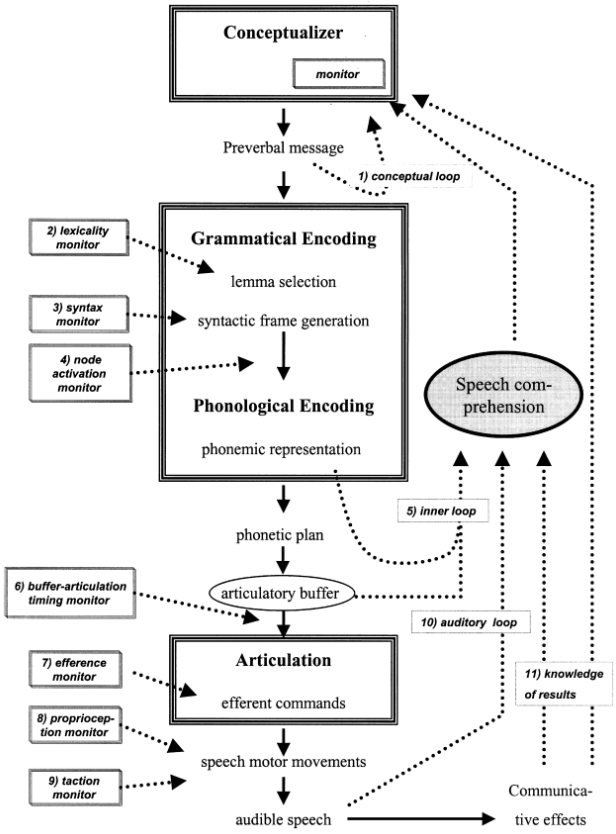


FIGURE 5.4: Monitoring channels in production (From Postma, 2000)

However, speech elicitation experiments using the changing world paradigm (where the speaker is describing a scene that suddenly changes) indicate that the situation is more complex: the location of the interruption depends on the repair; the repair and the interruption are co-ordinated. Tydgat, Stevens, Hartsuiker, and Pickering (2011) conducted a series of experiments using a picture naming task and the changing world paradigm. They found that "in contrast to the predictions of Levelt (1983), people stopped both within correct and incorrect words. The proportions of word-internal interruptions and word completions were influenced by the task instructions" (i.e. whether the subjects have been instructed to resume their speech or not, depending on the presence of an interlocutor etc.). They argue that speakers make a strategic choice and sometimes decide to postpone their interruptions and complete their words instead of interrupting as fast as possible. Hartsuiker, Pickering, and de Jong (2005), using a picture naming task and the changing word

paradigm, found that "planning a new word can begin before the initial word is abandoned, so that both words can be processed concurrently".

Seyfeddinipur, Kita, and Indefrey (2008) performed a corpus analysis of 448 self-repairs and measured the interval between the interruption point and the repair. They found that "speakers interrupted themselves not at the moment they detected the problem but at the moment they were ready to produce the repair". They argue that this is explained based on the communicative effects that each choice would have. Speakers who interrupt their speech immediately will maximise accuracy, speakers who continue until being able to produce a repair will maximise (perceived) fluency. Seyfeddinipur et al. (2008) concluded that "speakers preferred fluency over accuracy".

The moment speakers will interrupt is related, but not necessarily identical to the moment they detect trouble. Hartsuiker, Catchpole, de Jong, and Pickering (2008) argue that difficult repairs take more time to plan and execute than easy repairs because they compete for the same limited cognitive resources. In the case of increased cognitive load, the production of the repair would be further delayed. We would therefore hypothesise that under high cognitive load, self-repairs will have a tendency (a) either to include more words in the reparandum following the words that will be eventually replaced in the reparans or (b) more frequent presence of silent and filled pauses in the interregnum.

5.3.3 *Repair*

Levelt (1989) initially proposed the "well-formedness rule" that states that the original utterance plus the repair is well-formed if a version of the original utterance, completed with a hypothetical constituent if needed, is well-formed. However, textcitevanWijk.1987 found many violations of this rule in their corpus. They suggested instead that there are two ways of repairing: reformulation (which respects the well-formedness rule because it entails syntactic encoding) and lexical substitution (which does not respect the well-formedness rule; the listener will perceive that the replaced word should be understood within the currently used syntactic structure).

Boland, Hartsuiker, Pickering, and Postma (2005) used picture naming of simple geometric objects, and changed the context to elicit substitutions (e.g. "the light blue square" instead of "the blue square"). They found that repairs that involved adding a word took much longer than repairs that deleted a word. They argue that the findings suggest that speakers access a stored representation of their speech plan and make changes to it, which take more time if they are more difficult to make. Hartsuiker, Pickering, and de Jong (2005) found that the semantic and phonological relatedness between the reparandum and the repair affects repair time (when the reparandum and the repair are related semantically or phonologically, the repair is easier to produce and takes less time).

We argue that, beyond the global observations that have been exposed so far, each speaker will have an individual profile of production of self-repairs, and more generally an individual disfluency profile. We expect that higher levels of cognitive load will increase the production of disfluencies, but we advance the hypothesis that this should be studied against the baseline of each individual speaker, given that each speaker will favour a different self-repair strategy.

5.4 PROSODY AND DISCOURSE STRUCTURING

In Chapter 4 we set the framework of examining prosody at four levels (acoustic, articulatory, perceptual, linguistic) and mentioned that prosody fulfils numerous linguistic and communicative functions. In this section we will focus on the function of prosody to signal discourse structure, from a production point of view (cf. section 6.4 for a discussion from the perceptual point of view).

Prosody signals structure: at the word level, prosodic cues are used by the speaker and the listener to segment continuous speech into lexical units; at the syntactic level, there is a complex interaction between prosody and syntax; and at the discourse level, prosodic structure is used to encode rhetorical structure. Prosody also signals meaning: prosodic cues are used to indicate information status (e.g. salient elements in an utterance), and the illocutionary force of utterances (e.g. questions, exclamations, citations etc.). According to the implicit prosody hypothesis, even during silent reading, readers generate a prosodic representation of sentences that guides their interpretation (J. D. Fodor, 2002). No model of speech production can be complete without an account of how a speaker produces prosody, or as J. D. Fodor (2002) puts it "psycholinguistics cannot escape prosody". However, there has been very little empirical investigation into the role of prosodic phrasing in speech planning.

Prosodic structuring is a central issue in the role of prosody in speech planning. Most theoretical frameworks postulate a hierarchically organised prosodic structure, in which units are grouped by prosodic boundaries of varying strengths. These systems differ on the number of levels they postulate, although there is general consensus on two or three levels: major intonation unit – intermediate phrase – prosodic word (e.g. Shattuck-Hufnagel & Turk, 1996; Cutler, 1997; Frazier et al., 2006). A related concept is prosodic prominence: some units stand out of their environment due to acoustic cues. Prominence is used to signal information status, but also to indicate prosodic boundaries. The way these cues are used by the speaker and interpreted by the listener is language-specific. Selkirk's Strict Layer Hypothesis postulates that prosodic structure is non-recursive (Selkirk, 1984; Nespor & Vogel, 1986) but there is experimental and corpus-based evidence against this hypothesis: studies have found that acoustic and articulatory cues to prosodic boundar-

ies can be produced with different strengths depending on the depth of their embedding (e.g. Byrd & Krivokapić, 2008). Acoustic cues to prosodic boundaries and prominence include subsequent pauses, intonation contours and lengthening of final segments. We review related work, especially related to French, in Part II of the present thesis.

Furthermore, the relationship between prosodic structure and syntactic structure is a complex one. Some theories postulate that syntax drives prosody, i.e. that the prosodic structure can be largely inferred from syntactic structure (e.g. Selkirk, 1984; Nespor & Vogel, 1986). Other theoretical frameworks describe prosodic structure on the basis of prosodic features, such as intonation, duration and pauses (Mertens, 2008; Di Cristo, 2011; Delais-Roussarie et al., 2015), and postulate that there may be congruence or incongruence between the prosodic and syntactic structure (e.g. Mertens & Simon, 2013; Christodoulides & Simon, 2015).

Studies on the effects of syntax on prosody have shown that complex syntactic structures lead to the production of longer pauses (e.g. Cooper & Paccia-Cooper, 1980; F. Grosjean, Grosjean, & Lane, 1979; F. Ferreira, 1991; Strangert, 1997). This has been interpreted as showing that complex syntactic structures are more demanding on the production system, and that syntactically complex utterances are preceded by longer pauses because the speaker needs more time to plan the utterance. For instance, F. Ferreira (1991) studied paused durations between a subject noun phrase and an object noun phrase, and found that they increase with the complexity of the object noun phrase, but only if the subject noun phrase is also complex. F. Ferreira (1991) interprets this finding as indicating that speech planning proceeds in chunks.

Other studies show that prosodic structure is better predicted when both prosodic information and syntactic structure are taken into account. Gee and Grosjean (1983) found that pause duration can be better predicted if both syntactic and prosodic structure is used, rather than just syntactic structure. Furthermore, prosodic structure can override syntactic structure (e.g. F. Ferreira (1993) studying pause duration and final lengthening), and there is evidence that phrasal length is an even more robust evidence of the influence of syntactic structure on pause duration and speech planning (Krivokapić, 2012).

What are the implications for speech production models? Firstly, the relationship between prosodic structure and syntactic structure is linked to the issue of incrementality. Speech planning and speech production is incremental: speakers do not plan the whole utterance before the onset of speech; instead, planning and production proceed simultaneously (V. S. Ferreira, 1996; F. Ferreira & Swets, 2002; Keating & Shattuck-Hufnagel, 2002; Levelt, 1989). There is no consensus on whether the speech production system is "architecturally incremental" which would mean that speakers will obligatorily start speaking the moment a minimal unit is encoded. F. Ferreira and Swets (2002) show that speech production is highly incremental; and that speakers prefer to plan further ahead instead of starting to articulated immediately after having en-

coded the first chunk of the utterance, when they are given the possibility to do so. They therefore argue that speech is strategically incremental, i.e. that speakers are able to plan more than one planning unit at a time and that, when possible, they plan several chunks ahead before the onset of speech.

A further issue is the relationship between prosodic planning and grammatical encoding. Levelt's model favours a prosody-last approach (Levelt, 1989, 1999) and suggests that speech production is strictly incremental and without look-ahead. Levelt's model includes a Prosody Generator component, which comes into play after the syntactic structure has been fully encoded. The possible locations of intonation phrase (major unit) boundaries are constrained by the (already available) syntactic structure; the speaker choses to insert a boundary one word at a time with very little look-ahead. An alternative theory is the prosody-first approach developed by Keating and Shattuck-Hufnagel (2002), suggesting a large look-ahead in speech production. According to this model, a rough outline of the prosodic structure ("skeletal default prosody", p. 139) is created based on syntactic information, before the phonological encoding stage; the length of a prosodic unit is determined by the lengths of the previous prosodic unit and the next prosodic unit.

Krivokapić (2012) performed an empirical investigation on the role for prosodic structural complexity on pause duration. In a series of experiments, she examined pause duration preceding very long phrases (28 syllables); the effects of phrases of different length (6 to 14 syllables) on prosodic structural complexity; whether these effects are observed when the utterances are strings of numbers (to study the influence of semantic information); and whether these effects are locally constrained or more global. Krivokapić's (2012) results show that speakers that speakers have a large scope of planning and that both local and distant prosodic phrases have an effect on the speech planning process. Krivokapić argues that prosodic structure determines the chunk to be planned by speakers, who are capable of a large look-ahead. Speakers do not start speaking as soon as a minimal production unit is ready, but rather plan a large chunk, at least to a certain degree, before they start articulating. These findings are interpreted as arguments in favour of strategic incrementality and against architectural incrementality (using the terminology of F. Ferreira and Swets, 2002), and as favouring Keating and Shattuck-Hufnagel's (2002) prosody-first approach. Prosodic complexity was also found to have an effect on speech planning time, such that for longer phrases, prosodically complex structures lead to shorter pauses, while for shorter phrases, a complex structure leads to longer pauses.

D. Watson and Gibson (2004) review several theories of how syntactic/semantic structure influences the placement of intonational boundaries. They conclude that previous theories may be quite successful in their predictions, but they are complex, and incompatible with recent evidence for incrementality in sentence production. They propose a simpler incremental model called the Left hand side/Right hand side Boundary hypothesis. According to this

hypothesis, two factors contribute to the likelihood of producing intonational boundaries at word boundaries: (1) the size of the recently completed syntactic constituent at a word boundary; and (2) the size of the upcoming syntactic constituent. These factors are further constrained by syntactic argument relationships. They also argue that discourse status of relative clauses is an additional factor in determining the placement of intonational boundaries. However, F. Ferreira (2007) argue that algorithms designed to predict phenomena such as pauses or intonational breaks are problematic because they tend to conflate prosody and planning. According to F. Ferreira (2007), experimental work suggests that prosodic effects are based on prosodic constituency to the left of a potential boundary, and hesitations are due to planning of syntactic and semantic constituents to the right. Therefore, she claims that any adequate algorithm must distinguish between prosody and performance, prosodic and syntactic-semantic constituency, and planning and execution effects. Finally, corpus studies (e.g. Degand & Simon, 2009) have shown that the relationship between prosodic and syntactic boundaries varies systematically across communicative situations (speech "genres"). However, this finding alone does not necessarily mean that the differences are due to different planning strategies; it may be due to execution constraints.

In light of the previous discussion, we formulate the hypothesis that under cognitive load, there will be a marked increase in mismatches between prosodic and syntactic boundaries. We expect to observe in increase in mismatches because the constraints imposed on working memory will interfere with the look-ahead process, as described for example in Krivokapić (2012). In quantitative terms, we expect to find an increase in the number of occurrences of major prosodic boundaries inside minor syntactic units (chunks), i.e. in positions where there are normally not expected. This is consistent with the more generic hypothesis about an increase in the duration of silent and filled pauses under cognitive load, but it is a more precise hypothesis: it is not only the number and duration of pauses (or "hesitation" pauses as they have sometimes been called) that is important, but rather their placement within the syntactic structure, and the consequences that this will have for the prosody-syntax interface.

6

SPEECH PERCEPTION AND COMPREHENSION

In this chapter we will briefly review aspects of speech perception and comprehension, as they relate to what has already been presented. More specifically, in section 6.1 we review basic findings about the auditory perception of speech. Subsequently, we present a summary of the literature on the on-line comprehension of spoken language (section 6.2). The next two sections, on the role of working memory and automaticity (section 6.3) and the role of prosody in discourse structuring (section 6.4) present the perception counterpart of what was described in the previous chapter. The theoretical background of this chapter is related to Study 3 (Simultaneous Interpreting, chapter 20) and Study 4 (more specifically the Syntactically Unpredictable Sentences repetition task, chapter 21).

6.1 AUDITORY COGNITION AND SPEECH PERCEPTION

Before we can process spoken language, the auditory system plays an important role in sensing and interpreting auditory patterns. Auditory processing can be divided into two main stages: auditory perception and auditory cognition.

Auditory perception requires both hearing, which is a sensory process, and interpretation, which is perceptual and cognitive process (Baldwin, 2012, p. 31). The auditory system is composed of two primary components: the peripheral hearing structures (the outer, middle and inner ear) and the central auditory pathways located primarily within the cerebral cortex (p. 37). Essential auditory processes include sound localisation; auditory space perception, which is our ability to develop a spatial representation of the world based on auditory information; auditory scene analysis, which is the process of interpreting this auditory space. A subject of debate has been whether auditory stream separation requires attention. Macken, Tremblay, Houghton, Nicholls, and Jones (2003) present evidence that auditory stream segregation happens outside focal attention (Baldwin, 2012, p. 50). However, performance in a task requiring attention, such as serial recall, is severely disrupted by babble speech (simultaneous presentation of two or more irrelevant speech signals) indicating that the process of segregation requires some mental re-

sources, even if it is mainly automatic. It has also been found that the disruption decreases when the number of simultaneous voices increases, which may indicate that the disruption is the result an automatically activated attempt to process language. Another essential process in auditory perception is the continuity effect, i.e. "our ability to perceive a continuous and coherent stream of auditory information in the face of disruption from simultaneously overlapping acoustical information" (Baldwin, 2012, p. 51). A special case of the continuity effect is phonemic restoration: when a phoneme in a sentence was replaced by a click or cough, the listeners would "fill in" the missing phoneme (Warren, 1970). The continuity effect provides evidence for the importance of context in auditory processing, and consequently reliance on top-down cognitive processes.

Auditory cognition begins when we attend to an acoustic stimulus. Selective listening is a form of selective attention, which has been studied extensively with the dichotic listening paradigm, and more recently with brain imaging techniques. In the dichotic listening paradigm, subjects are asked to attend to only one of two simultaneously presented sound messages (often each to each ear), and answer questions or shadow the attended message. It has been found that the more two messages resemble each other, the more difficult they are to separate (Moray, 1969, cited in Baldwin, 2012, p. 55). Listeners are generally only able to report overall physical characteristics of unattended messages, such as whether it was speech by a male or female speaker. Humans have the ability to select and attend to a specific auditory pattern in the midst of competing patterns; the chosen pattern can be located and held in memory long enough to interpret it along several simultaneous dimensions (Baldwin, 2012, p. 70). When the auditory pattern is speech, these dimensions include cues about the age, gender and emotional state of the speaker. In summary, bottom-up and top-down processing, auditory stream segregation, auditory selective attention and temporary storage of acoustic patterns are all processes that demand cognitive resources; to the extent that aspect of these processes are less automatic and more attention-demanding, cognitive load may have a disruptive effect on them.

6.2 LANGUAGE COMPREHENSION

Language processing is something most of us engage do with great speed and skill, naturally and without finding it particularly challenging. However, the study of language reveals a great complexity: language is structured on many different levels and each of them is inherently ambiguous. In order to understand an utterance, initially, the listener needs to recognise speech, to segment a continuous sound wave into segments and select the most plausible interpretation of these segments, a process which is driven by both bottom-up perception of acoustic cues, and by top-down expectations of the listener about the lexical items that are more probable in a given context. Next, the

syntactic structure of the utterance needs to be established. Once again, the speaker will have to select between different possible interpretations, as the same lexical form may have different syntactical functions depending on its context and its relationships and dependencies with other lexical items. Finally, discourse comprehension is performed by the listener, who needs to resolve referential expressions (e.g. pronouns), link referents to objects and concepts and continuously integrate new information into a coherent whole. The listener constantly draws upon background knowledge to comprehend speech. "Perhaps the most fundamental issue in language comprehension is understanding how 'top-down' knowledge, not provided by the relevant aspect of the stimulus, is accessed and used" (Garrod & Pickering, 1999, p. 3).

In order to study language comprehension, the standard psycholinguistic approach is to break down this complex system into smaller components and describe each of them: lexical processing, syntactic processing, discourse processing, and dialogue and interaction management. We will review central issues in each of these stages.

6.2.1 *Lexical processing*

Lexical processing tries to explain how the meaning of each word is accessed and recovered during language comprehension. There is consensus on the special status of words as the main unit of the mental lexicon, but there are different theories about the minimal unit of lexical processing. A structural approach defends the position that the process of language comprehension is modular (cf. J. A. Fodor, 1983) and is performed sequentially, starting with morphemes, moving up to words and sentences. However, Marslen-Wilson and Tyler (1987) point out that the task is so complex that the only way for the human brain to be efficient at it is to use all available resources simultaneously and in parallel. According to this position, lexical access also draws upon syntactical, semantic and pragmatic information; the process is not modular, as this would preclude using information from other levels.

Spoken words "are encountered as parts of a transient speech stream with the component sounds heard in a temporal sequence" (Garrod & Pickering, 1999, p. 5), so any model of speech comprehension has to take into account the temporality of spoken language (Auer, 2009). The cohort model of spoken word recognition (Marslen-Wilson & Welsh, 1978) postulates that all words consistent with the pattern of the speech segments recognised so far are activated; recognition occurs when the cohort of activated words is reduced to one. In a modular system, we would expect all meanings of all activated words to be activated before recognition (since the lexical processing stage would not have access to disambiguating semantic information). An alternative view is the interactive view of lexical recognition, according to which meaning and context may be activated before the word itself has been recognised (there is interaction between levels of linguistic representation). The experimental

evidence supports the latter view (e.g. Moss & Gaskell, 1999). The historical debate between modular and interactionist theories in language comprehension has been largely settled in favour of interactionism. Current research focuses on how exactly the different sources of information are combined, and the time course of this process.

6.2.2 *Syntactic Processing*

At the level of syntactic processing, words are combined into meaningful sentences. A large body of work has examined the question of how the processor resolves syntactic ambiguities. These ambiguities may be global, i.e. remain after the entire sentence structure has been constructed, or local, i.e. one analysis seems possible in the beginning of reading or listening to the sentence, and after encountering a word, the reader or listener has to select an alternative analysis. There are two main questions related to on-line syntactic processing. Firstly, whether these analyses are considered one at a time and sequentially (the listener therefore re-analyses the sentence shortly after the word resolving ambiguity has been recognized) or considered in parallel. And secondly, whether the listener initially favours the simplest analysis, the most (semantically) plausible analysis, or the most frequent analysis; and whether all or only some sources of information can be used in the time course of parsing decisions.

There is clear evidence that language comprehension, and sentence analysis in particular, is extremely incremental: both syntactic analysis and semantic interpretation take place as soon as every new word is encountered (Garrod & Pickering, 1999, p. 7). This extreme incrementality in language processing has implications for our analysis on the role of prosody, and more particularly on the role of prosodic units and their relationship with syntactical units. It means that, while units may be incrementally constructed, as new lexical, syntactic and prosodic cues are encountered, we should not hypothesise that higher levels of processing are deferred until each unit is complete. Higher levels of processing, such as discourse processing, may be engaged early on, and perform their analysis on the basis of temporarily incomplete units. Drawing an analogy with the interaction between lexical and syntactical processing (where syntactical processing may modify lexical selection), on-line discourse comprehension may have a bidirectional interaction with the integration of prosodic and syntactical cues.

6.2.3 *Discourse processing*

Full comprehension of language is not limited to lexical and syntactic processing. The listener (or the reader) has to relate the meaning of the different elements to the world and between them (i.e. to resolve references); he has to establish the significance of each individual sentence or utterance and link

it to the message that the speaker is trying to convey, possibly refining the significance in the process (e.g. decide whether a word should be understood literally, metaphorically, or as part on an idiom); and finally, the listener needs to integrate the incoming stream of sentences in a cohesive and coherent message, and keep track of dialogue interaction.

A key issue is whether these aspects of discourse processing are specific to language, or whether they reflect more general cognitive processes that are used to interpret and understand other modalities, such as pictures and films (Garrod & Pickering, 1999, p. 8). It seems logical that general cognitive constraints also apply to the understanding of language. Gernsbacher (1991) describes the structure building framework, according to which the goal of comprehension is to build a coherent, mental representation, or structure, of the information being comprehended. According to this framework, the comprehender will initially create a foundation for their mental structures; then they will map incoming information that is coherent or related to previous information to that structure. However, if the incoming information is less coherent or related, they will shift and initiate a new substructure; thus, most representations comprise several branching substructures. Gernsbacher's model is also based on the mechanisms of suppression and conceptual enhancement.

Finally, Garrod and Pickering (1999) argue that the study of dialogue uncovers aspects of language processing not revealed by the study of either language comprehension or language production in isolation. "In a conversational setting there is an immediate and intimate interaction between the process of producing and utterance and its concurrent interpretation in the dialogue context. Production is constantly modified by feedback from the listener; and the listener's subsequent production is strongly influenced by what he or she has just interpreted as a listener" (p. 10).

6.3 WORKING MEMORY AND AUTOMATICITY IN LANGUAGE COMPREHENSION

Under normal circumstances, recognising spoken words seems to be automatic, based on the criteria of awareness (of the process itself, instead of the result), intentionality, efficiency and interruptibility proposed by Bargh (1994). However, understanding low-intelligibility speech, such as when listening to a speaker with an unfamiliar foreign accent, hearing impairment or dysarthria, or when listening to speech synthesised by a computer, speech perception is more effortful.

Francis and Nusbaum (2009) conducted three experiments to examine the effects of intelligibility on working memory demands imposed by the perception of synthetic stimuli. They used a dual-task paradigm: the primary task was a speeded word recognition task, and the secondary task was designed to modulate the availability of working memory resources during speech perception. In the first experiment, speech intelligibility was varied by training

a group of listeners to make better use of available acoustic cues (in synthesised speech stimuli), compared to a control group. Training was found to significantly improve intelligibility and recognition speed, and working memory load significantly slowed recognition. There was a significant interaction between training and WM load, indicating that the benefit of training on recognition speech was observed only under low memory load. In two further experiments, speech intelligibility was manipulated by changing synthesisers and no special training was offered to listeners. Improving intelligibility without training improved recognition accuracy and increasing memory load still decreased it, but more intelligible speech did not produce more efficient use of available working memory capacity. The authors conclude that their results suggest that perceptual learning modifies the way available capacity is used, perhaps by increasing the use of more phonetically informative features and/or by decreasing use of less informative ones.

In recent years there has been increased interest in the role of working memory in sentence processing. The key question in this debate has been on whether the resources used during syntactic processing are different from the resources used for other more conscious verbal tasks (van Gompel & Pickering, 2007, p. 297). A shared resources account, such as the one proposed by Just and Carpenter (1992) postulates that all linguistic processes draw from the same limited pool of working memory resources. When the working memory capacity, either for storage or for processing, is exceeded this will lead to either a slow-down, or a failure to maintain linguistic information in memory. According to this model, individual differences in verbal working memory lead to individual differences in sentence processing. An alternative account for the role of working memory in sentence comprehension is given by Caplan and Waters (1999). They put forward the dedicated resources account, which "assumes that the working memory resources dedicated to obligatory and automatic linguistic processes such as sentence processing are different from those used for more strategic and controlled linguistic processes, such as those used in a reading span test" (van Gompel & Pickering, 2007, p. 297).

6.4 PROSODY AND DISCOURSE STRUCTURING

Speer and Blodgett (2006, p. 513) summarise current work on the relationship between prosodic phrasing and syntactic processing as follows: there is a consistent, early influence of a prosodic representation on the production and comprehension of syntactic meaning in sentences in discourse. More specifically, prosody has an immediate, rather than a delayed, effect on a range of syntactic parsing decisions, suggesting that prosody can determine the incremental construction of a syntactic representation on the fly during sentence comprehension. This suggests that prosodic structure plays a primary role in the process of recovering sentence structure. Furthermore, prosodic pro-

cessing effects are attributable to the use of a global prosodic representation, rather than to local markers in a lexico-syntactic representation (p. 514). For example, the effect of a particular local prosodic boundary is dependent on its position in the overall prosodic structure of an utterance.

Kjelgaard and Speer (1999) conducted experiments to investigate the extent to which co-operating or conflicting prosodic phrasings facilitate or interfere with the process of language comprehension. They used phrases with a local syntactical ambiguity ("garden-path sentences" such as "when Roger leaves the house is dark" / "when Roger leaves the house it's dark") that were uttered with a co-operating prosodic form (i.e. facilitating the resolution of the local syntactical ambiguity – e.g. "when Roger leaves | the house is dark"), baseline and conflicting prosodic form (i.e. interfering with the resolution of the local syntactical ambiguity – e.g. "when Roger leaves the house | is dark"). Kjelgaard and Speer (1999) found that "the prosody of an utterance can have early and immediate effects on language comprehension. More specifically, the prosodic phrasing of an utterance can facilitate or interfere with early syntactic processing" (Speer & Blodgett, 2006, p. 516). Furthermore, Blodgett (2004) shows that intonation phrase boundaries have an immediate influence on syntactic processing regardless of lexical biases.

Cutler (1997, p. 185) notes that "the evidence at this processing level speaks against any direct availability of syntactic information in prosodic structure; prosodic hierarchies encode prosodic and not syntactic relationships. Furthermore, a given utterance may allow alternative prosodic structures which are equally likely and equally acceptable. It is only recently that studies of the role of prosody in syntactic analysis have accepted this to be the case, and have begun to approach the issue without the hypothesis of a one-to-one mapping". She further notes that the large body of research on attachment preferences based on the visual modality (reading) may have been misguided, given that such ambiguities would not arise in the spoken modality.

In the area of discourse structuring, research clearly suggests that prosodic information contributes to the creation of a discourse model, by indicating the information status and salience of elements through prosodic prominence; however it has not been shown that prosody directs the construction of a discourse model (Cutler, 1997, p. 185).

7

COGNITIVE LOAD ASSESSMENT

In this chapter we review methods proposed in the literature for assessing the cognitive load imposed on a person by a specific task. Given that cognitive load is a multidimensional phenomenon depending both on the demands imposed by the task and on the skills and cognitive abilities of the person performing the task, measuring cognitive load is not straightforward like measuring a physical quantity. The methods reviewed in this chapter can be grouped in four main categories:

- Analytical methods, relying on subjective data (expert opinion) and/or mathematical models. These methods, however, do take into account individual differences.
- Subjective methods, using self-reported data to quantify the performer's perception of the difficulty of a task (e.g. introspection, retrospective or concurrent description).
- Performance methods, using a primary and a secondary task and trying to establish to what extent the latter affects the former (the dual-task paradigm).
- Psycho-physiological methods, monitoring the physiological processes that have been known to vary with changes in cognitive load. These methods can be more or less intrusive.

Brünken, Plass, and Leutner (2003) propose a classification of the various methods available for assessing cognitive load across two dimensions: objectivity (subjective or objective) and causal relation (direct or indirect). The objectivity dimension describes whether the method uses subjective, self-reported data or objective observations of behaviour, physiological conditions, or performance. The causal relation dimension classifies methods based on the type of relation of the phenomenon observed by the measure and the actual attribute of interest. Brünken et al. (2003) focus on the cognitive load imposed by instructional materials on learners, but their classification is generalizable; their examples of cognitive load assessment methods are presented in Table 7.1.

	Indirect	Direct
Subjective	Post-treatment questionnaires in which learners are asked to report the amount of mental effort invested	Rating of the difficulty of the materials (by experts)
Objective	• Performance outcome measures (in the area of multimedia learning, these are the knowledge acquisition scores) • Behavioural patterns and their relationships with the learning process • Physiological measures known to correlate with cognitive load (e.g. eye-tracking analysis)	Neuroimaging techniques (although the authors note that "the connection between memory load and prefrontal cortex activity is not fully understood")

TABLE 7.1: Examples of cognitive load assesment methods (From Brünken, Plass, & Leutner, 2003)

7.1 ANALYTICAL AND SUBJECTIVE METHODS

According to Paas, Tuovinen, Juhani, E., Tabbers, and van Gerven (2003, p. 66) "analytical methods are directed at estimating the mental load and collect subjective data with techniques such as expert opinion and analytical data with techniques such as mathematical models and task analysis". As an example, Seeber (2011) proposes a decomposition of the process of simultaneous interpreting into four sub-tasks (storage, perceptual auditory verbal processing, cognitive verbal processing, verbal response processing and interference) and constructs hypotheses about which syntactic structures in the original speech would impose greater cognitive load to the interpreter (e.g. verb-final constructions are more cognitively demanding because they would tax the storage sub-component).

Subjective indirect methods rely on questionnaires that collect self-reporter subjective measures. Paas, Tuovinen, Juhani, E., et al. (2003, p. 66) notes that "most subjective measures are multidimensional in that they assess groups of associated variables, such as mental effort, fatigue, and frustration, which are highly correlated (for an overview, see Nygren, 1991). Studies have shown, however, that reliable measures can also be obtained with unidimensional scales (e.g. Paas, van Merrienboer, & Adam, 1994). Moreover, it has been demonstrated that such scales are sensitive to relatively small differences in cognitive load and that they are valid, reliable, and non-intrusive (e.g. Gimino, 2002; Paas, van Merrienboer, & Adam, 1994)". As an example, Khawaja

(2010, p. 39) mentions the use of a 7-point or a 9-point Likert scale, inviting participants to "rate their level of mental effort in completing this task" (from "very, very low effort" = 1, to "medium effort" = 5, to "very, very high effort" = 9), or to "rate the difficulty of the task" (ranging from "very easy" = 1 to "neither easy nor difficult" = 4, to "very difficult" = 7).

Although self-ratings may appear questionable, it has been demonstrated that people are quite capable of giving a numerical indication of their perceived mental burden (Gopher and Braune, 1984, cited in Paas, Tuovinen, Juhani, E., et al., 2003, p. 66).

7.2 PERFORMANCE-BASED AND BEHAVIOURAL METHODS

Task- and performance-based techniques include two subclasses: primary task measurement, which is based on task performance, and secondary task methodology, which is based on the performance of a task that is performed concurrently with the primary task (Paas, Tuovinen, Juhani, E., et al., 2003, p. 66).

A performance measure on a primary task is based on the assumption that the complexity of the task is directly related to the cognitive load experienced by the performer, and that there will be negative effects on performance due to overload. This assumption however may be problematic: two persons may have identical performance on the task while one of them is making more effort than the other. Furthermore, if the subject is a highly-skilled expert on the specific task, it may be the case that no effects on performance will be detected using a typical task difficulty manipulation.

For this reason, the dual-task paradigm is frequently used, where the subject is performing a secondary task concurrently with the primary task. "In this procedure, performance on a secondary task is supposed to reflect the level of cognitive load imposed by a primary task. Generally, the secondary task entails simple activities requiring sustained attention, such as detecting a visual or auditory signal. Typical performance variables are reaction time, accuracy, and error rate". (Paas, Tuovinen, Juhani, E., et al., 2003, p. 66).

In a component-based model of working memory, simple secondary tasks that can be performed automatically (i.e. with low levels of control, cf. Chapter 3 and Section 5.2) may not be suitable, as they may not sufficiently interfere with the central executive and induce weak cognitive load. For this reason, secondary tasks that include a processing component (e.g. taking a decision based on a stimulus) are more appropriate for taxing the central executive (instead of merely the storage subsystems). Performance measures, like most physio-psychological measures are continuous, and therefore more useful when studying a phenomenon that is dynamically developing in time, such as speech production and perception.

Behavioural measures have been used when assessing the cognitive load imposed on users of complex multi-modal human-computer interfaces. The

underlying assumption is that, regardless of their level of performance and their subjective appreciation of their mental effort (i.e. regardless of data that could be collected with performance-based and subjective methods), users will interact differently with the system when under high cognitive load. Examples of behavioural measurement techniques are key logging (recording the precise timing of all keys pressed by someone typing a text on a computer), recording mouse movements and clicks on a computer or similar device, control inputs on a flight simulator etc. These methods have proven their use in the human-computer interaction (HCI) community, but are limited to situations in which a person is using a system sophisticated enough to induce cognitive load, and having enough inputs or controls that can be recorded.

7.3 PHYSIO-PSYCHOLOGICAL METHODS

Physio-psychological methods are based on the assumption that changes in the cognitive functioning are reflected in physiological measures. Physiological techniques include the measurement of heart rate activity, brain activity, and eye activity (for a review of some of these techniques, see Kramer, 1991). An important feature of physiological measures is that most of them are continuous, with varying degrees of temporal resolution; for instance, electroencephalographic (EEG) data has a more detailed temporal resolution than measurements of the galvanic skin response (GSR). Continuous data can be used to calculate averages over periods of time, over participants and over tasks. In the following, we will review studies that have used physio-psychological measures of cognitive load.

Brookings, Wilson, and Swain (1996) simulated three scenarios of varying difficulty with air traffic controllers: in the first scenario, they varied the traffic volume (number of aircraft to be handled), in the second scenario varied traffic complexity (arriving to departing flight ratio, pilot skill and mixture of aircraft types) and in the third scenario they induced overload by requiring controllers to handle a larger number of aircraft in a limited amount of time. They compared three different methods of assessing cognitive load: performance measures, subjective measures (using the NASA standardised TLX questionnaire) and physiological measures (EEG, eye blink, heart rate, respiration, and eye saccades). Their findings indicate that performance measures were affected by the load induced in different scenarios (thereby validating their method), and that the most sensitive measure were the reported subjective load in the TLX questionnaire, eye blink rate, respiration rate and EEG measures. Furthermore, they find that only the EEG data was associated with the main effects for the type of traffic, arguing that different workload measures have different sensitivity in measuring complex tasks. We will return to findings regarding eye blink rate in the following section.

Shi, Ruiz, Taib, Choi, and Chen (2007) used the galvanic skin response (GSR) to evaluate the stress and arousal levels of users while using unimodal (visual display only) and multimodal (including speech and gesture) versions of a computer user interface. The galvanic skin response is defined as changes in the electrical conductivity of the skin, reflecting autonomic nerve responses that trigger the sweat glands. Shi et al. (2007) report preliminary results showing that GSR measurement increase significantly when using a multimodal interface compared to a unimodal interface, and claim that GSR can "serve as an objective indicator of user cognitive load in real time, with a very fine granularity". However, it is known that skin conductance is primarily affected by sweating, which may be the result of several factors, many of them unrelated to cognitive load. Note that the so-called "lie detector" devices combine GSR measures with heart rate measures, and that these devices are notorious for their low levels of reliability.

Grimes, Tan, Hudson, Shenoy, and Rao (2008) present an experiment conducted to explore the feasibility of using electroencephalograph (EEG) data for classifying working memory load. They used a 32-channel active electrode EEG device; the tasks used to create different levels of working memory load were variations of the n-back task. Using "a novel feature selection scheme that seems to alleviate the need for complex drift modelling and artefact rejection", they report "classification accuracies of up to 99% for 2 memory load levels and up to 88% for 4 levels". They also report that the models built on one task could potentially generalise to other similar tasks.

In a study by Paas, van Merrienboer, and Adam (1994), heart-rate variability was measured and used to estimate the level of cognitive load. Paas, Tuovinen, Juhani, E., et al. (2003, p. 66) reports that they "found this measure to be intrusive, invalid, and insensitive to subtle fluctuations in cognitive load". Other studies had better success with electrocardiographic (ECG) data: for example, Haapalainen, Kim, Forlizzi, and Dey (2010) report that "the electrocardiogram median absolute deviation and median heat flux measurements were the most accurate at distinguishing between low and high levels of cognitive load, providing a classification accuracy of over 80% when used together".

While these findings show that physiological measures can be very successful as objective indicator of various types of cognitive load, they also highlight some important limitations. Some of these measures are intrusive or perceived as intrusive (e.g. using heart and EEG electrodes), which precludes their use in real-world situations, and may cause additional stress to subjects in an experimental situation. Physiological measures may be affected by other bodily functions (e.g. sweating, heart rate and respiration rate), and individual differences may override any effects of cognitive processing; and a physiological measure that yields acceptable performance in one controlled setting or task, may not be generally valid in other cases. A set of promising

non-intrusive physiological measures is provided by eye tracking and pupillometry devices, which is the focus of the next section.

7.4 PUPILLOMETRY

The task-evoked pupillary response (TERP) is a reflexive response of the eye's pupil caused by cognitive load. The eye's pupil dilates (its diameter increases) as a result of a decrease in the parasympathetic activity in the peripheral nervous system. It has been found that the pupil's dilation caused by the TERP is linearly correlated with the increase of the load placed on working memory. Beatty (1982) studied the TERP and found that it "provides a reliable and sensitive indication of within-task variations of processing load. It generates a reasonable and orderly index of between-task variations in processing load. It reflects differences in processing load between individuals who differ in psychometric ability when performing the same objective task". Beatty and Lucero-Wagoner (2000) identified three useful task-evoked pupillary responses (TEPRs): mean pupil dilation, peak dilation, and latency to the peak.

Pomplun and Sunkara (2003) note that although pupil size can be measured by most video-based eye-tracking system, it is rarely used in usability studies due to two factors: first, the pupil's dilation can be influenced by a variety of factors, such as changes to the luminosity in the visual field; and second, the fact that the size of the pupil as seen by the eye-tracker camera depends on the person's gaze angle, and therefore eye movements cause distortion to the size data. They propose algorithms to calibrate eye-tracking systems and reliably extract pupil size data, eliminating the geometry-based distortion. Klingner (2010) notes that "cognitive pupillometry has been generally limited to experiments using auditory stimuli and a blank visual field, because the pupil's responsiveness to changes in brightness and other visual details interferes with load-induced pupil dilations". He also proposes algorithms to smooth the raw data and use the output of video-based eye trackers for pupil size estimation.

Apart from pupil size, other measures of eye activity have also been proposed in the literature as methods to estimate cognitive load. Chen, Epps, Ruiz, and Fang (2009) note that previous TERP studies have focused on eye activity in settings atypical of human-computer interaction, and study the effects of cognitive load induced by the interaction with a computer system on eight measures derived from eye activity: blink latency, blink rate, mean pupil size, standard deviation of pupil size, fixation time, fixation rate, saccade size and saccade speed. Their results show that "all features are capable of discriminating different cognitive load levels. In particular, saccade size and speed were shown to provide very high levels of discrimination between low and medium levels of mental effort. Correlation analysis among the various

pairs of improvements in discriminating different effort levels can be made by combining multiple features".

Cognitive pupillometry has been successfully used to study language processing. Engelhardt, Corley, Nigg, and Ferreira (2010) focuses on language comprehension and investigated the "processing effort by measuring peoples' pupil diameter as they listened to sentences containing a temporary syntactic ambiguity". In their first experiment, the prosody of short sentences was manipulated, and the results showed that "when prosodic structure conflicted with syntactic structure, pupil diameter reliably increased". In a second experiment, they manipulated both prosody and visual context, finding that "when visual context was consistent with the correct interpretation, prosody had very little effect on processing effort. However, when visual context was inconsistent with the correct interpretation, prosody had a large effect on processing effort". They conclude that there is an interaction between the visual context (when a sentence refers to that context) that can modulate the effects of linguistic information such as prosody.

Zekveld and Kramer (2014) studied the effects interfering (masking) speech on language comprehension, across an intelligibility range from 0% to 99% correct. They found that "pupil dilation was largest at intermediate intelligibility levels, smaller at high intelligibility, and slightly smaller at very difficult levels. Participants who reported that they often gave up listening at low intelligibility levels had smaller pupil dilations in these conditions. Participants who were good at reading masked text had relatively large pupil dilation when intelligibility was low". This indicates that the pupil response is sensitive to language processing load and possibly reflects cognitive overload in difficult listening conditions.

8

METHODOLOGICAL CONSIDERATIONS

In this chapter we will examine some of the challenges in obtaining speech material suitable for the study of the effects of cognitive load. This challenge should not be underestimated. It is not uncommon to find references to cognitive load as an explanation for observations in studies on phonetics, prosody, or foreign language teaching and evaluation, but without a theoretical backing or definition of cognitive load these claims may be overly simplistic (e.g. counting the number of one type of disfluencies and attributing an observed increase to "some kind" of cognitive load). We argue that in order to study speech, multiple techniques and data should be used, each with their strengths and weaknesses (cf. P. Wagner, Trouvain, & Zimmerer, 2015).

8.1 SPEECH ELICITATION

Niebuhr and Michaud (2015) describe speech data acquisition as an underestimated challenge, and provide a review of issues and factors that need to be considered by researchers engaging both in speech elicitation in the laboratory, and in fieldwork. From the onset, one must recognise that "every speaker is unique, that no two recording situations are fully identical, and that human subjects participating in the experiments are no 'vending machines' that produce the desired speech signals by paying and pressing a button" (p. 2). Factors that need to be taken into consideration include: individual speaker differences (physiological, social and cognitive factors, differences in linguistic experience and skills, strategies and preferences, relationship of the speaker with the researcher), the tasks used to elicit speech (recording settings, strong and weak points of different tasks, within-task differences and potential artefacts), and recording conditions (equipment, selection of participants, materials, procedure, context).

Individual differences in speech may reflect physiological factors, such as age, body type and gender. For instance, there are physiological and anatomical differences between the male and female speech production apparatus and female voices tend to be higher-pitched and breathier. However, "gender differences in speech do not merely have a biophysical origin. Some differences are also due to learned, i.e. socially evoked behaviour, and the dividing

line between these two sources of gender-related variation cannot always be easily determined" (Niebuhr and Michaud, 2015, p. 4, citing Simpson, 2009). Furthermore, speakers adapt their speech to the speech of their interlocutors, a phenomenon known as phonetic entrainment, or phonetic accommodation. Accommodation in dialogue affects pitch and intensity (in absolute terms and their range), voice quality, speaking rate, vowel qualities and speech reduction (Pardo, 2006; De Looze, Scherer, Vaughan, & Campbell, 2014), as well as lexical and syntactical features of utterances (Nenkova, Gravano, & Hirschberg, 2008). Musical training also affects speech production and perception: for example, Schon, Magne, and Besson (2004) show that musical training facilitates pitch processing in both music and language (musicians are better than non-musicians in detecting pitch violations), and Parbery-Clark, Skoe, Lam, and Kraus (2009) show that musical training facilitates speech comprehension in noise (both cited in Niebuhr and Michaud, 2015, p. 4).

In a speech elicitation experiment, it is important to collect information about the linguistic background of the participants. An individual's speaking style is affected not only from situational and communicative factors (that the experimenter may attempt to control) but equally importantly from sociolinguistic factors. When studying a language like French where there is significant geographical variation (cf. Avanzi, 2014; Durand, Laks, & Lyche, 2009; Simon, 2012), or when drawing a subject population from a multilingual country such as Belgium, the cultural issues surrounding the speaker's perception of the language(s) they speak become even more important. The author had first-hand experience of this fact when inviting "French native speakers" to participate in experiments, only to discover how diverse the university's participant pool can be: given the important regional prosodic variation phenomena that have long been documented, it was necessary to limit the participants to those born in French-speaking areas of Belgium and France (the regional variation remains non-negligible, but it is hypothesised that cognitive load effects will override it, especially in the prosodic variables under study). An individual's contact with different languages and different regional varieties of these languages influences the way he or she speaks; some persons are more susceptible to influences of language contact than others (Niebuhr & Michaud, 2015, p. 7).

It is also essential to establish good communication between the investigator and the participants in an experiment. This is important to understand to what extent the differences observed reflect individual speaker strategies, and to what extent they were caused by different understandings of the experimental task. Different speakers may interpret the experiment instructions differently, and may consequently adopt different communicative strategies (Niebuhr & Michaud, 2015, p. 10). The laboratory setting, and the expectations of speakers about a laboratory setting, may influence their speaking style. For instance, participants may feel obliged to hyper-articulate and ad-

opt more expressive prosody "since this is a recording that others may listen to". The use of a microphone may prompt speakers to adopt a more formal speaking style, depending, among others, on their personality and individual factors such as being extrovert or not, feeling comfortable speaking in public or not etc.

Every speech elicitation task has its strong and weak points. Niebuhr and Michaud (2015, pp. 12-13) propose a typology with six types of speech materials: isolated logatoms or words; isolated sentences; read monologues; read dialogues; unscripted monologues; and unscripted dialogues. Their typology is based on five factors: the degree of control over experimental variables (i.e. dependent and independent variables) as well as other variables (control variables); event density: the number of analysable tokens per time unit; expressiveness; communicative intention: the speaker's concern to actually convey a message; and homogeneity of behaviour: the probability that the elicitation condition is defined in such a way that it leads speakers to behave in a comparable way (facilitating cross-speaker comparisons and replicability). There is no task that is ideal across all five factors.

Isolated word, logatoms or sentences provide the maximum experimental control, are selected on the basis of subsequent analyses, but they severely lack in communicative intention and, usually, expressiveness; they are best suited to study precisely defined phonetic phenomena under highly controlled conditions, but one should be careful about generalising the findings on other communicative situations or uses of language.

Unscripted dialogues (e.g. the CID corpus, Bertrand et al., 2008) may have a low event density, and therefore analysing them is time consuming. Niebuhr and Michaud (2015) believe that "the lack of control over experimental and other variables is particularly the case in the case of unscripted dialogues" (p. 13) and propose the use of read dialogues instead. They claim that read dialogues "combine an informal, expressive speaking style – which can be enhanced by using a corresponding orthography and font type – with relatively high degrees of communicative intention, homogeneous behaviour, event density, and a relatively high degree of control over experimental and other variables" (p. 14).

However, as P. Wagner, Trouvain, and Zimmerer (2015, p. 5) point out "one could certainly ask the question why mimicry may be regarded as problematic or even acceptable at all. We believe that acted speech does not constitute 'bad data'. We simply argue that mimicked speech in specific styles is a very poor substitute for authentic speech produced in these styles because we have reasons to believe they affect both perception and production". They give the example of a corpus of read dialogues between undergraduate students simulating business meetings, constituted ignoring the fact that the student did not yet have experience in the professional world.

Event density can be enhanced in unscripted dialogues by including prompts and adding control elements. A number of tasks take advantage of this design:

the Map Tasks (Anderson et al., 1991), the 'Shape Display Task' (Fon, 2006), appointment-making tasks, role-playing tasks and quizzes (e.g. the 'Picture Difference Task', Turco, Gubian, and Schertz (2011); the 'Joint Crossword Puzzle Solving' task, Crawford, Brown, Cooke, and Green (1994)). A more general family of tasks includes all instruction-giving tasks (e.g. the 'Picture-Drawing task', Spilková, Brenner, Öttl, Vondricka, and van Dommelen (2010); the 'Card Task', Maffia, Pellegrino, and Pettorino (2014)). According to Niebuhr and Michaud (2015, p. 15), instruction-giving tasks perform better across the dimensions of expressiveness and communicative intention than role-play tasks.

Unscripted monologues can be based on retelling picture stories (e.g. the Frog story), or can be elicited by asking speakers to recite lyrics, poems or traditional texts that they know. Another way to elicit expressive unscripted monologues is to record speakers during or after computer games (e.g. Mixdorff, 2004), a technique that allows the elicitation of keywords, or even manipulating the outcome of the game to generate emotional speech (Niebuhr & Michaud, 2015, p. 16).

P. Wagner, Trouvain, and Zimmerer (2015, p. 10) propose to examine the following list of key questions before embarking on experimental data collection:

- To what extent does the method (experimental setting or recording situation, level and kind of control, experimental hardware and limitations of measurement techniques) influence the result?
- Are our results gained from controlled settings supported by less controlled settings and vice versa? That is, do we need to question the external or internal validity of our data?
- Are our data and results relevant for communicative interaction 'in the wild', i.e. what is the ecological validity of our research?
- How can the observer's paradox be minimized?
- May small changes in the experimental setting lead to different results?

According to P. Wagner, Trouvain, and Zimmerer (2015, p. 10), the data used for research "need to be assessed independently of the hypothesis at question. We argue that the best way to approach research questions is by embracing a certain methodological pluralism. Once our results are similar or even identical across different methodological settings, our conclusions can be made with more confidence".

In speech elicitation, the experimenter must control for within-task differences. For example, time pressure, e.g. imposed on speakers asked to produce sentences at different speaking rates, may induce cognitive load and put the speakers under stress. In any task, fatigue and boredom may introduce within-task differences. These within-task differences may actually be

desired, if for example our research question concerns the effects of time pressure and cognitive load on reading sentences, or the effects of boredom on speech produced in the laboratory; otherwise, care should be taken to minimise them.

Recordings should be performed using suitable professional equipment, especially if the data is to be shared for further research and when phonetic and prosodic analyses are envisaged. Microphone characteristics (dynamic range, frequency response curve, polar pattern, pop filtering etc.) should be taken into consideration; usually head-mounted, directional or cardioid microphones are necessary to separately record participants in a dialogue, so that the signals remain analysable in the presence of back-channelling and overlap. The recording environment must have appropriate acoustic characteristics (e.g. avoid reverberation, even in fieldwork).

Niebuhr and Michaud (2015, p. 22) point out that depending on the objective of the study it may be necessary to select participants, or to avoid particular profiles: "producing spontaneous speech in the lab and producing spontaneous-sounding read speech both require a certain extroversion, fluency, language competence, and self-confidence; speakers should be preselected accordingly". Every recording situation will make the speakers more aware of the way they speak. This may trigger changes in the speech behaviour, or even an effort to please the experimenter by exaggerating these speech features that the speaker believes are sought by the experimenter (the observer's paradox). Debriefing participants is an appropriate way to detect individual strategies or beliefs that they may have developed over the course of the experiment; it is also standard ethical practice.

Finally, care should be taken to avoid artefacts, or unwanted phenomena such as those introduced by translating materials from other languages without appropriate cultural adaptation, or using monotonous tasks for long periods of time.

8.2 USE OF CORPORA

Speech corpora are valuable resources, and the advent of easier to use and cheaper recording devices, as well as the use of the Internet as a source for recordings, has created an explosive increase in the available data.

Recent research in prosody focuses on phono-stylistic situational variation in large corpora and abandons binary distinctions, such as the dichotomy of "read" and "spontaneous" speech. Goldman, Prsir, et al. (2014) analysed 9 speaking situations using four dimensions: audience (whether the speaker is physically present before of an audience), media (whether the speech is broadcast, creating an indirect audience), preparation (which indicates the degree of preparation afforded to the speaker), and interactivity (which indicates whether the main speaker may be interrupted). The results show that while no single prosodic measure is sufficient to separate and classify speak-

ing styles, a linear combination of several measures leads to a robust clustering of samples belonging to different genres. It also showed that an iterative, adaptive procedure is necessary to define speaking styles: "while the initial, top-down approach was to select samples in order to create a balanced corpus (based on a predefined array of situational features), subsequent data analysis led to the observation that samples within a given genre could and should be further classified in sub-genres. Therefore, the interplay of prosodic measures and situational features gave rise to an a posteriori subdivision of genres (bottom-up approach) in order to ensure compact definitions and to reduce the excessive heterogeneity of some speaking situations. Results show that sub-genre groupings transcend genre differences, and that some of them are related to common, controlled situational features. They also present evidence for groupings due to unpredicted, or hidden, situational features, like 'external time pressure', 'speech sequence duration', or 'solemnity / ritual conventions', that belong to the prototypical image of speaking style" (p. 108).

It would be tempting to construct a *meta-corpus*, i.e. a collection of samples from existing speech corpora, based on the hypothesis that speaking in specific communicative situations will necessarily induce "cognitive load" on the speaker. However, this would ignore the importance of individual variation in linguistic ability, cognitive processing and working memory, as well as the importance of practice for skilled behaviour. A speaking situation that is stressful for one individual is not necessarily (equally) stressful for another one. On the basis of typical corpus metadata alone, it would be impossible to attribute the differences observed in speech to cognitive load; multiple alternative explanations would be possible (e.g. social factors or stress).

We have nevertheless opted to use corpora extensively in two alternative ways: as a baseline for observations of speech and prosodic patterns across speaking styles; and as a development testbed for automatic annotation and analysis tools. For this reason, some contributions of this thesis are more general or even unrelated to the study of speech under cognitive load.

9

PREVIOUS STUDIES ON SPEECH PRODUCTION AND PERCEPTION UNDER COGNITIVE LOAD

In this chapter we will review some important previous studies on the effects of cognitive load on speech production and speech perception. Many of the production studies, presented in the first section, were undertaken with a view to training automatic classification systems to recognise cognitive overload in operators of complex systems (e.g. in aviation). The study of speech perception under adverse conditions, such as noise, acoustic degradation, or hearing loss, is a field related to speech perception under cognitive load, in the sense that it can be hypothesised that these conditions cause cognitive load, but as we will see in the second section, the effects may differ.

9.1 STUDIES FOCUSING ON PRODUCTION

Lively, Pisoni, van Summers, and Bernacki (1993) conducted a study on both production and perception of short utterances under varying levels of cognitive load, induced by a parallel visual tracking task. They asked subjects to produce short utterances in the form of "say h(vowel)d again" (e.g. "say had again", "say hood again"). The sentences were displayed on a computer screen while subjects were performing the Allen and Jex (1968) visual tracking task: using a joystick to keep a dot on the screen from hitting two vertical boundaries, while the computer inserts random deflection. Acoustic measurements were taken over the utterance and on the target vowels, in order to compare speech under different workloads. They found that, under the high workload condition, some speakers produced utterances with increased amplitude and amplitude variability, decreased spectral tilt and fo variability, and increased speaking rate. No changes in the first three formants (F1, F2, F3) of the target vowels were observed across conditions, for any of the speakers. In a subsequent perceptual identification experiment, they found small but significant advantages in intelligibility for the utterances that had been produced under workload for speakers who showed robust changes in speech production. They conclude that, during speech production, increased workload caused both laryngeal and sub-laryngeal adjustments in articulation and modification of the absolute timing of articulatory gestures; and that the ma-

jor factor affecting intelligibility seems to be the changes in amplitude and amplitude variability for utterances produced under workload. Lively et al. (1993) interpret their results in the context of Lindblom's Hyper- and Hypo-articulation (H & H) Theory (Lindblom, 1990), which stipulates that speakers adapt their speech to suit the demands of the environment and these modifications are designed to maximize intelligibility. However, note that this study was conducted with only 5 participants.

Berthold and Jameson (1999) point out that the use of computing devices may lead to attentional distractions and thus to an excessive cognitive load. An adaptive human-computer interface could modulate its output to avoid such overload, by monitoring the user's speech input for indicators of high cognitive load. They synthesise the results of 11 studies[1] on the effects of cognitive load on speech, which can be grouped into two categories: effects on production quality and effects on production rate. With regards to production quality, increased cognitive load was correlated with: an increase in sentence fragments (syntactical interruptions) in 4 out of 5 studies that included this measure; an increase in the number of false starts (2 out of 4 studies); an increase in syntax errors (1 out of 1 study); an increase in the number of self-repairs in 2 out of 7 studies, no change in the number of self-repairs in 1 out of 7 studies and a reduction in the number of self-repairs in 4 out of 7 studies. With regards to production rate, increased cognitive load was correlated with: a decrease in articulation rate (in 7 out of 7 studies that included this measure); a decrease in speech rate (7 out of 7 studies); an increase of the onset latency of responses to a prompt (in 9 out of 11 studies); an increase in the number of silent pauses (4 out of 5 studies); an increase in the duration of silent pauses (in 8 out of the 10 studies); an increase in the number of filled pauses (4 out of 6 studies); an increase in the duration of filled pauses (1 out of 2 studies); and an increase in the number of repetitions (5 out of 6 studies). Berthold and Jameson (1999) note that these tallies are based on the number of studies that found a given tendency, which was in most but not all cases statistically significant.

Following up on this work, Jameson et al. (2009) present the report of an eight-year project to evaluate how a system can automatically recognise situationally determined resource limitations of its user, and in particular time pressure and cognitive load. They present four experiments investigating the use of speech features to this end. The experimental conditions differed on whether the user (a) was required to produce utterances quickly or not; and (b) was 'navigating' in computer simulation of an airport terminal, or not. In one of the experiments, additional distraction was added by simulated loudspeaker announcements. Thirty-two participants took part in each of the experiments and 7 measures were taken on their speech samples: number of syllables in an utterance, articulation rate, the total duration of silent pauses

1 Berthold and Jameson (1999) present a meta-analysis; the references to the original studies can be found in Berthold (1998)

in an utterance divided by the number of words (using threshold of 200 ms for silent pauses), the total duration of filled pauses in an utterance divided by the number of words, the number of pauses shorter than 200 ms divided by the number of words in an utterance (the authors call this measure "hesitations"), onset latency (the time interval between the presentation of a pictorial stimulus and the start of the first syllable of the response), and the number of disfluencies. Disfluencies included self-repairs (of both syntax and content), lexical false starts and syntactical interruptions; Jameson et al. (2009) argue that since these phenomena are relatively infrequent, it is more useful for a prediction model to group them (calculate the total count of all these types of disfluencies per utterance, with the exclusion of filled pauses that were counted separately).

Jameson et al. (2009) trained dynamic Bayesian networks on the data collected in their experiments, to evaluate the extent to which specific speech features can be used to estimate the cognitive load of the speaker. They evaluate the impact of leaving out each of the 7 measures on classification accuracy. The most discriminative measure was the number of syllables per utterance. Regarding the other indicators, they report that "the sum of the changes that result from leaving individual indicators out is much smaller than the extent to which recognition exceeds the chance level of 50%. This fact shows that the contributions of the indicators are not simply additive; it may be possible to leave out one indicator without much loss of accuracy because the information that it contributes is largely supplied by other indicators" (p. 28).

Yin, Ruiz, Chen, and Khawaja (2007) compiled a spoken corpus of answers to reading comprehension questions related to 3 texts of increasing difficulty. Additionally, for the third level of text difficulty, subjects had to perform a secondary task (counting random numbers they heard over earphones). Text difficulty was assessed using the Lexile framework. In total, 15 speakers participated in their study. The design of Yin et al. (2007) was the inspiration for Study 2 (Reading Comprehension and Question Answering Corpus), therefore it will be presented in more detail in the relevant chapter. Yin et al. (2007) extracted features typically used in speech recognition (Mel-Frequency Cepstral Coefficients, Delta Cepstral Coefficients, Delta-Delta Cepstral Coefficients, Shifted Delta Cepstral Coefficients, and pitch estimate using autocorrelation) and used them to train Gaussian Mixture Models to classify speech data as belonging to one of 3 levels of cognitive load; they report 71.1% classification accuracy for a speaker-independent system.

Yap, Ambikairajah, Choi, and Chen (2009) used the Stroop test to construct a corpus with 3 levels of load (reading congruent colour names; reading incongruent colour names; and reading incongruent colour names with time pressure). This corpus evolved into the Cognitive Load Speech and EGG (electroglottograph) database (CLSE), which also includes 80s seconds of neutral reading per speaker (baseline speech), and a reading span task. In Yap, Ambikairajah, et al. (2009) they propose to add phase-based features to an ex-

isting automatic cognitive load measurement system (based on MFCC and prosodic features, using a Gaussian Mixture Model and described in Yin et al. (2007); see above) in order to improve performance. The additional features proposed are group delay features, all-pole model based FM features and zero crossing count based FM features. Decrease in performance is observed when phase based features are considered individually or when concatenated with baseline features. However, significant performance improvement is observed when group delay features are fused with baseline features using linear combination score level fusion.

In Yap, Epps, Choi, and Ambikairajah (2010), again using the CSLE database, they investigate the effects of cognitive load on glottal features. They report that "recent research on automatic speech-based measurement system indicates that cognitive load information is more prominent in the frequency region below 1 kHz" (p. 5234). The glottal source measures studied were: open quotient, normalized amplitude quotient and speed quotient. Analysis of the glottal parameter distributions suggests that an increase in cognitive load can be related to a more creaky voice quality. In automatic classification using three classes of cognitive load (low, medium, high), the inclusion of the glottal source features improved performance, compared to the baseline system (using MFCCs, pitch, intensity and shifted delta cepstra) from 79% to 84%.

In Yap, Epps, Ambikairajah, and Choi (2011), a follow-up work to the previous two studies, they investigate the effects of cognitive load on voice source features, using the CSLE database. Given that the reliability of glottal flow features depends on the accuracy of the glottal flow estimation, which is a non-trivial process, they propose the use of acoustic voice source features extracted directly from the speech spectrum (or cepstrum) for cognitive load classification. They also present pre-processing and post-processing techniques to improve the estimation of the cepstral peak prominence (CPP). With regards to automatic classification, tests with 3 classes show that CPP outperforms glottal flow features. They report that "score-level fusion of the CPP-based classification system with a formant frequency-based system yielded a final improved accuracy of 62.7%, suggesting that CPP contains useful voice source information that complements the information captured by vocal tract features". We note that the baseline systems used between Yap, Epps, Choi, and Ambikairajah (2010) and Yap, Epps, Ambikairajah, and Choi (2011) are not the same. The envisaged applications though are different: a system such as the one described in Yap, Epps, Choi, and Ambikairajah (2010) would require reliable glottal source features (captured with an EGG device, as was the case in the CSLE database), while Yap, Epps, Ambikairajah, and Choi (2011) describes a system that works only on the basis of the speech signal (the features used can be estimated by analysing the voiced part of the signal).

Gorovoy, Tung, and Poupart (2010) present an experimental study where 10 undergraduate kinesiology students performed 6 tasks: counting up and

down by 1s, 3s, and 7s from random start numbers. These tasks were meant to induce increasing levels of cognitive load, assuming that participants will find counting by 1s easy (low cognitive load), 3s moderately difficult (intermediate difficulty), and 7s difficult (high cognitive load). Each trial (counting) was repeated 3 times and lasted approximately 20 seconds each, giving 18 short speech samples per participant. The speech features found to be affected by these tasks were: articulation rate, pause rate, and pause duration.

Finally, Huttunen, Keränen, Väyrynen, Pääkkönen, and Leino (2011) explored how three types of intensive cognitive load typical of military aviation (load on situation awareness, information processing, decision-making) affect speech. The utterances of 13 male military pilots were recorded during simulated combat flights. The features studied were: articulation rate (number of syllables per second), the first formant (F1) and the second formant (F2) from first-syllable short vowels in pre-defined phoneme environments. Articulation rate was found to correlate negatively (with low coefficients) with loads on situation awareness and decision-making but not with changes in F1 or F2. Changes were seen in the spectrum of the vowels: mean F1 of front vowels usually increased and their mean F2 decreased as a function of cognitive load, and both F1 and F2 of back vowels increased. The strongest associations were seen between the three types of cognitive load and F1 and F2 changes in back vowels. The authors interpret the results in the context of aviation safety, noting that temporal and spectral changes may affect intelligibility; it is thus important to use standard aviation phraseology to minimise the probability of a misunderstanding. While the authors did not address the question directly, an interesting direction for applied research would be to analyse standard aviation terminology, examining the extent to which it is robust to the changes in speech production under high cognitive load.

9.2 STUDIES FOCUSING ON PERCEPTION

In this section we will review a series of studies on the effects of cognitive load on speech perception. There is abundant literature on the related fields of perception and acoustics, covering speech recognition under adverse environments, and including questions ranging from word recognition in noisy and acoustically degraded environments (e.g. the effects of reverberation in halls, signal distortion in mobile telephones etc.) to the psychoacoustics of hearing loss. In this chapter we will focus on studies that have tried to differentiate between the effects of various types of perceptual and cognitive load on speech perception and comprehension.

The distinction between energetic and informational masking is a useful one for distinguishing between different types of load. Energetic masking occurs when the audibility of a target is reduced by a distractor due to blending of their acoustic signals at the periphery, and in the same ear (Mattys, Brooks, & Cooke, 2009, p. 205), whereas informational masking, is often con-

ceptualised as anything that "reduces intelligibility once energetic masking has been accounted for" (Cooke, Garcia Lecumberri, & Barker, 2008, p. 415). The same masker signal may be causing both energetic and informational masking. Informational masking may be caused by competition for attentional resources, and may be more severe when the signal is intelligible (and therefore activates the language comprehension processes). As an example, "in sentence transcription experiments, babble noise made of a large number of talkers (e.g., $N \geq 6$, which is fairly unintelligible) is shown to cause less informational masking than babble noise with few talkers, i.e., $N \leq 3$" (Carhart, Johnson, and Goodman, 1975; Freyman, Balakrishnan, and Helfer, 2004; cited in Mattys, Brooks, and Cooke, 2009, p. 207).

Mattys, Brooks, and Cooke (2009) argue for a psycholinguistic approach to speech recognition in adverse conditions that "draws upon the distinction between energetic masking, i.e., listening environments leading to signal degradation, and informational masking, i.e., listening environments leading to depletion of higher-order, domain-general processing resources, independent of signal degradation" (p. 203). In other words, they propose that the different stages of speech perception and comprehension will be differentially vulnerable to different types of adverse conditions (perceptual, attentional, or mnemonic). Under no particular load, listeners perform connected speech segmentation "by relying primarily on lexical-semantic knowledge and paying less attention to sub-lexical cues" (Gow and Gordon, 1995; Mattys, White, and Melhorn, 2005; cited in Mattys, Brooks, and Cooke, 2009, p. 204). Mattys, Brooks, and Cooke (2009) used short phrases, where the acoustic and lexical cues to segmentation were either congruent (i.e. the acoustic cues coincided with the most probable lexical segmentation), incongruent (i.e. acoustic cues drove the majority of listeners to opt for a lexically improbable segmentation) or ambiguous. In a series of experiments they tested different types of perceptual load (one-talker babble noise; eight-talker babble noise; speech-modulated noise) and cognitive load (playing the stimulus on one ear and the masker on the other; asking participants to perform a secondary task). They show that "severe energetic masking, such as that produced by background speech or noise, curtails reliance on lexical-semantic knowledge and increases relative reliance on salient acoustic detail. In contrast, informational masking, induced by a resource-depleting competing task (divided attention or a memory load), results in the opposite pattern" (p. 203).

In a follow-up to this work, Mattys, Davis, Bradlow, and Scott (2012) present a review of the effects of adverse conditions on the perceptual, linguistic, cognitive, and neurophysiological mechanisms underlying speech recognition. They propose a classification of adverse conditions across two dimensions: their origin and their effect (see Figure 9.1). An adverse condition may be caused by degradation at the source (production of a non-canonical signal), degradation during signal transmission (interfering signal or medium-induced impoverishment of the target signal), and receiver limitations (peri-

pheral, linguistic, cognitive). An adverse condition may affect perceptual processes, mental representations, attention, and/or memory functions. The shade of each cell in Figure 9 indicates their estimation of approximate frequency of occurrence and the importance of each origin-effect combination (light grey: rare/mild; dark grey: common/moderate; black: frequent/severe).

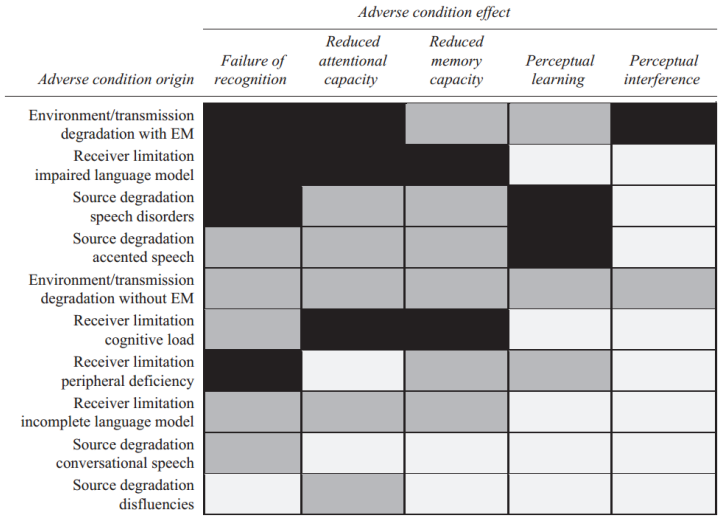


FIGURE 9.1: Summary of effects of adverse conditions to speech perception (From Mattys, Davis, Bradlow, & Scott, 2012)

Regarding the interplay between speech perception and cognitive functions, Mattys, Davis, et al. (2012, p. 967) note that within cognitive hearing sciences, the Ease of Language Understanding model (Ronnberg, Rudner, Foo, & Lunner, 2008) suggests that, "in cases of segmental-lexical mismatches due to a degraded input, working memory is a key predictor of intelligibility, owing to its role in retrospectively and prospectively reconstructing missing information. Thus, in that conceptualisation, it is possible to envisage individual differences in perception of degraded speech as a manifestation of individual differences in memory functions. Likewise, research on compensation (or failure to compensate) for low-level hearing deficits has often limited its scope to the contribution of other sources of linguistic information, such as lexical or syntactic knowledge". Furthermore, "adverse conditions can provide an insight into the degree to which various processes involved in speech recognition are subject to active attentional control or, instead, automatic. For instance, speech tasks requiring active inhibition (e.g., ignoring voice characteristics) or selection (e.g., choosing 'bat' rather than 'pat' when hearing '?at') are shown to be particularly sensitive to divided attention and working memory load (e.g. Nusbaum and Schwab (1986))" (p. 968).

Zekveld and Kramer (2014) investigated changes in speech recognition and cognitive processing load due to masking, and manipulated the similarity between the target and masker speech, by using masker voices with either the same (female) gender as the target speech or different gender (male) and/or by spatially separating the target and masker speech. They used pupillometry as a technique to estimate cognitive load. They report that "the performance benefit from different-gender compared to same-gender maskers was larger for co-located masker signals. The performance benefit of spatially-separated maskers was larger for same-gender maskers. The pupil response was larger for same-gender than for different-gender maskers, but was not reduced by spatial separation". They also found associations between better perception performance and better working memory, better information updating, and better executive abilities; the pupil response was not associated with cognitive abilities.

An unexplored research question is that of the effects of cognitive load (working memory load, attentional load) on the perception of prosodic structure, i.e. on the perception of unit boundaries, beyond lexical segmentation.

10

GLOBAL HYPOTHESES

In the previous chapters, we have formulated hypotheses about the effects of cognitive load on speech production and perception, informed by the preceding theoretical discussion or previous studies. In this chapter, we summarise our hypotheses. Additional, more specific hypotheses are formulated for each of the four studies, depending on the more specific aspects and methods of the studies.

1. With respect to global prosodic characteristics of speech produced under cognitive load, we expect to find a different frequency distribution of silent pause length; silent pauses will be more numerous and longer in duration.

2. Filled pauses will be more numerous, but the frequency distribution of their length will not significantly vary.

3. We expect to find greater variability in all temporal measures, including articulation rate, and a greater variability in segment durations.

4. We expect to find an increase in the pitch range of speakers under high levels of cognitive load.

5. Depending on the individual speaker we expect to find different compensation strategies, which will strike a balance between reducing the length of utterances and decreasing speech rate.

6. Each speaker will have an individual profile of production of self-repairs, and more generally an individual disfluency profile. We expect that higher levels of cognitive load will increase the production of disfluencies, but out hypothesis is that this should be studied against the baseline of each individual speaker, given that each speaker will favour a different self-repair strategy.

7. Under cognitive load, there will be a marked increase in mismatches between prosodic and syntactic boundaries. In quantitative terms, we expect to find an increase in the number of occurrences of major prosodic boundaries inside minor syntactic units (chunks), i.e. in positions

where there are normally not expected. Silent and filled pauses will be placed incongruently with the syntactic structure and we also expect to find a decrease in discourse coherence measures.

8. We expect that speech produced under cognitive load will have less syntactically complex structures. We do not expect to find significant differences in the frequency distribution of morphosyntactic categories, but we do expect a tendency to use more formulaic language.

9. With respect to perception, we do not expect to detect effect at the level of sentence processing, but rather interference with higher-order discourse processing. The speech comprehension and dialogue interaction issues will therefore be addressed in the relevant studies.

Part II

METHODOLOGY, TOOLS AND BASELINE STUDIES

11

ANALYSING SPEECH

This chapter is an introduction to the second part of the thesis. We first describe the methodological choices leading us to design and conduct the four experimental studies presented in Part III. Then we outline aspects of spoken language (prosody and its interfaces with other linguistic levels) that we sought to analyse. In several cases, we decided to do preparatory work and improve the automatic annotation and analysis tools available for (French) spoken corpora: this work is described in detail in the remaining chapters of Part II. Finally, we describe pre-existing spoken language corpora that we used in order to develop and test these automatic tools, and in order to test our methods and obtain baseline results in speech produced under a wide variety of communicative situations. These corpora were constructed for different research objectives. As discussed in section 8.2, it would not be appropriate to hypothesise about the cognitive load experienced by the speaker solely on the basis of the communicative situation, in the absence of subjective and objective measures (cf. Chapter 7) and disregarding individual differences. However they provide a baseline, for various speaking styles, against which we can compare the prosodic feature data collected in our experimental studies.

11.1 CHOICE OF EXPERIMENTAL STUDIES ON SPEECH UNDER COGNITIVE LOAD

We conducted four experimental studies in order to collect speech produced under cognitive load in a variety of communicative situations and tasks. Table 11.1 is a feature grid, showing for each study:

- the type of cognitive load primarily induced on the participants: taxing working memory, attention and executive control resources, or both;
- the task used to induce cognitive load;
- whether the participants engage in speech production, speech perception or both;

- whether there is interaction between participants (dialogue) or not (monologue);
- the type of speech material collected.

The grid additionally shows which aspects of prosody (and its interfaces with other linguistic levels) have been studied on the basis of the data and speech corpora collected in each study.

Short title	Study 1 CLSE^2-FR	Study 2 Monologues	Study 3 Interpreting	Study 4 Driving
Dimension				
CL type (primarily)	Working Memory	Working Memory	WM + Attention	WM + Attention
Task inducing CL	Stroop, Reading Span	Memorisation	SI itself	ConTRe task
Production	✓	✓	✓	✓
Perception			✓	✓
Interaction	Monologue	Monologue	Monologue	Monologue, Dialogue
Speech material	Words, Phrases	Read, Spontaneous	Read, Spontaneous	Read, Spontaneous
Analysis				
Glottal source features	✓			
Pauses		✓	✓	✓
Temporal measures		✓	✓	✓
Disfluencies			✓	✓
Boundaries			✓	✓
Dialogue-related				✓
Perception of fluency			✓	

TABLE 11.1: Feature grid of the four experimental studies

Note: Simultaneous interpreting is in itself a cognitively demanding task, that involves both speech perception and production.

11.2 MEASURES AND AUTOMATIC TOOLS FOR THE ANALYSIS OF SPOKEN CORPORA

During the course of writing this thesis, we have developed new automatic annotation tools, as well as a common platform to host them. Much of this work was undertaken in collaboration with colleagues, and working on a variety of spoken corpora. Oftentimes the research questions tackled were broader than the focus of this thesis: the tools developed can be used in various domains of speech processing and prosody. In this section we outline this work, its motivation and links with the annotation and analysis of speech under CL collected through the experimental studies. This work is presented in the remaining chapters of Part II.

Previous studies (mainly on English) report effects of cognitive load on the temporal organisation of speech (cf. section 9.1). Global hypotheses 1, 2 and 3 are related to the temporal organisation of speech (see Chapter 10). Consequently, Chapter 16 is devoted to the methodological aspects of describing the temporal structure of speech, and more specifically, the statistical distribution of silent pause length, and other temporal measures (speech rate, articulation rate, run length etc.). We adopt a method described in the literature that models the log-transformed length of silent pauses as a mixture of Gaussian distributions (GMM). This technique is used in Studies 2, 3 and 4. We have also developed an analysis tool that automatically calculates approximately 50 temporal prosodic parameters in multi-speaker dialogues: this tool is used in Study 4.

We have formulated the hypothesis that under cognitive load, there will be a marked increase in mismatches between prosodic and syntactic boundaries, and that we expect to find an increase in the number of occurrences of major prosodic boundaries inside minor syntactic units (chunks). For this reason, it will be beneficial to automate, to the extent possible, the annotation of prosodic boundaries and of chunks in French spoken corpora. This objective is covered in four interrelated chapters (12, 13, 14 and 15).

Chapter 12 describes the challenges of performing a morphosyntactic analysis of spoken language, and Chapter 13, focuses on the issue of annotating disfluencies. We have developed DisMo, an automatic part-of-speech, multiword unit and disfluency annotator to automate this process. Additionally, we present a detailed annotation protocol that was validated by applying it on a 7-hour corpus, and propose a system for the automatic detection of disfluencies on the basis of a transcription and the corresponding speech signal. These methods and tools have been used in Studies 3 and 4.

In Chapter 14 we present our work on the detection of prosodically prominent syllables based on their acoustic correlates and lexical/syntactical information, and in Chapter 15 we present our studies on the acoustic and syntactic correlates of perceived prosodic boundaries. Prosodic prominence and boundaries are indispensable in understanding how speakers segment their

flow of speech and how listeners perceive this segmentation. The combined use of the DisMo and Promise tools facilitates our study of the relationship between prosodic and syntactic boundaries in the speech material collected.

The speech elicited in the context of the experimental studies amounts to several hours of recordings; corpora used for tool development and baseline studies are also large and contain several layers of annotation. It became apparent that a streamlined approach to the management of such data is necessary. In Chapter 17 we present Praaline, a new software tool that we developed, that allows its user to perform corpus management, manual and automatic annotation, as well as visualisation, querying and statistical analysis of speech data.

Finally, in the remaining chapters of Part II, we also present four studies on the following subjects: a corpus-based study on the frequency, co-occurrence and prosodic characteristics of disfluencies in spoken French; a corpus-based study on the acoustic and syntactical correlates of prosodic boundaries in French as these boundaries had been annotated by experts; followed up by a perceptual experiment to examine the naïve listeners' on-line perception of prosodic boundaries; and a corpus-based study on the effects of speaking style on the distribution of silent pauses.

11.3 SPEECH CORPORA REFERENCED IN THIS THESIS

In this section we present the corpora on which we have worked, in order to perform baseline analyses, and/or develop and test automatic annotation tools, and to which we have contributed new annotations. In the interest of brevity, other chapters in Part II will reference corpora by their name, without repeating the description of their compositions and design principles. The present section serves as a reference for next chapters. The corpora presented here include CPROM-PFC (Avanzi, 2014); LOCAS-F (L. J. Martin, Degand, & Simon, 2014); C-PhonoGenre (Goldman, Prsir, et al., 2014); Rhapsodie (Lacheret, Kahane, et al., 2014); these corpora have been used in studies and in the development of automatic annotation tools. We also briefly present large corpora to which we have contributed annotations, including the PFC corpus (Durand et al., 2009); OFROM (Avanzi, Béguelin M.-J., & Diémoz, 2015); and the Valibel collection of corpora (Simon, Francard, & Hambye, 2014).

11.3.1 *The CPROM-PFC corpus*

The CPROM-PFC corpus (Avanzi, 2014) consists of speech material recorded in 14 geographical areas, spread over 3 European French-speaking countries: 5 regional varieties spoken in Metropolitan France (Béthune, Brécey, Lyon, Paris and Ogéviller); 5 regional varieties spoken in Switzerland (Fribourg, Geneva, Martigny, Neuchâtel and Nyon) and 4 regional varieties spoken in

Belgium (Brussels, Gembloux, Liège and Tournai). The material was mainly extracted from the database of the project "Phonologie du français contemporain" (see below) and additional recordings performed by M. Avanzi. For each of the 14 sites, 4 female and 4 male speakers were selected; they were born and raised in the city in which they were recorded. The age of the speakers varies between 20 and 80 years. The corpus is stratified into four age groups: this parameter is controlled for each of the 14 groups of speakers ($F(13, 84) = 0.308$), between male and female speakers ($F(1, 84) = 0.110$, n.s.) and between male and female speakers across the 14 groups ($F(13, 84) = 0.114$, n.s.).

Each speaker was recorded in a reading text task (the text is 398 words-long) and in semi-directed sociolinguistic interviews, in which the informant has minimal interaction with the interviewer. The entire reading text and a stretch of 3 minutes of spontaneous speech for each speaker were orthographically transcribed and automatically aligned within Praat (Boersma & Weenink, 2016) with the EasyAlign script (Goldman, 2011), which provides a 3-layer annotation in phones, syllables and words. All alignments were manually verified and corrected when necessary by inspecting both spectrogram and waveforms. Several additional levels of annotation were added using the tools presented in this thesis. The corpus was imported into Praaline (Christodoulides, 2014, ; see Chapter 6) allowing for precise control of the annotation structure: levels of annotation and their associated attributes are defined as a database schema, protecting the referential integrity of the data (for example, the fact that an annotation attribute is linked to the level of syllables or to the level of tokens is no longer dependent on keeping tier boundaries precisely aligned in Praat TextGrids, but becomes part of the database schema definition). Part-of-speech and disfluency annotation was added to CPROM-PFC using the DisMo annotator (see Chapters 2 and 3); this corpus served as the training and development corpus for the first public version of DisMo. Trained experts annotated the corpus for syllabic prosodic prominence, which in turn was used to develop the Promise automatic annotator (see Chapter 4). Dedicated annotation attributes indicates overlapping or short non-audible segments that are unusable for prosodic measure extraction. In total, the corpus is approximately 11 hours-long, and includes approximately 113 thousand tokens (63k in semi-directed interviews and 47k in reading). Table 11.2 presents the basic properties of this corpus. CPROM-PFC is not yet publicly available, and access to it was granted by means of collaboration with Dr Mathieu Avanzi.

11.3.2 *The LOCAS-F corpus*

The LOCAS-F corpus (L. J. Martin et al., 2014) is a corpus of spoken French with samples of 14 different speaking styles (discourse genres). Each speaking style is represented with approximately the same amount of speech ma-

Area	Region	Age		Dur. (sec)	Nb. syll.	Nb. tokens
		Min-Max	Mean (sd)			
Belgium	Brussels	27-65	44 (15)	2810	11446	8565
	Gembloux	22-76	42 (21)	2821	11677	9135
	Liège	21-76	48 (24)	2951	9692	7400
	Tournai	19-82	44 (26)	2837	10518	8031
France	Béthune	21-89	46 (25)	2925	11153	8571
	Brécey	19-80	47 (22)	3110	11505	8659
	Lyon	21-74	42 (21)	2677	10866	7783
	Paris	24-86	50 (22)	2896	10088	8188
	Ogéviller	23-93	58 (24)	3023	10685	8101
Switzerland	Fribourg	20-82	43 (24)	2895	10865	8186
	Geneva	21-61	41 (18)	2863	10720	8062
	Martigny	22-80	49 (28)	2963	10199	7726
	Neuchâtel	25-78	53 (24)	2960	10201	7625
	Nyon	30-70	46 (17)	2929	9948	7637
Totals	14 points	19-93	46 (21)	11.2 h	149563	113669

TABLE 11.2: Composition of the CPROM-PFC corpus

terial. Its duration is 3.5 hours and it contains 43.000 tokens. The discourse genres covered are the following: official speech in an academic setting [aca], scientific conference presentation [conf], face-to-face dialogue in a formal [conv-f] or informal [conv-i] setting, political debate [deb], sermon / homily [hom], radio interview [int-rad], sociolinguistic interview [int-soc], informal TV interview [int-tv], monologue narration of life event [mono-n], semi-prepared radio monologue [mono-r], radio news bulletin [news], political public address [pol], and neutral reading of a newspaper text [read]. The corpus composition is detailed in Table 11.3, including the number of total duration of samples, their number of tokens, the number of major and intermediate prosodic boundaries, and the number of functional sequences coded.

The LOCAS-F corpus contains parts of the C-PROM corpus (Avanzi, Lacheret, & Victorri, 2010), which is a publicly available corpus of 24 recordings (70 minutes in total) covering 7 speaking styles, produced by speakers from Belgium, Switzerland and France. More specifically, the overlapping samples between C-PROM and LOCAS-F are: 3 radio news broadcasts (code JPA in C-PROM), 3 political public addresses (code POL in C-PROM), 3 scientific conference presentations (code CNF in C-PROM), 2 radio interviews (code INT in C-PROM), and 3 monologue narrations of life events (code NAR in C-

Genre	Nb	Duration	Tokens	PB //	PB ///	Synt. Seq
ACA	3	15:16	2332	161	401	361
CONF	4	16:43	2939	318	321	556
CONV-F	3	12:51	2714	354	269	883
CONV-I	3	12:24	2945	280	377	1251
DEB	4	19:17	5216	883	529	1463
HOM	3	13:21	1759	120	344	428
INT-RAD	4	20:28	4313	746	476	1032
INT-SOC	3	15:23	2958	453	335	766
INT-TV	3	15:31	4003	517	482	1333
MONO-N	3	10:20	2367	288	135	862
MONO-R	3	13:22	2591	293	278	564
NEWS	4	14:44	2902	463	207	564
POL	5	20:23	2889	307	475	610
READ	3	15:17	3151	445	414	485
Total	48	3:35:20	43079	5628	5043	11158

TABLE 11.3: Composition of the LOCAS-F corpus

PROM). The C-PROM corpus was one of the first publicly available corpora to contain an annotation of prosodic prominent syllables, cross validated by 3 expert annotators; we have extracted the prosodic prominence annotation from the original data and injected it in the LOCAS-F corpus, which does not contain a manually corrected annotation of prominent syllables.

The LOCAS-F corpus was primarily constituted to study the relationship between prosodic and syntactic boundaries, including the properties of dislocated structures. It has been orthographically transcribed in Praat, and a phonetic transcription was automatically produced and aligned with the speech signal using the EasyAlign script. The aligned segmentation in phones, syllables and words was manually corrected by two experts. The corpus contains a detailed annotation of prosodic boundaries and syntactic structures, performed by trained phoneticians using a double-blind methodology. The prosodic boundary annotation distinguishes between major and intermediate boundaries; Mertens and Simon (2013) explain the rationale behind the choice of these two levels. The syntactic annotation is articulated in two levels: functional sequences and dependency clauses. The encompassing level delimits dependency clauses, the maximal syntactic unit that is any root element (most often a verb) and its dependent elements. Dependency clauses are analysed into functional sequences: Verb Sequence (SV), Subjet Sequence (SS), Dependent Sequence (SR) on the left or right periphery of the root element,

Incomplete Sequences (I), Adjuncts (A), Other (mostly non-verbal sequences), and Discourse Markers (DM). This syntactical annotation has been inspired by Bilger and Estelle Campione (2002) and Blanche-Benveniste (1996), and is described in detail in Tanguy, van Damme, Degand, and Simon (2012). Degand and Simon (2009) present a study on the effects of speaking styles on the relationship between prosodic and syntactic boundaries, based on an earlier version of the LOCAS-F corpus.

Our contribution to the LOCAS-F corpus was manifold. First, the corpus was imported into Praaline, which helped discover data integrity errors (misaligned segments, invalid tags left behind by human annotators etc.). It also allows a better representation of multi-speaker recordings: while originally transcribed in separate Praat TextGrids, the speech and corresponding annotations pertaining to different speakers is locked to the same timeline in Praaline. This allows a clearer visualisation of the annotations and a better analysis of those corpus samples where more than one speakers interact. We subsequently applied the DisMo automatic annotator to the entire corpus, and we manually corrected the annotation of disfluencies by applying the protocol described in Chapter 3. We furthermore used Prosogram extract a multitude of prosodic measures for each syllable, and store this information in the corpus database. These measures were correlated with the expert prosodic annotation, in order to study the acoustic and syntactic correlates of the boundaries perceived by experts, as described in Chapter 5. The LOCAS-F corpus is not yet publicly available, and access was granted by means of collaboration with Prof Anne Catherine Simon and Prof Liesbeth Degand (Université catholique de Louvain).

11.3.3 *The C-Phonogenre corpus*

The C-PhonoGenre corpus was compiled to study situation-dependent speaking styles, or phonogenres and the associated prosodic variation. It contains data from 8 speaking styles: instructional speech [DIDA]; spontaneous narration [NARR]; speeches during "Question Time" at the French parliament [PARL]; sermons [RELG]; radio press reviews [RPRW]; three kinds of sports commentary [SPOR]: rugby, basketball and football; presidential New Year's wishes [WISH] and weather forecasts [WFOR]. The average sample duration per speaker is 5:30 min. The corpus composition is presented in Table 11.4.

The corpus samples were selected using the methodology detailed in Goldman, Prsir, et al. (2014). The corpus contains recording of both female and male speakers, originating from 3 different French-speaking areas: Metropolitan France, Belgium and Switzerland. Speaking situations were described by features on four dimensions: audience, media, preparation and interactivity; each dimension had 3 different states. Audience = 1 indicates that the speaker is physically present before an audience. Media = 1 indicates speech directed to an individual or a small group, yet in front of a microphone or camera

PhonoGenre	Nb	Duration (min)	Nb. syll.	Nb. tokens
DIDA	17	100	26 304	18 717
NARR	10	44	11 396	9 546
PARL	10	20	5 710	3 613
READ	16	36	9 932	6 648
RELG	7	54	8 726	6 141
RPRW	15	95	26 359	17 531
SPOR	5	35	7 601	5 305
WFOR	10	9	2 861	1 947
WISH	15	98	18 614	12 578
Total	105	491	117 503	82 026

TABLE 11.4: Composition of the C-PhonoGenre corpus

(indirect audience). Preparation = 1 indicates semi-prepared speech, situated between spontaneous and read speech. In the case of parliamentary debates, a question is prepared, while the answer is semi-prepared. Interactivity = 1 indicates that the main speaker may be interrupted. The values for each dimension and each speaking style in the C-PhonoGenre corpus can be found in Table 11.5.

The C-PhonoGenre corpus was manually transcribed orthographically and a phonetic transcription and segmentation was obtained using EasyAlign. The alignment was manually corrected. A single annotator manually added speech delivery information containing four types of annotation: i) disfluencies, articulation and phonological phenomena: schwa; vowel lengthening (whether associated to hesitation or not); creaky voice; liaison and elision; ii) symbols to distinguish between complete silence, audible and less audible breaths, and mouth noises; iii) indices of paralinguistic phenomena (laugh, cough) and external sounds; iv) overlapping segments and syntactic plan interruptions.

Our contribution to C-PhonoGenre was the following. During its construction, we applied the DisMo annotator to the entire corpus; the disfluency annotations for some of the samples were corrected (cf. Chapter 3). The entire corpus, along with multiple annotations from semi-automatic prosodic analysis tools: Prosogram (Mertens, 2004), ProsoProm (Goldman, Avanzi, Auchlin, & Simon, 2012) and DurationAnalyser (Dellwo, 2010) were stored in a Praaline corpus database. We collaborated with Pierre Menetrey to create a bridge beween the Praaline database and a web-based corpus access system; as a result C-PhonoGenre is publicly available on the web. After developing the automatic annotation tools describe in this thesis, we have applied them to C-PhonoGenre in order to enrich the available data, and used the

Phonogenre		Audience	Media	Preparation	Interaction
DIDA	Radio	1	2	2	2
	TV	0	2	2	0
	Lecture	2	0	1	0
NARR	Narration	1	0	0	2
PARL	Question	2	1	2	1
	Answer	2	1	1	1
READ	Reading	0	0	2	0
RELG	Internet mass	0	1	2	0
	Sermon on TV	2	1	2	0
RPRW	Radio press review	0	2	2	0
SPOR	Basket	0	2	0	0
	Rugby/football	1	2	0	2
WFOR	Weather forecast	0	2	2	0
WISH	Pres. New Year	0	1	2	0

TABLE 11.5: Situational features in C-PhonoGenre

corpus for testing. Access to C-PhonoGenre prior to its public dissemination was granted by means of collaboration with Prof Antoine Auchlin and Jean-Philippe Goldman (Université de Genève).

11.3.4 *The Rhapsodie corpus*

The Rhapsodie corpus is a publicly available corpus containing multiple speaking styles and created with the objective to study the relationship between prosodic phrasing and syntax in French. The corpus samples were mainly collected from existing French corpora, including the PFC corpus (Durand et al., 2009), C-PROM (Avanzi et al., 2010) and CFPP (Branca-Rosoff, Fleury, Lefeuvre, & Pires, 2012). The corpus contains 57 short samples (average sample duration is 5 minutes) for a total of 3 hours and 33.000 tokens. The corpus samples were balanced across four dimensions: the degree of speech planning, the degree of interactivity, the communication channel, and the main discourse strategy used by the primary speaker (oratory, argumentative, descriptive, or procedural); the corpus contains both monologues and dialogues. The syntactic annotation is articulated in two levels, called "micro-syntactic" and "macro-syntactic" by the authors; the main theoretical framework posits the use of "pile structures" to represent the syntactic relations of short segments of continuous speech, including self-corrections and other types of disfluencies. Despite the fact that this syntactic annotation is quite idiosyncratic,

the corpus does provide a coarse part-of-speech annotation and dependency syntax structures in the CoNLL format. The prosodic annotation includes: prosodically prominent syllables annotated by experts based on their perception, using two levels (weak and strong); an annotation of disfluencies at the syllable level. (e.g. lengthening); and an automatically derived prosodic structure annotation composed of three levels: intonational periods, rhythmic groups and metrical feet. A perceptual boundary annotation was abandoned by the project due to poor inter-annotator agreement (Lacheret, Kahane, et al., 2014, p. 4). The Rhapsodie corpus contains expert-validated annotations and its objectives are almost identical with those of the LOCAS-F corpus. Since the corpus is already publicly available (unlike LOCAS-F) we have used it to test the findings of our studies on prosodic boundaries (Chapter 15) and as a testing corpus for the automatic annotation tools DisMo (Chapter 12) and Promise (Chapter 14).

12

MORPHOSYNTACTIC ANALYSIS OF SPOKEN LANGUAGE

The morphosyntactic annotation of spoken corpora poses specific challenges, caused by the specificities of spoken language and by the nature of transcription. For example, when developing a part-of-speech tagger or a syntactic parser for written texts, it is possible to use punctuation in order to segment the input: sentences are already delimited, even in languages where word-level segmentation is not straightforward. This is not the case for transcriptions of spoken language; given a non-punctuated transcription, different segmentations are possible and represent alternative interpretations of the input. Furthermore, spoken language is characterised by the presence of disfluencies, a class of phenomena that reflect the real-time construction of the message by the speaker (and often serve other purposes; we return to the subject of disfluencies in the next chapter). These phenomena, along with "non-canonical" syntactic structures (e.g. frequent use of interrupted structures, parentheticals and clefts), pose additional problems for part-of-speech tagging and syntactical analysis.

We take the position that, because of all these challenges, there is a need to develop automatic annotators specifically for spoken language. However, it is often beneficial to be able to compare the morphosyntactic annotation of a spoken corpus with the morphosyntactic annotation of a written corpus, in order to discover commonalities and differences (e.g. phenomena specific to spoken language, or having a different frequency of occurrence depending on the modality). We therefore posit that a useful tagger for spoken language should be constructed on the basis of a "least common denominator" of features (e.g. a basic annotation tag-set), shared between spoken and written corpora, and enriched with additional annotation features to fully capture spoken language phenomena.

In this chapter we describe the structure and features of the DisMo (Christodoulides, Avanzi, & Goldman, 2014) multi-level annotator for spoken language that we developed based on this design philosophy. DisMo is designed to be used with transcriptions of spoken corpora, and provides multiple levels of annotation: part-of-speech tagging, lemmatisation, multi-word unit detection, automatic detection and annotation of several types of disfluencies, detection of potential discourse markers, and chunking. In this chapter we fo-

cus on the version for French (e.g. in giving details about tag-set choices) as this is the version used in this thesis. A version for English is also available, and it is possible to easily train new language models given a gold-standard annotated corpus.

12.1 RELATED PREVIOUS WORK

Previous work on the morphosyntactic annotation of spoken French has focused on three areas: part-of-speech tagging, chunking and dependency parsing. It has been based on two main techniques: pre-processing spoken language transcriptions in order to use an automatic annotator designed for written language; or training statistical models on a (frequently small) spoken language corpus.

Valli and Véronis (1999) note that spoken language transcriptions are challenging to automatic part-of-speech taggers because of the particular conventions necessary to represent phenomena that are only present in speech; they particularly mention hesitations, lexical false starts ("bribes") and repetitions. However they report that the performance of a typical POS tagger is acceptable, provided that the transcription is normalised, i.e. provided that these phenomena are removed from the input to the tagger. They therefore encourage the construction of large-scale POS-annotated corpora of spoken French. Dister (2007) explored the application of context-free grammars, expressed as a cascade of Unitex graphs, to normalise spoken language transcriptions, followed by rule-based part-of-speech tagging. Dister (2007) applied these techniques on a subset of the Valibel collection of spoken language corpora (Simon, Francard, & Hambye, 2014). Building up on this work, Blanc, Constant, Dister, and Watrin (2008) propose a chunker for spoken French, which runs in two phases: initially false starts and repetitions are identified: repetition detection is rule-based, and false starts are already annotated as part of the Valibel transcription conventions; then a cascade of finite-state automata, expressed as Unitex graphs, is applied to the normalised transcription, in order to produce a segmentation in "super-chunks" (a term coined by the authors for segments larger than a typical chunk, that may include multi-word expressions and extend to grammatical constituents). Language resources are used in these approaches, often enriched with POS lexica extracted from corpora of transcriptions.

An alternative approach to normalising the input to a tagger designed for written texts is training statistical models directly on an annotated spoken corpus. Eshkol, Tellier, Taalab, and Billot (2010) explore this approach in depth, working on a subset of the ESLO ("Enquête sociolinguistique d'Orléans") corpus. Based on a manually corrected corpus of 18k tokens, using a part-of-speech tag-set that was enhanced with additional tags for spoken language, and using Conditional Random Fields (CRF) models (Lafferty, McCallum, & Pereira, 2001), they report an accuracy of 85% to 90%. Tellier, Duchier,

Eshkol, Courmet, and Martinet (2012) explore the development of an automatic chunker for French, based on CRF statistical models trained on the French Treebank (Abeillé, 2003). The French Treebank is a gold standard treebank of French newspaper texts, containing part-of-speech tags and a syntactical analysis in constituents. In a follow-up study, Tellier, Eshkol, Dupont, and Wang (2014) explored the use of the uncorrected output of a part-of-speech tagger to chunk spoken language transcriptions. The reported accuracy varies greatly with the type of the chunk (ranging from an F-Measure of 69% for adjectival chunks to 85% for verbal chunks).

Benzitoun, Fort, and Sagot (2012) present the TCOF-POS corpus, a freely available spoken French corpus, with manually corrected part-of-speech annotation. After training two widely used part-of-speech taggers on this corpus, they obtain an accuracy of 93,6% (TreeTagger) and 94,3% (Melt). They also propose a tag-set for spoken French part-of-speech annotation (see below: comparison of tag-sets).

Several authors have explored the use of linguistic resources to improve morphosyntactic annotation, chunking and parsing. The use of external dictionaries to enhance part-of-speech tagging for French is explored in Denis and Sagot (2012): on the basis of experiments with written texts, they report a 25% relative error reduction over the same tagger without lexical information. Constant and Sigogne (2011) show that the performance of a part-of-speech tagging system based on CRF models is improved by including multi-word unit detection in the annotation pipeline. Manning (2011) notes that seeking to improve the performance of part-of-speech taggers already at a very high level ("from 97% to 100%", as the title reads) depends on better descriptive linguistics, i.e. improving the taxonomies used in the training corpora, and incorporating more explicit linguistic rule knowledge in the systems (as opposed to increasing the size of training corpora even further).

Finally, a body of work has sought to apply dependency parsing techniques to spoken language. Deulofeu, Duffort, Gerdes, Kahane, and Pietrandrea (2010) present the syntactic annotation scheme used in the Rhapsodie project, whose main philosophy is summarised by the authors as follows: "we don't consider that syntax can be reduced to dependency, and we have to define the delimitation of functional relations as well as the delimitation of so called 'macro-syntactic' phenomena such as dislocation and colon effect that go beyond dependency. Our complete annotation therefore includes units joined by dependency, paradigmatic sub-units, and higher-level relations that are still syntactic and not purely discursive. We propose a well-defined distinction between syntax based segmentation, called 'dependency units', and pragmatically based segmentation, called 'illocutionary units'" (p. 275). Bawden, Botalla, Gerdes, and Kahane (2014) detail the process by which the 33k token treebank was annotated in dependency relations, paradigmatic structures ("piles", Gerdes and Kahane (2009)), and discourse-level relations.

In designing the DisMo annotator, we have tried to combine techniques previously used in the studies summarised in this section. More specifically, DisMo uses lexical resources, includes a pre-processing stage, has been trained on manually-corrected spoken language transcriptions, integrates multiword unit detection and specifically tackles the annotation of disfluencies. In the next section, we present a method to structure annotations in multiple levels that enhances the system's capacity to describe different types of linguistic phenomena, while keeping the annotations simple and expressing interrelations.

12.2 A MULTI-LEVEL ANNOTATION SCHEME

The DisMo annotator accepts many different types of input. For a complete analysis, the user should provide an orthographic transcription aligned to the speech signal, at least at the utterance level. If such data is available, the system will include automatically calculated prosodic parameters into its analysis (e.g. silent pauses as indices of speech segmentation; speech rate and pitch for disfluency detection). In the case of disfluency detection (examined in more detail in the next chapter), a speech-text alignment at the syllable level will permit finer annotations (e.g. detection of lengthening). It is also possible to perform morphosyntactic annotation on a transcription without the corresponding speech signal, or on a written text, in which case the prosody-related modules are deactivated. DisMo is capable of handling multi-speaker transcriptions, considering speech turns as indices to sequence segmentation. A customisable pre-processing script ensures that information contained in transcription conventions (e.g. noting lexical false starts with a special symbol) will be preserved and used in the subsequent annotation stages. The annotator's output is structured on multiple levels, and each level contains attributes describing a different set of linguistic phenomena, as described in Table 12.1.

Structuring the annotation on multiple levels has several advantages over simple taggers that only output one tag per lexical unit (Schmid, 1994, e.g. TreeTagger;). Most importantly, it allows for a principled approach to the process of tokenisation and identification of multi-word units and expressions. At the level of minimal tokens, the tokenisation will continue to split as long as the resulting lexical units exist as autonomous units in the lexicon of a given language. As an example, for French, the string "tout à fait" will be split into three minimal-level tokens, while "aujourd'hui" and "parce que" will remain as single minimal-level tokens (actually, for French, these are the only two cases of minimal-level tokens containing a splitting character). Using a separate multi-word unit level, DisMo is able to accurately represent the fact that a multiword unit expression may have a different syntactical role than its constituent parts; while at the same time facilitating the study of patterns of part-of-speech constituent tokens in multiword units (e.g. retriev-

Annotation Level	Annotation Attributes	
`tok-min` minimal tokens	pos-min	part-of-speech tag (with 2 or 3 levels of detail; see next section)
	pos-ext-min	extended morphosyntactic information (e.g. gender, person)
	lemma-min	lemma of the minimal token
	disfluency	disfluency annotation coding
	part-of-mwu	a flag indicating whether the minimal token is part of a multi-word unit
	(internal info)	information regarding the method that DisMo used to select the various tags and a confidence measure
`tok-mwu` multi-word units	pos-mwu	part-of-speech tag of the multi-word unit as a whole
	pos-ext-mwu	extended morphosyntactic information
	lemma-mwu	lemma of the multi-word unit
	(internal info)	method and confidence measure, as above
`discourse` potential DMs	Note that a potential discourse marker may span many minimal tokens and may or may not correspond to a multi-word unit.	
`chunk`	Grouping of minimal token into chunks, i.e. non-recursive minimal syntactical units (see below).	

TABLE 12.1: Levels of annotation and related attributes added by the DisMo annotator

ing all multiword units in a corpus that are composed by a verb + a noun). We illustrate these points with examples from the PFC corpus:
1. « je n'ai pas pu obtenir de poste à Lyon tout de suite (pos-mwu ADV) donc j'ai été exilé » (PFC Lyon; 69aag1gg)
2. « Ils vendent tous les objets euh que des objets euh russes ça c'est fabrication russe. Tous ce qui était objet russe oh il y en avait certains c'était euh donc des fournisseurs de Paris » (PFC, Aveyronnais à Paris, 75xlv1lg)

In example (1) the expression "tout de suite" is a multiword adverbial expression and will be annotated with the tag ADV on the entire string at the tok-mwu level; still, the constituent parts of the adverbial expression will receive their individual part-of-speech tags on the minimal token level (in this case: "tout/ADV de/PRP suite/NOM:com"). Similarly, in example (2), the multi-level annotation scheme allows us to annotate "c'est", "il y en avait" and "c'était" as multi-word expressions introducing constituants ("introducteurs"), while keeping the fine-grained part-of-speech annotation intact at the tok-min/pos-min level (e.g. "c'/PRO:per:sjt était/VER:impf", "il/PRO:per:sjt y/PRO:per:obji a/VER:pres").

Multiword units having a different morphosyntactic function when taken as a whole, compared to the morphosyntactic properties of their constituent parts, include numerals, adverbial expressions, composite words etc. As an example, we could envisage a large corpus study on the morphosyntactic

patterns of adverbial expressions in French (by searching for the tok-min-level annotation of constituent tokens belonging to tok-mwus annotated with the tag ADV). Another example (relevant to the work presented in this thesis) is the annotation of prosodic prominence in multi-word adverbial phrases: the fact that these are already marked as multi-word units by DisMo is taken into account in subsequent processing stages (e.g. the detection of prominent syllables and boundaries by Promise, cf. 14).

Beyond multiword units, the multi-level annotation scheme used by DisMo allows for a clean separation between the morphosyntactic function and the pragmatic or discursive function of a token or a string of tokens. For example: 3. « bon (pos-min: ITJ; discourse: MD) déjà j'ai dû passer un un euh un concours un test d'entrée un concours c'est-à-dire qu'on a été plusieurs et il y en a qu' ont pas été pris quoi (pos-min ITJ; discourse: MD) on a été genre une dizaine ils en ont pris genre six tu vois (pos-mwu: VER:pres; discourse: MD) un truc comme ça » (PFC Paris; 75cab1gg)

In this example, the tokens "bon", "quoi", "tu vois" will be annotated as discourse markers at the discourse and chunks level, while keeping their original part-of-speech tags at the tok-min level. As an example, using a corpus annotated with DisMo, it is trivial to perform a statistical study on the morphosyntactic properties of potential discourse markers; a more difficult task, of course, is to decide whether these potential discourse markers are actually discourse structuring devices, on the basis of linguistic principles. In the next section, we will present a comparison the choices made for the tag-set of the French version of DisMo.

12.3 COMPARISON OF PART OF SPEECH TAG-SETS FOR FRENCH

The part of speech tag-set of DisMo is composed of 12 main categories: verbs (VER), nouns (NOM), adjectives (ADJ), adverbs (ADV), conjunctions (CON), determinants (DET), pronouns (PRO), numerals (NUM), prepositions (PRP), interjections (ITJ), prefixes (PFX) and foreign words (FRG). Most of these categories are further broken down, on a second and sometimes third level. For example, we have chosen a broad interpretation of the class of determinants (in line with the reasoning put forward by the Universal POS Tagset Project, Petrov et al., 2011). The class of determinants includes definitive articles (DET:def), demonstratives (DET:dem), indefinite articles (DET:ind), interrogative adjectives (DET:int), exclamatives (DET:exc), partitive articles (DET:par) and possessive articles (DET:pos). In the case of verbs, the second and third levels of the tag are used to indicate verb tense and aspect. A comparison of various tag-sets that have been proposed for French in the literature can be found in Table 12.2.

Part-of-Speech class	Universal POS tagset	TreeTagger	Stanford POS tagger	French Treebank	MElt	Cordial adapté	TreeTagger TCOF-POS	DisMo
Abréviation	*	ABR	_	_	_	_	X:sig*	X:acr*
Adjectif	ADJ	ADJ	A	A	ADJ	ADJ	ADJ	ADJ
Adjectif interrogatif	ADJ	ADJ	A	_	ADJWH	ADJ	PRT:INT*	DET:int
Adverbe	ADV	ADV	ADV	ADV	ADV	ADV	ADV	ADV
Adverbe, acronyme	ADV	ADV	ADV	ADV	ADV	ADV	ADV	ADV:acr
Adverbe interrogatif	ADV	ADV	ADV	ADV	ADVWH	PIINT	ADV	ADV:int
Adverbe de négation	ADV	ADV	ADV	ADV	ADV	ADVNEG	ADV	ADV:neg
Conjonction de coordination	CONJ	KON	C	CC	CC	CONJCOO	KON	CON:coo
Conjonction de subordination	CONJ	KON	C	CS	CS	CONSUB	KON	CON:sub
Déterminant défini	DET	DET:ART	D	D	DET	DET/DEF	DET:def	DET:def
Déterminant démonstratif	DET	?	D	D	DET	DET/DEM	DET:dem	DET:dem
Déterminant indéfini	DET	DET:ART	D	D	DET	DET/IND	DET:ind	DET:ind
Déterminant interrogatif	DET	?	D	D	DETWH	DET/INT	DET:int	DET:int
Déterminant exclamatif	DET	?	D	D	DET	DET/IND	DET:ind	DET:exc
Déterminant partitif	DET	DET	D	D	DET	DET/IND	DET:par	DET:par
Déterminant possessif	DET	DET:POS	D	D	DET	DET/POSS	DET:pos	DET:pos
Mot étranger	X	_	ET	ET	ET	ET	ETR	FRG
Interjection / Marqueur discursif	X	INT	I	I	I	INT	INT	ITJ
Interjection	X	INT	I	I	I	INT	INT	ITJ:itj
Introducteur	_	_	_	_	_	_	_	INTROD
Onomatopée	X	INT	I	I	I	UEUPH	INT	ITJ:ono
Nom acronyme	NOUN	NOM	N	NC	NC	NC	NOM:sig	NOM:acr
Nom commun	NOUN	NOM	N	NC	NC	NC	NOM	NOM:com
Nom propre	NOUN	NAM	N	NP	NPP	NP	NAM	NOM:pro
Numéral cardinal, déterminant	DET	NUM	D	D	DET	DET	NUM	NUM:crd:det
Numéral cardinal, adjectif	ADJ	NUM	D	D	DET	DET	NUM	NUM:crd:adj
Numéral cardinal, pronom	PRON	NUM	D	D	DET	DET	NUM	NUM:crd:pro
Numéral cardinal, nom	NOUN	NUM	D	D	DET	DET	NUM	NUM:crd:nom
Numéral ordinal, adjectif	ADJ	NUM	D	D	DET	DET	NUM	NUM:ord:adj
Numéral ordinal, pronom	PRON	NUM	D	D	DET	DET	NUM	NUM:ord:pro
Numéral ordinal, nom	NOUN	NUM	D	D	DET	DET	NUM	NUM:ord:nom
Préposition	ADP	PRP	P	P	P	PREP	PRP	PRP
Préposition + Déterminant (contracté)	ADP	PRP:det	P	P	P+D	PREP	PRP:det	PRP:det
Préposition + Pronom	ADP	PRP	P	P	P+PRON	PREP	PRP	PRP
Préfixe	PRT	_	_	PREF	PREF	_	_	PFX
Ponctuation	.	PUN	PUNC	PONCT	PONCT	PCT	SENT	PNC
Ponctuation citation	.	PUN:cit	PUNC	PONCT	PONCT	PCT	SENT	PNC
Pronom démonstratif	PRON	PRO:DEM	PRO	PRO	PRO	P/DEM	PRO:dem	PRO:dem
Pronom indéfini	PRON	PRO:IND	PRO	PRO	PRO	P/IND	PRO:ind	PRO:ind
Pronom interrogatif	PRON	PRO	PRO	PRO	PROWH	P/INT	PRO:int	PRO:int
Pronom personnel, sujet	PRON	PRO:PER	PRO	PRO	CLS	PPERSUJ	PRO:cls	PRO:per:sjt
Pronom, impersonnel, sujet	PRON	PRO:PER	PRO	PRO	CLS	PPERSUJ	PRO:clsi	PRO:per:sjt

Part-of-Speech class	Universal POS tagset	TreeTagger	Stanford POS tagger	French Treebank	MElt	Cordial adapté	TreeTagger TCOF-POS	DisMo
Pronom personnel, objet direct	PRON	PRO:PER	CL	Cl	CLO	PPERCOMPL	PRO:clo	PRO:per:objd
Pronom personnel, objet indirect	PRON	PRO:PER	CL	Cl	CLO	PPERCOMPL	PRO:clo	PRO:per:obji
Pronom possessif	PRON	PRO:POS	PRO	PRO	PRO	PPOSS	PRO:pos	PRO:pos
Pronom relatif	PRON	PRO:REL	CL	Cl	PROREL	PREL	PRO:rel	PRO:rel
Pronom réflexif	PRON	PRO:PER	CL	Cl	CLR	PPERCOMPL	PRO:clo	PRO:ref
Pronom personnel tonique	PRON	PRO:PER	PRO	PRO	PRO	PPERTON	PRO:ton	PRO:per:ton
Symbole	.	SYM	_	_	_	_	_	PNC
Verbe conditionnel	VERB	VER:cond	V	V	V	VCON	VER:cond	VER:cond
Verbe conditionnel auxiliaire	VERB	VER:cond	V	V	V	VCON	AUX:cond	VER:cond:aux
Verbe futur	VERB	VER:futu	V	V	V	VF	VER:fut	VER:fut
Verbe futur auxiliaire	VERB	VER:futu	V	V	V	VF	AUX:fut	VER:fut:aux
Verbe gérondif	VERB	VER:ppre	V	V	VPR	V	VER	VER:ger
Verbe impératif	VERB	VER:impe	V	V	VIMP	VPIMP	VER:impe	VER:impe
Verbe impératif auxiliaire	VERB					VPIMP	AUX:impe	VER:impe:au
Verbe imparfait	VERB	VER:impf	V	V	V	VI	VER:impf	VER:impf
Verbe imparfait auxiliaire	VERB	VER:impf	V	V	V	VI	AUX:impf	VER:impf:aux
Verbe infinitif	VERB	VER:infi	V	V	VINF	VINF	VER:infi	VER:inf
Verbe infinitif auxiliaire	VERB	VER:infi	V	V	VINF	VINF	AUX:infi	VER:inf:aux
Verbe participe passé	VERB	VER:pper	V	V	VPP	VPPPAS	VER:pper	VER:ppas
Verbe participe passé auxiliaire	VERB					VPPPAS	AUX:pper	VER:ppas:au
Verbe participe présent	VERB	VER:ppre	V	V	VPR	VPPRES	VER:ppre	VER:ppre
Verbe participe présent auxiliaire	VERB					VPPRES	AUX:ppre	VER:ppre:aux
Verbe présent	VERB	VER:pres	V	V	V	VPINP	VER:pres	VER:pres
Verbe présent auxiliaire	VERB	VER:pres	V	V	V	VPAUX	AUX:pre	VER:pres:aux
Verbe présentatif	VERB	_	_	_	_	PRES	_	VER:pres:ent
Verbe passé simple	VERB	VER:simp	V	V	V	VPPPAS	VER:simp	VER:simp
Verbe passé simple auxiliaire	VERB	VER:simp	V	V	V	VP	AUX:simp	VER:simp:aux
Verbe subjonctif imparfait	VERB	VER:subi	V	V	VS	VPSUB	VER:subi	VER:subi
Verbe subjonctif imparfait	VERB	VER:subi	V	V	VS	VPSUB	AUX:subi	VER:subi:aux
Verbe subjonctif présent	VERB	VER:subp	V	V	VS	VPSUB	VER:subp	VER:subp
Verbe subjonctif présent auxiliaire	VERB	VER:subp	V	V	VS	VPSUB	AUX:subp	VER:subp:aux
Amorce de nom commun	_	_	_	_	_	_	NAM:trc	(separate
Amorce de verbe	_	_	_	_	_	_	NOM:trc	encoding)
Amorce de nom propre	_	_	_	_	_	_	VER:trc	
Amorce d'adjectif	_	_	_	_	_	_	ADJ:trc	
Marqueur de discours	_	INT	I	I	I	MD, MDEUH, MDINT	INT	MD

TABLE 12.2: Comparison of part of speech tag-sets for French: Universal POS tag-set, French Treebank, TreeTagger, Stanford POS tagger, Melt tagger, modified version of the Cordial tag-set, tag-set used in TCOF-POS, and the part of speech tag-set used by DisMo.

The comparison of tag-sets shows that the tag-set proposed by DisMo is one of the most detailed ones in the literature. At the same time, by using only the first level of the tags, it is possible to perform a coarse analysis, with higher accuracy. Table 12.2 is also a valuable tool for interoperability: using these correspondences it is possible to convert between part-of-speech annotations in corpora. If the source tag-set is finer-grained than the target tag-set the conversion is straightforward; if the target tag-set is more detailed than the source tag-set, we can use this table to identify those occurrences of tags that will need to be manually corrected (made more specific). The 12 main categories in DisMo roughly correspond to the major POS categories in the Universal POS Tagset (Petrov, Das, & McDonald, 2012). The Universal POS Tagset project has compiled gold-standard POS-annotated corpora in 22 languages, and using DisMo it is easy to contribute to this effort and perform coarse-level cross-linguistic studies.

The part-of-speech tags shown in Table 12.2 are applied to the pos-min and pos-mwu level, with one supplementary tag (INTROD) being only applicable to multi-word units. DisMo also provides lemmatisation, using its dictionary (see next section on language resources used). The dictionary is also the source for the extended morphosyntactic information (e.g. gender and person). However, it must be noted that at this stage of development of the system, there is no module to apply grammatical agreement rules; the pos-ext field contains all possible combinations of gender and person for a given lexical form (e.g. the tag "ms:fs" would indicate that the form is either masculine singular or feminine singular in this context). At the level of minimal tokens, the attribute "disfluency" is used to encode several types of disfluencies: hesitations, drawls, lexical false starts, repetitions, interruptions, different types of self-corrections (substitutions, insertions), etc. In the next chapter we present the detailed annotation scheme used by DisMo for annotating disfluencies.

With regards to the annotation tag-set used for chunking, we based our choices on the study by Tellier, Duchier, et al. (2012), which presents a simplified version of the constituent-level annotation tag-set used in the French Treebank (Abeillé, 2003). The French Treebank contains approximately 200k tokens of texts from the newspaper Le Monde, with manually corrected part-of-speech tags and a syntactical analysis using a constituency grammar. We converted the POS tags in the French Treebank to the corresponding first-level (coarse) POS tags used in DisMo, and used the French Treebank corpus as a training corpus for the chunker (for the technical details of the chunker, see next section). Table 12.3 summarises the tag-set used by DisMo's chunking algorithm.

According to Abney (1991), a chunk is a "non-recursive core of an intra-clausal constituent, extending from its beginning to its head"; chunks are non-overlapping. Given that speakers frequently perform interruptions of the syntactical structure, as they incrementally construct their message, chunks

Code	Chunk Description
AP	Adjectival chunk (including modifier adverbs)
AdP	Adverbial chunk
NP	Nominal chunk (nouns, pronouns and dependent determinants)
PP	Prepositional chunk (nominal chunk introduced by a preposition; or verb qualifier introduced by a preposition)
VN	Verbal chunk (this includes the categories VN, VPinf and VPpart of the French Treebank)
CONC	Co-ordinating conjunction
CONS	Subordinating conjunction

TABLE 12.3: Chunk tag-set used by DisMo

are a useful intermediate representation for the syntactical analysis of spoken data. Chunk parsing is a technique whereby a parser seeks to establish relations between chunks, rather than between words, in an effort to reduce computational complexity. In future developments of DisMo, we plan to explore this technique: therefore, the chunking module provides an intermediate representation of the, often "non-canonical", syntactical structure of utterances.

12.4 DISMO SYSTEM ARCHITECTURE

DisMo is designed with a modular structure, where each module adds and modifies annotation information based on the output of the previous steps. The modules communicate through a shared memory structure (the token unit list). The cascade of modules runs as follows:

- Tokenisation. This module is customisable to allow for different transcription conventions.
- Application of language resources. The token list is run through the lexicon, non-ambiguous units are annotated, and a list of possible part-of-speech tags is created for the rest of the units. Certain simple disfluencies and unambiguous multi-word units are detected at this stage. The data structure is enriched with flags indicating that a series of tokens is a potential multi-word unit or a potential discourse marker.
- First pass of the part of speech annotation, using a statistical model based on Conditional Random Fields (CRF).
- Disfluency detection and annotation (using both rule-based and statistical techniques, outlined in the next chapter).

- Chunking (optionally), using a CRF statistical model. This model was trained on the French Treebank and uses the BIO (beginning-inside-outside) system to model chunking as a sequence labelling problem.
- Second and final pass part of the part-of-speech annotation, including detection of multi-word units, using a CRF statistical model.
- Post-processing and application of rules to ensure the coherence of annotations.

DisMo is written in C++ and uses multiple third-party open source libraries (most notably the Qt platform and Taku Kudo's CRF++ toolkit for training statistical models using conditional random fields). The language resources used in the version of DisMo for French were created by combining open source resources, including the DELA dictionaries (Courtois, 1990; Courtois et al., 1997), the GLÀFF dictionary (Sajous, Hathout, & Calderone, 2013) and smaller hand-crafted dictionaries of named entities. The detection of potential discourse markers is performed using the specialised dictionary LexConn (Roze, 2009).

During the first-pass POS annotation, only coarse part-of-speech categories are attributed to different tokens. These tags are used in the disfluency detection module and in the chunking module. Simple repetitions of function words (such as "le le le") are detected at the pre-processing stage, and are removed from the input to downstream modules. Note that this removal is only temporary, using a flag in the token list data structure; the tokens remain intact and annotated as repetitions in the system's final output.

12.5 EVALUATION

We have evaluated the performance of the part-of-speech tagging (at the coarse level) for different sizes of the training corpus. We are focusing on this module because its output will have significant impact on the other modules. For this experiment, the training corpus consists of samples from the PFC corpus (Paris and Lyon); and the testing corpus is the CPROM-PFC corpus.

The 127k tokens of the Paris and Lyon sections of the PFC corpus have been manually corrected by Dr Giulia Barreca. The DisMo annotation of the CPROM-PFC corpus has been corrected by François Delafontaine and Dr Mathieu Avanzi; the CPROM-PFC corpus has served as a training corpus for a the first public version of DisMo (cf. Christodoulides, Avanzi, & Goldman, 2014), which was subsequently used to annotate the Paris and Lyon sections of the PFC corpus. By using this incremental corpus annotation method, DisMo for French is currently using a training corpus of approximately 180k tokens.

The evaluation results are presented in Table 12.4. We observe that the application of lexical resources (notably the word forms dictionary) that lim-

Training Corpus Size			C	Training time (min)	Accuracy without LRs	Accuracy using LRs
Tokens	Sequences	Features				
			1	15,5	93,72%	94,98%
50 k	4055	6761860	2	19,1	93,73%	95,00%
			3	15,7	93,81%	95,05%
			1	16,4	93,78%	95,00%
60 k	4904	7798071	2	21,0	93,90%	95,06%
			3	27,4	93,91%	95,07%
			1	26,2	94,19%	95,28%
70 k	5771	8730189	2	30,9	94,19%	95,29%
			3	33,5	94,21%	95,31%
			1	28,9	94,40%	95,44%
90 k	7321	10269306	2	36,8	94,46%	95,47%
			3	44,7	94,39%	95,44%
			1	38,9	94,53%	95,52%
100 k	7866	11266500	2	47,3	94,49%	95,50%
			3	50,4	94,51%	95,51%
			1	45,1	94,59%	95,58%
120 k	9384	12962304	2	58,6	94,58%	95,57%
			3	65,0	94,57%	95,57%
			1	58,8	94,57%	95,58%
140 k	10859	14757462	2	71,6	94,58%	95,57%
			3	73,2	94,59%	95,58%

TABLE 12.4: DisMo accuracy as a function of the training corpus size, with and without the use of lexical resources (LRs)

its the search space of the initial part-of-speech annotation improves the accuracy of the first-pass POS annotation; however the gain diminishes as the training corpus becomes larger. The detection of disfluencies (especially repetitions) and downstream processing further improves the accuracy of the system. After using all data available for training, and activating all the system's modules, the accuracy of DisMo (as measured with 5-fold cross validation) is approximately 98% for the coarse POS tag-set and 97% for the fine-grained tag-set, a result that is comparable to the performance of taggers for written language.

12.6 APPLICATION TO LARGE CORPORA

The collaboration with the colleagues working on CPROM-PFC and the PFC corpus has led to substantial improvements in the performance of the DisMo annotation. The possibility of annotating large spoken corpora with a standardised and detailed tag-set opens up research possibilities (for example in

the study of regional variation in syntax, in phonetics, fluency etc.). Four major French corpora have been annotated with the latest version of DisMo:

- The Phonologie du Français Contemporain (PFC) corpus (Durand et al., 2009), containing 1,4 million tokens.
- The collection of corpora in the Valibel speech database (Simon, Francard, & Hambye, 2014), totalling approximately 6 million tokens.
- The Corpus Oral de français de Suisse Romande (OFROM) (Avanzi, Béguelin M.-J., & Diémoz, 2015), containing approximately 0,5 million tokens.
- The Enquête sociolinguistique d'Orléans (ESLO) corpus, with approximately 2 million tokens.

13

DISFLUENCIES

An important characteristic of spoken language is the prevalence of a class of phenomena called disfluencies, such as filled pauses, repetitions and false starts. Disfluencies are an interesting phenomenon to study as such, especially as they relate with the psycholinguistic processes of speech production and perception (cf. chapter I.3.3 of the present thesis). At the same time, it is necessary to take into account this class of phenomena when applying natural language processing techniques to spoken language corpora. In this chapter we will describe our work on speech disfluencies: a detailed annotation protocol was developed and refined through its application initially to the CPROM-PFC corpus (in collaboration with Dr Mathieu Avanzi), and subsequently to other corpora including LOCAS-F and C-PhonoGenre. Furthermore, we developed a specialised annotation interface to facilitate the adoption of this annotation protocol, as well as a semi-automatic system for the detection of disfluencies that integrates with the DisMo annotator presented in the previous chapter.

13.1 RELATED PREVIOUS WORK

Disfluencies can be considered as disruptions of the ideal delivery of speech, as "cases in which a contiguous stretch of linguistic material must be deleted to arrive at the sequence the speaker 'intended', likely the one that would be uttered upon a request for repetition" (Shriberg, 1994). However, as we have seen in chapter I.3 on speech production, this "ideal" delivery is often elusive, even in speaking styles that allow the speaker to prepare and rehearse his speech extensively (e.g. a planned political address where the speaker is reading a rehearsed text). As speech is the most natural expression of language, and given the varying functions that they perform, disfluencies should be studied as an integral part of language production and perception, and not as a nuisance that should be dismissed and eliminated from our corpora (cf. Shriberg, 2001).

An alternative approach is to consider disfluencies as linguistic devices used to manage the flux of time, facilitating cognitive processes both for the speaker and the listener, while they are incrementally constructing a

message through a series of steps including planning and self-monitoring (Levelt, 1989; Hartsuiker, 2014, see also chapter I.3.3). Moniz, Ferreira, Batista, and Trancoso (2015) characterise disfluencies as "online editing strategies with several (para)linguistic functions" (p. 1). Several communicative functions have been correlated with 'dis'fluencies. J. E. Arnold, Fagnano, and Tanenhaus (2003) and Bosker, Pinget, Quene, Sanders, and de Jong (2013) show that disfluencies, and especially filled pauses, may be used to draw the listener's attention to new or complex information. Clark and Fox Tree (2002) and Moniz, Trancoso, and Mata (2009) show how disfluencies can be used to manage dialogue interaction, and thus be perceived as fluent communicative devices in specific contexts.

Corpus-based studies on disfluencies in French include Adda et al. (2007) and Boula de Mareüil, Habert, et al. (2005) who investigate the interplay between disfluencies and overlaps in broadcast political interviews, as a strategy to manage interaction (e.g. hold the floor during an argument). Henry and Pallaud (2004) present a qualitative and quantitative analysis of false starts and repetitions. Pallaud, Rauzy, and Blache (2013) study the correlation of self-corrections and interruptions with other dialogue events in the CID corpus of spontaneous free-theme dialogues. And finally Vasilescu, Candea, and Adda-Decker (2004) studied the acoustic properties of filled pauses in 8 languages including French.

Work on the automatic detection of disfluencies has focused on modelling to improve the performance of automatic speech recognition (ASR) systems (e.g. Adda-Decker et al., 2004, for French) or to improve parsing performance on transcriptions of spoken language (e.g. Jørgensen (2007) for English; and Peshkov, Prévot, Rauzy, and Pallaud (2013) for French). Various feature sets and machine learning algorithms have been used: CART decision trees with exclusively lexical features (Moreno & Pineda, 2006, working on the Spanish DIME corpus); or exclusively prosodic features (Shriberg, Bates, & Stolcke, 1997, English).

Liu, Shriberg, and Stolcke (2003) show that "the detection of disfluency interruption points is best achieved by a combination of prosodic cues, word-based cues and POS-based cues" while "the onset of a disfluency is best found using knowledge-based rules" and "specific disfluency types can be aided by modelling word patterns". In Liu, Shriberg, Stolcke, and Harper (2005) and Liu, Shriberg, Stolcke, Hillard, et al. (2006), a comparison of the performance of Hidden Markov Models (HMM), maximum entropy models and Conditional Random Fields (CRF) models in detecting disfluencies using both lexical and prosodic features shows that discriminative models generally outperform generative models. Working on the English Switchboard corpus, Georgila (2009) proposed a system for disfluency detection based on CRF models combined with Integer Linear Programming (ILP) rules. In Georgila, Wang, and Gratch (2010) CRF models are compared to ILP, showing that ILP performed better when there was relatively few domain-specific data

for training. Mieskes and Strube (2008) investigate the gains from including dialogue interaction data into an automatic disfluency event classifier.

Working on French, Dutrey, Rosset, Adda-Decker, Clavel, and Vasilescu (2014) present the results of a series of experiments with CRF-based automatic disfluency detection, working on a corpus of recordings of customer service interactions; the objective is to render the manually-transcribed data more readable in order to improve automated opinion mining. Bouraoui and Vigouroux (2009) have worked on automatic disfluency detection in French air traffic control conversations (a linguistically constrained, standardised type of interaction). Peshkov et al. (2013) demonstrate that it may be more efficient to perform the tasks of syntactic chunking and disfluency detection simultaneously.

Finally, Germesin, Becker, and Poller (2008) suggest using hybrid systems for disfluency detection, i.e. "different detection techniques where each of these techniques is specialised within its own disfluency domain"; we have adopted this approach for the disfluency detection module in DisMo.

In the following section, we will present a detailed, language-independent protocol for the annotation of disfluencies in speech corpora.

13.2 A MULTI-LEVEL ANNOTATION PROTOCOL FOR DISFLUENCIES

The annotation protocol is designed so that it can be integrated with the multi-level annotation scheme of the DisMo annotator, presented in the previous chapter. The term "disfluencies" is used both for phenomena that are phonetic/prosodic in nature (i.e. are related with the articulation of speech and its super-segmental properties) and for syntactical phenomena (i.e. are related with the order of words in an utterance). Consequently, the annotation protocol is articulated over two levels of annotation: the syllable level is used to annotate phonetic/prosodic features of disfluencies, and the minimal token level is used to define stretches (sequences) of one or more tokens that are affected by disfluencies. The same tokens may be linked to several types of disfluency, as these phenomena co-occur: for example, we may annotate a repetition as affecting two tokens, separated by a filled pause, and concurrently, annotate a perceived hesitation-related lengthening on the last syllable of the first token. In the following, we will provide examples that show how many different types of disfluencies can be annotated concisely, by using the codes corresponding to each level. Table 13.1 summarises the different types of disfluencies, and gives typical examples of each type, drawn from the annotation of the CPROM-PFC corpus.

The DisMo annotation protocol for disfluencies combines aspects from the systems described in Shriberg (1994), Brugos and Shattuck-Hufnagel (2012) and Heeman, McMillin, and Yaruss (2006). We have also consulted previous work on annotation schemes for disfluencies in spoken corpora: in English, notably the annotation scheme proposed by the Linguistic Data Consortium

LEVEL 1: Simple disfluencies are those affecting only one minimal token		
FIL	Filled pauses	c' est pour ça que j' hésite euh un peu en parler FIL
LEN	Hesitation -related lengthening	au cercle d'oenologie de= Bruxelles LEN
FST	Lexical false start	comme infirmière so/ sociale FST
WDP	Silent pause within word	il m' a dit ça su+ __ +ffit WDP
LEVEL 2: Repetitions, i.e. one or more tokens repeated in exactly the same form		
REP	Repetition	les disques et et lancer les jingles REP* REP_ il a il a il a dit que REP:1 REP:2 REP:1 REP*:2 REP_ REP_ c' est pas c' est pas un système génial REP:1 REP:2 REP*:3 REP_ REP_ REP_
LEVEL 3: Structured editing disfluencies		
DEL	Deletion	c' est vraiment un en tout cas la parole DEL DEL DEL DEL*
SUB	Substitution	cette personne était enfin c' est un ami de SUB* SUB:edt SUB_ SUB_
INS	Insertion	c' est vrai que Béthune euh vivre à Béthune a été INS* INS+FIL INS_ INS_ INS_
LEVEL 4: Complex disfluencies are a combination of several structured ones that cannot be described by combining the codes of the previous 3 levels		
COM	Complex	les ac/ les actions enfin les activités enfin professionnelles COM COM COM COM COM COM COM COM COM

TABLE 13.1: Annotation Scheme for disfluencies in DisMo

in the context of the MDE project that was used for the annotation of the Switchboard corpus (Mateer & Taylor, 1995); and in French, the work of the VoxForge project (Clavel et al., 2013) and the disfluency annotation protocol for the CID corpus (Pallaud et al., 2013). In the following, we will describe the different types of disfluencies and the annotation codes used to describe them.

1. Simple disfluencies are those affecting only one minimal token, and include:
    - Filled pauses, including both autonomous fillers and epenthetic vowels. Note that given the corresponding phonetic transcription, it is easy to separate autonomous fillers from epenthetic vowels, and to study the acoustic properties of vowels used in filled pauses.

- Hesitation-related syllable lengthening. In corpora containing a phonetic transcription and syllabification, it is possible to indicate precisely which syllable(s) are lengthened, by using the LEN annotation code at the syllable level as well. According to the annotation protocol, this information is duplicated: a LEN code on a token indicates that it is being perceived as containing one or more lengthened syllables (those bearing LEN codes at the syllable level).
- Lexical false starts, i.e. fragments of words whose articulation was interrupted by the speaker
- Intra-word silent pauses.
- Mispronounced words. The actual and reference phonetic form may be stored in an additional attribute of the tok-min level.

2. Repetitions are defined as strings of one or more tokens that are repeated by the speaker in exactly the same form, possibly interspersed with filled or silent pauses, as long as this repetition does not serve a grammatical or emphatic purpose. For example, in the utterance "le le le chien", we consider that the definite article "le" is repeated; however in the utterance "il est très très joli", the repetition of the adjective "très" is emphatic and thus not annotated as a disfluency (cf. Dister, 2014). The REP code is attributed to all the tokens in the repetition sequence; intervening silent or filled pauses are therefore annotated as REP+SIL and REP+FIL.

3. Structured disfluencies include:
   - Substitutions: the speaker backtracks and modifies some tokens already uttered, using the same syntactic structure (e.g. "normalement je louais enfin je loue toujours un appart"),
   - Insertions: the speaker backtracks and adds tokens, using the same syntactic structure (e.g. c'est vrai que Béthune vivre à Béthune. . .)
   - Deletions, also called syntactical interruptions: the speaker abandons a series of tokens and continues starting a new syntactic structure (e.g. c'est vraiment en tout cas la parole...).

Note: To be annotated as such, structured disfluencies must follow the reparandum – interregnum – repair pattern; otherwise we use the annotation scheme for complex disfluencies.

Repetitions and structured disfluencies can be described as a sequence of three contiguous regions, in line with Shriberg (1994):

```
(reparandum) * interruption point (interregnum, including optional editing terms) (repair)
```

The reparandum is the part of the utterance that is repeated or that will be corrected, edited, or deleted. The interruption point is the point between the reparandum and the interregnum; this instance in time does not necessarily coincide with the moment the speaker detected the trouble or his intention to alter the utterance. The interregnum is the part between the reparandum and the repair. It may optionally include explicit editing terms, i.e. words or phrases used by the speaker to signal the correction (e.g. "enfin"). The repair is the continuation of the message that follows the disfluency, so that if the first two regions are removed the remainder is lexically fluent (Shriberg, 2001). We have chosen not to include discourse markers as "disfluencies", given that they are annotated on a separate annotation level, and that information regarding the concurrence of disfluencies with discourse markers can be easily extracted with queries crossing these annotation levels.

Whenever more than one token are repeated, we use numbering to indicate the repetition pattern: the first token in the repeated sequence is annotated as REP:1, the second as REP:2 etc. Given the series of tokens affected by a repetition (from the reparandum to the repair), it is possible to use pattern matching to find the interruption point and assign the numbering. For repetitions and structured disfluencies (codes REP, SUB, INS and DEL), we use extension to the tags to indicate these regions: an asterisk (*) is appended at the interruption point, i.e. at the end of the tag of the last token of the reparandum; explicit editing terms are indicated by appending ":edt" to the main tag; and the repair region is signalled by appending an underscore (_) to the main tag. Note that for deletions the repair is not annotated, since anything that follows the material deleted could be considered as the repair.

The annotation scheme is hierarchical, in the sense that Level 1 annotation codes (simple disfluencies) may combine with Level 2 and Level 3 codes; and Level 2 codes (repetitions) may combine with Level 3 codes (structured disfluencies). In the example of an insertion given above, a filled pause produced within the interregnum region of the insertion is annotated as INS+FIL. This allows us to model the co-occurrence of simple disfluencies within the interregnum part or near the interruption point of structured disfluencies. Complex disfluencies are the result of combining several structured disfluencies. Unlike Shriberg (1994) we do not limit the annotation scheme to nested complex disfluencies, but adopt the "backtracking table" notation proposed by Heeman et al. (2006).

13.3 BASELINE CORPUS STUDY: DISFLUENCIES IN SPONTANEOUS FRENCH SPEECH

The annotation protocol described in section 3.2 was applied by two trained annotators on the 7-hour Spontaneous Speech sub-corpus of CPROM-PFC, in order to study the distributional and prosodic characteristics of disfluencies in spoken French (Christodoulides & Avanzi, 2014, the results of this

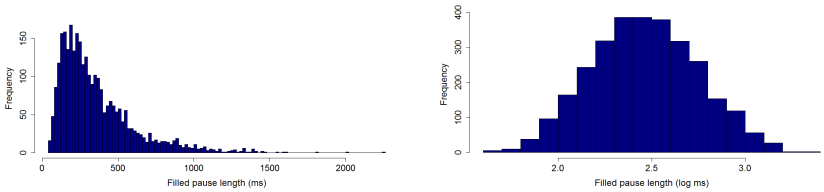
study are presented in). Our findings were generally in line with previous studies on the phonetic and prosodic features of disfluencies (e.g. for English, Shriberg (1999, 2001); for French, Vasilescu et al. (2004); for Portuguese, Moniz, Batista, Trancoso, and Mata (2012)).

Regarding the relative frequency of different types of disfluencies, we found that filled pauses (both autonomous fillers and epenthetic vowels) are the most common single-token disfluencies, followed by lengthening (drawls). Furthermore, 82% of lexical false starts co-occur at the interruption point of structured disfluencies. Among the structured disfluencies, repetitions (especially 1- and 2- token) are the most prevalent, followed by deletions and substitutions. The most common patterns of co-occurring disfluencies are show in Table 13.2.

Pattern	Description	Occurrences
`FIL SIL:l`	Filled pause followed by a long silent pause	446
`REP* REP_`	Repetition of one token	186
`LEN SIL:l`	Lengthening followed by a long silent pause	121
`FIL SIL:b`	Filled pause followed by a short silent pause	100
`REP* REP+SIL REP_`	Repetition of one token, with a silent pause in-between	90
`LEN LEN`	Consecutive lengthening of two tokens	74
`LEN FIL`	Lengthening followed by a filled pause	68
`REP:1 REP*:2 REP_ REP_`	Repetition of two tokens (A B A B)	58
`REP:1 REP*:2 REP+SIL REP_ REP_`	Repetition of two tokens (A B A B), with a silent pause in-between	39
`DEL DEL*`	Deletion of two tokens	38
`REP* REP+FIL REP+SIL REP_`	Repetition of one token, with a one filled and one silent pause in-between	34
`SUB* SUB_`	Substitution of one token with another one	33
`REP:1 REP*:1 REP_`	Double repetition of the same token (A A A)	33
`DEL*`	Deletion of one token	31
`LEN SIL:b`	Lengthening followed by a short silent pause	30
`LEN FIL SIL:l`	Lengthening followed by a filled pause and then a long silent pause	30
`DEL DEL* SIL:l`	Deletion of two tokens followed by a long silent pause	26

TABLE 13.2: Patterns of co-occurrence of disfluencies in the CPROM-PFC Spontaneous Speech sub-corpus (consisting of 7 hours of sociolinguistic interviews)

Regarding the prosodic and phonetic characteristics of disfluencies, we found that in the reparandum region of structured disfluencies, the syllables affected are those immediately preceding the interruption point (lengthening or trailing fillers or short silent pauses). Using a machine learning algorithm for sequence labelling (a CRF model) it is possible to differentiate between



(a) Original distribution (b) Log-transformed distribution

FIGURE 13.1: Filled pause length distribution in the CPROM-PFC Spontaneous Speech sub-corpus: left: the original distribution; right: the distribution of the logarithm of filled pause length

prosodically prominent syllables and syllables lengthened due to hesitation. The duration of filed pauses follows a unimodal log-normal distribution, as can be seen in Figure 13.1. Filled pauses exhibit falling intonation contours, with a mean pitch close the speaker's mean pitch; however the formant values of vowels in filled pauses are significantly different from those of the same vowels in fluent contexts.

Finally, we studied prosodic cues that may signal the interruption point of a disfluency: articulation rate, pitch and energy increase in the repair region (compared to the reparandum) possibly signalling the boundaries of each region. Three measurements (articulation rate, mean pitch and mean peak intensity) were taken 500 ms before the interruption point and 500 ms after the start of the repair region. The means of the differences (repair – reparandum) were 0.77 syll/s, 0.38 ST and 0.96 dB respectively (in all cases: $p<0.001$; pairwise t-test). These cues are not always present, but may aid listeners, as well as automatic detection systems.

13.4 AUTOMATIC DETECTION OF DISFLUENCIES

In this section, we describe the automatic disfluency detection system in DisMo. The results have been reported in (Christodoulides & Avanzi, 2015).

Disfluency detection in DisMo is organised in different modules, with each module being responsible for detecting and annotating specific types of disfluencies, using the most appropriate method for each type. We have chosen a design similar to the one proposed by Germesin et al. (2008) because the analysis of the CPROM-PFC corpus (section 3.3) indicated that some types of disfluencies are clearly more frequent than others.

The detection of filled pauses (FIL) and lexical false starts (FST) from the raw output of an automatic speech recognition system is outside the scope of this paper: we assume that the input already contains these phenomena and that transcription conventions are sufficient to identify them: for example,

filled pauses are transcribed as either "euh" or "euhm" (generally, a fixed list of tokens), and lexical false starts are followed by a slash character, as in "mar/" (generally, a transcription convention that can be identified using regular expressions). Although this limits the applications of our system, it is a reasonable assumption given that our initial objective is to facilitate the annotation of existing large French spoken language corpora. Similarly, intra-word pauses (WDP) are identified only on the basis of string matching.

The detection of hesitation-related lengthening of syllables is optional, used for corpora that include a reliable automatic or manual syllabification. It is based on a Support Vector Machine (SVM) classifier that takes into account the following prosodic features: the length of the syllable relative to windows of ± 3 neighbouring syllables (in order to normalise for articulation rate), its structure (consonants-vowels), its position within the token, and relative pitch over the same windows. The objective is to distinguish increases in syllable length that will be perceived as a hesitation, rather that acoustic cues of prosodic prominence.

The following step in the cascade is the detection of repetitions using a Conditional Random Fields (CRF) model, with the following features: word form, part-of-speech (DisMo level-1 tag, i.e. one out of 12 main POS categories, and level-2 tag, i.e. one out of the 64 specific POS tags), whether the token is equal to the following 1, 2... 7 tokens. Combinations of such features are included in the model, over windows of 1 to 3 tokens. The model labels sequences as belonging or not belonging to a repetition cluster (R or O). The repetition clusters identified by the CRF model are post-processed to find patterns and assign the detailed REP annotation codes.

Using the results of the previous steps (annotation of simple disfluencies and repetitions), the next processing step is to identify editing disfluencies. Both lexical and prosodic information is input to a CRF model in order to predict the onset of the disfluency, its reparandum region and the interruption point. The lexical features include token, POS tag (as above), and edit distance with the following 1...7 tokens. An SVM classifier produces hypotheses regarding possible interruption points, based on the following prosodic features: difference in articulation rate, mean pitch and mean intensity 500 ms before and after each possible interruption point (defined as the end of the last syllable of each token); these hypotheses are also input into the CRF model as features of the corresponding token. The output of the model is post-processed to assign the detailed INS, SUB or DEL codes. The user may choose to discard sequences that cannot fit into these patterns, or to leave them annotated with a generic DIS code (depending on whether manual intervention is envisaged or not).

Having separate modules deal with each class of phenomena allows us to fine-tune their parameters to the specificities of each class. We can also use the predictions of one module into the next one: in our system the automatic detection is performed in three steps: (1) repetitions (based on lexical features,

SIL and FIL codes); (2) interruption point prediction (taking into account lexical features, POS, and the results of the repetitions module – the fact that a token forms part of a repeated structure is added as a feature to the CRF model) and (3) editing type disfluencies (taking into account the interruption point predictions, which are entered as a feature to the CRF model). The LEN module is optional.

13.5 EVALUATION

We performed an evaluation of the performance of each module separately. Table 2 summarises the accuracy, precision and recall measures (as applicable) of the different modules. As expected, the types of disfluencies occurring more frequently in our corpus are better detected by a system based on probabilistic models. In all cases we have used 5-fold cross validation: the 63k corpus is divided in 5 "folds", each containing an approximately equal number of sequences to annotate (sequences are segmented at silent pauses over 500 ms in length). Four folds are used as a training corpus and the resulting model is used to annotate the fifth one that functions as an evaluation; the reported results are the averages of applying this process with all 5 possible combinations.

Disfluency type / Method	Precision	Recall	F-measure
LEN – SVN classifier	78.2%	87.4%	82.5%
REP – CRF model	84.3%	75.8%	79.8%
IP – Interruption point hypotheses	76.7%	52.0%	62.0%
SUB, INS, DEL – CRF models	(see Table 13.4)		

TABLE 13.3: Overall evaluation of the automatic disfluency detection system

The measures presented in Table 13.3 are calculated on the token level (i.e. number of tokens correctly/incorrectly classified as being part of this specific type of disfluency): this is because LEN affects a single token, while IP is added to the single token that is an interruption point. For repetitions, the CRF model just outputs one code (REP) indicating that the token in question is part of a repeated sequence. The system then uses the iterative Diff algorithm to calculate the exact repetition codes.

The detection and annotation of editing-type disfluencies has proven to be much more challenging. Inspired by Dutrey et al. (2014), we have conducted experiments to evaluate four possible BIO (begin-inside-outside) schemes that may be the desired output of the CRF model of this module. Table 2 presents these options: the editing terms are annotated as a separate region (method 1, 3), or included in the reparandum region (method 2, 4); it may be

desirable to also predict the repair region (method 1 and 2) or not (method 3 and 4).

We have then evaluated the performance of the alternative BIO systems, in two experiments: in the first one, the correct interruption point was always given to the CRF model (setting an upper limit of the performance); in the second one, the predicted interruption points were used (actual performance). Table 13.4 summarises the results in terms of precision and recall. The results show that the strategy of considering the reparandum and the editing terms as one contiguous region, contrasted with the repair region, yields marginally better F-measure results.

	Reparandum and Editing terms	Repair region	Precision	Recall	F-measure
Gold standard Interruption Points (upper limit)					
1	Separate	Predict	77.6%	51.4%	61.9%
2	Merged		74.7%	44.7%	55.9%
3	Separate	Ignore	82.4%	62.8%	71.3%
4	Merged		76.9%	53.2%	62.9%
Predicted Interruption Points (actual performance)					
1	Separate	Predict	54.3%	36.5%	43.7%
2	Merged		48.6%	31.3%	38.0%
3	Separate	Ignore	62.6%	42.1%	50.3%
4	Merged		59.2%	36.2%	44.9%

TABLE 13.4: Evaluation of the editing disfluency CRF models

In this chapter we presented a detailed annotation protocol for disfluencies in spoken corpora. The protocol is language-independent. We have applied this protocol to a 63k token corpus (the Spontaneous Speech sub-corpus of CPROM-PFC), and developed an annotation tool to facilitate human annotators. The annotation tool automatically inserts the correct codes, given a sequence of tokens and the type of disfluency desired by the user: the tool analyses the token sequence to find an interpretation compatible with the disfluency type selected. Finally, we have presented the results of training and evaluating different machine learning algorithms for the detection and annotation of disfluencies. Our results indicate that an automatic detection of the majority of speech disfluencies in a transcribed corpus is feasible, even if, as expected, some phenomena are more easily recognizable automatically than others.

14

PROSODIC PROMINENCE

Prosodic prominence is "an umbrella term encompassing various related but conceptually and functionally different phenomena, such as phonological stress, paralinguistic emphasis, lexical, syntactic, semantic or pragmatic salience, to mention a few" (P. Wagner, Origlia, et al., 2015). A most generic definition is given by Terken (1991): "we say that a linguistic entity is prosodically prominent when it stands out from its environment by virtue of it prosodic characteristics". Prosodic prominence has been operationalised on various levels of linguistic description (e.g. at the syllable level or at the word level), and has been annotated using different scales (categorical, multi-level or continuous).

P. Wagner, Origlia, et al. (2015) is the collaborative report of a multinational meeting on prosodic prominence that took place in Capri, Italy, where phoneticians, phonologists and engineers exchanged views on the subject. Reviewing a varied body of work on prosodic prominence, P. Wagner, Origlia, et al. (2015) organise it around three perspectives. The functional perspective focuses on communicative and core linguistic functions of prominence, "its realisation being indicative of information structure, contextual givenness, phrasal stress, word order or lexical class (Baumann & Roth, 2014; Bocci & Avesani, 2011; D'Imperio, 1998; Vainio & Järvikivi, 2006)", including paralinguistic aspects, such as emotion or attitude, that "significantly contribute to the pertaining impressions of both prominence placement and strength and may be confounded with the linguistic core functions of prosodic prominence" (p. 2). The physical perspective treats prominence as a psycho-acoustic rather than a communicative effect, describes it as a continuous rather than a categorical phenomenon, and focuses on established signal correlates to perceived prominence: fundamental frequency excursion and shape, duration, voice source features such as spectral tilt, open quotient, excitation strength, intensity and loudness, hyper-articulation, and multimodal cues such as eyebrow and head movements and gesture (Rietveld and Gussenhoven, 1985; Turk and Sawusch, 1996; de Jong, 1995; Al Moubayed, Beskow, and Granström, 2009; Chasaide, Yanushevskaya, Kane, and Gobl, 2013 — cited in P. Wagner, Origlia, et al., 2015, p. 2). The cognitive perspective focuses on perceptual processing, i.e. "the low-level neural pathways and psycho-acoustic

processing mechanisms that contribute to higher-level cognitive processing (Ludusan, Origlia, & Cutugno, 2011; McAngus Todd & Brown, 1996). Such high-level processes are known to be strongly shaped by linguistic knowledge including linguistic and paralinguistic functions as well as situation-specific expectations (Cole, Mo, & Hasegawa-Johnson, 2010; P. Wagner, 2005; Lacheret, Simon, Goldman, & Avanzi, 2013)" (p. 2).

Each perspective in isolation will provide only a narrow picture of the nature of prosodic prominence (the authors evoke the Indian tale of blind men trying to describe an elephant). "A physical signal perspective not taking into account function related categorical judgements or function-related contextual embedding will fail to model what makes prominence 'prosodic'. This is enhanced by findings showing that prominence perception is driven largely by top-down expectations and at least partly independent of signal correlates, e.g. prominence may be perceived because 'it belongs there'. Likewise, a purely functional perspective may miss out on signal related aspects if it relies on simple 1:1 function-signal mappings or on overly simplistic signal correlates. It may also fail to acknowledge the additional impact of processing constraints influencing prominence perception such as rhythmic expectations or attentional processes which may even lend themselves to be exploited by a language's phonological system. The cognitive perspective relies on linguistic categories as these either constrain its models or explain its data-driven results" (P. Wagner, Origlia, et al., 2015, p. 3). They conclude with a set of methodological recommendations as a roadmap for research on prosodic prominence.

We apply this roadmap to describe the work presented in this chapter. In this series of studies, undertaken in collaboration with Dr. Mathieu Avanzi, prosodic prominence was approached from a primarily functional and secondarily physical perspective. We seek to discover the acoustic and syntactic correlates that influence the perception of prominent syllables, based on annotations by experts of a large corpus (CPROM-PFC). We are therefore interested in both bottom-up signal correlates and top-down expectations; however, there were essentially no perceptual or attentional constraints placed upon the listeners (in expert annotation, the annotator may revise his choices after listening to an excerpt multiple times). We define prominent syllables as those perceived to stand out of their environment due to a combination of acoustic signal properties and top-down, linguistic expectations. We are studying this perception in French, where syllabic prosodic prominence contributes substantially to prosodic grouping and boundary demarcation (e.g. Mertens, 1991); it is therefore linked to further work on the perception of prosodic boundaries (see next chapter).

We compare different machine learning techniques for the automatic detection of prominent syllables, using prosodic features (including pitch, energy, duration and spectral balance) and lexical information (including part-of-speech categories). We explore the differences between modelling the de-

tection of prominent syllables as a classification or as a sequence labelling problem, and combinations of the two techniques. An automatic system is trained and evaluated on the expert prominence annotation of the CPROM-PFC corpus, which consists of almost 100 different speakers (balanced for age and gender) and covers three regional varieties of French. The result of this work is an automatic annotator of prosodically prominent syllables, based on statistical models, called Promise, which can be easily used to annotate new corpora through an interface provided by Praaline (see Chapter 17).

14.1 RELATED PREVIOUS WORK

As demonstrated by the review of P. Wagner, Origlia, et al. (2015), there is general agreement that prominence can be defined as the phenomenon of a linguistic unit (syllable, word, or even a larger stretch of speech) being perceived as standing out of its environment (Terken, 1991; Terken & Hermes, 2000). The perception of prominence is influenced both by bottom-up acoustic cues and by top-down expectations (D. Arnold & Wagner, 2008; Cole et al., 2010; P. Wagner, Tamburini, & Windmann, 2012). It is possible to detect syllables that will be perceived as prosodically prominent by listeners based on a set of acoustic parameters (including pitch, duration, energy and spectral features) and lexical information (part-of-speech tags). In French, syllabic prominence is of crucial importance because it essentially contributes to mark the boundaries of prosodic groups (unlike variable-stress languages, such as English or Dutch, where "stress" is a lexical property of a specific syllable in a word, while "accents" are used to signal the information status of a linguistic unit).

Experiments on the perception of prosodic prominence have shown that the reliability between two or more annotators can reach a satisfactory level (e.g. Buhmann et al., 2002 for Dutch; Mo, Cole, and Lee, 2008 for English; Smith, 2011 for French). It is estimated that the accuracy between trained and experienced human raters lies between 85% and 90% in the best case scenario (Wightman & Ostendorf, 1994; Avanzi et al., 2010; D. Arnold, Wagner, & Baayen, 2013). When compared with human annotation, the accuracy of unsupervised methods for prominence detection barely exceeds 80% (e.g. Tamburini, 2003; Goldman, Avanzi, Simon, Lacheret, & Auchlin, 2007). Supervised methods, i.e. algorithms trained on transcribed material that has been labelled for prosodic prominence by experts, may reach an accuracy of 80% to 90% (accuracy being the ratio of units correctly labelled to the total number of units).

In previous work to create automatic labelling systems for detecting syllabic prosodic prominence, Avanzi et al. (2010) used a corpus-based learning method (inferring rules from an analysis of a labelled corpus) and obtain an F-measure of almost 80%; the result of this work was consolidated in the Analor annotator. Goldman, Avanzi, Simon, et al. (2007) present a rule-based

algorithm, where rules were inferred by studying a corpus annotated by three experts; the result of this work led to the creation of the ProsoProm annotator, which provides a binary classification (a syllable is labelled as prominent or non-prominent). In follow-up work, Goldman, Avanzi, Auchlin, and Simon (2012) present a script called ProsoGrad that provides a four-level classification score for syllabic prosodic prominence. Most of the prosodic features used in both ProsoProm and ProsoGrad are extracted by applying the Prosogram (Mertens, 2004) collection of scripts for Praat (Boersma & Weenink, 2016). All three annotators for French mentioned above are rule-based classifiers, created on the basis of supervised training.

D. Arnold, Wagner, and Baayen (2013) explore the use of Random Forest classification and derive a statistical model that explains 85% of the observed variance in their corpus using a reduced set of acoustic features. Cutugno, Leone, Ludusan, and Origlia (2012) compare the use of statistical models based on Conditional Random Field (CRF) and Latent Dynamic CRFs (LDCRF), reporting an F-measure of 73.3% and 75.1% respectively. Using various classification models and linear discriminant analysis (LDA) to select the most promising features, Obin, Rodet, and Lacheret (2009) reach an F-measure between 84.1% and 87.5%.

These datasets on which the above-mentioned algorithms are evaluated and/or trained include different speaking styles (read speech and spontaneous speech, for most studies), and up to 30 speakers. The corpus size is measured in syllables, ranging between 6k and 17k syllables; they correspond to corpora between 20 minutes and 120 minutes long. Our study was conducted using a significantly larger corpus, annotated by experts for syllabic prosodic prominence. We evaluate several different algorithms for automatic prominence detection, framing the task either as a classification problem, or as a sequence labelling problem.

14.2 PROMISE: A TOOL FOR AUTOMATICALLY ANNOTATING PROSODIC PROMINENCE

We used the CPROM-PFC corpus to train and test statistical models predicting which syllables will be perceived as prosodically prominent. The CPROM-PFC corpus is presented in chapter 1; for this study we used the 2014 version of the corpus (Avanzi, 2014), which is approximately 11 hours-long, and includes approximately 113 thousand tokens (63k in semi-directed interviews and 47k in reading).

Prominent syllables and syllables associated with disfluency (fillers, lengthened syllables due to hesitations, false starts, repairs, etc.) were identified independently by two experts on the basis of their perceptual judgment only, following the C-PROM methodology, presented in detail in Avanzi et al. (2010). A third expert intervened in cases of disagreement between the two annotators and decided the final value of the syllable (+/- prominent, +/- associ-

ated to a disfluency) in a dedicated tier. Data labelling was performed over a period of almost three years, and four couples of annotators took turns. Kappa statistics (Cohen, 1960) were used to assess the reliability for each pair regarding prominence annotation, and lead to Kappa values varying between 0.61 and 0.88, with a mean of 0.72, which is considered as "substantial" according to Landis and Koch (1977).

The dataset was processed by using Praaline (Chapter 17), and using its interface we applied Prosogram (Mertens, 2004) for pitch stylisation on the entire corpus. Prosogram's algorithm operates in two phases; for each syllable, vocalic nuclei are detected based on intensity and voicing. The fo curve on the nucleus is then stylised into a static or dynamic tone, based on a perceptual glissando approach. After this pre-processing step, a number of acoustic, prosodic and lexical features were extracted for each syllable, including:

- Syllable duration (ms);
- Minimum, maximum and mean pitch (stylised fo, in semitones);
- Pitch movement (within syllable and between successive syllables);
- Peak intensity within the syllabic nucleus;
- Spectral balance;
- Token (word) to which the syllable belongs, and Part-of-Speech tag of the token;
- Presence and duration of subsequent pause;
- Syllabic structure (C/V, whether the syllable ends with a schwa);
- Position of the syllable relative to the token (word) in which it belongs: initial, final, penultimate, mono-syllabic word.

The selection of the above-mentioned features is based on the findings in previous research relating to the perception of syllabic prominence in spoken French. Relative measures were calculated for each syllable, with different contexts: the mean value of a measure is calculated over a symmetric window of 2, 3, 4 and 5 syllables (before and after the current one); silent pauses block the context window. Pitch and intensity measures are in logarithmic scales (pitch is converted to semitones relative to 1 Hz; intensity is in dB). A z-score transformation was applied to normalize the data over each corpus recording (sample), thus taking into account speaker variation (e.g. the differences in pitch register between different speakers). We followed the definition of spectral balance outlined in Sluijter and van Heuven (1996). According to this approach, the mean energy of four adjacent frequency bands is measured (0-500 Hz, 0.5-1 kHz, 1-2 kHz and 2-4 kHz). All calculated features were stored in the database.

Several machine learning algorithms were tested on the task of predicting whether a syllable should be labelled as prominent or non-prominent. Despite the fact that previous research has shown that prominence is perceived as a gradual phenomenon (P. Wagner, Tamburini, & Windmann, 2012; Goldman, Avanzi, Auchlin, & Simon, 2012, e.g.), we frame the problem as a binary classification question, because (a) the experts in charge of the data labelling were asked to annotate syllables as prominent or not; and (b) this is the most practical approach for an automated system whose outputs will be used in later stages of processing (e.g. for minor prosodic phrasing annotation).

This task is traditionally treated as a classification problem: we thus tested a Decision Tree classification algorithm, where attribute selection was performed based on the Information Gain criterion (Quinlan, 1986); Support Vector Machines (C-SVM with a radial basis function kernel); Neural Network classification (ANN with 20 hidden layer neurons); and a Random Forest of 30 decision trees (Breiman, 2001). However, the annotation of prosodic prominence can also be defined as a sequence labelling problem: given a sequence of syllables, separated by silent pauses, a machine learning method attributes a prominent / non-prominent label to each syllable, based on its features and its context. A particularly suitable method for this type of annotation is based on Conditional Random Fields (Lafferty et al., 2001).

We conducted an initial testing of the performance of classification algorithms using a reduced set of features (see infra) using the Orange (Demšar et al., 2013) data mining framework. The most promising classifiers (Random Forest and SVM), as well as the CRF sequence labelling method, were selected for further testing. To fine-tune the parameters of each algorithm and perform detailed cross-validation tests, we developed a plugin for Praaline, in C++ using the open source libraries RF-ACE, libSVM and CRF++. This tool allows for the selection of different sets of features, training models on different subsets of the dataset and testing the prediction accuracy, precision, recall and F-score.

14.3 EVALUATION AND DISCUSSION

The results of the preliminary testing of the performance of different algorithms are shown in Table 14.1. A sub-corpus of 49k syllables and a reduced set of features (relative duration, relative max pitch, relative mean pitch, relative intensity within a context of 2, 3 and 4 syllables; presence of pause; position of syllable within word) were used for training. The evaluation was performed using 5-fold cross validation.

We then selected the more robust versions of the Random Forest and SVM classifiers (with the full feature set and after model tuning) and compared their performance with that of a CRF model. For each feature in position i, its context attributes include i – 2, i – 1, i, i +1, i + 2, i | i – 1, i | i + 1. In order to train the CRF model it was necessary to discretize the numerical features:

Algorithm	Accuracy	Precision	Recall	F1 measure
Random Forest	83.3%	86.3%	69.2%	76.8%
SVM	83.1%	82.8%	72.8%	77.5%
Neural Network	84.2%	83.9%	74.8%	79.1%
Classification Tree	77.0%	71.0%	71.6%	71.3%

TABLE 14.1: Performance of classification algorithms on 49k syllables (19 features) using 5-fold cross-validation

we used 10 equally-spaced bins (note that acoustic features such as pitch are already log-transformed; a z-transformation is applied to the relative syllable duration attributes).

Additional features were included in this second set of experiments. In the CRF model, we added lexical attributes, such as the part-of-speech tag (category, such as adjective or pronoun; and full tag), and the word itself. Hesitations and disfluencies are not excluded from the CRF sequence (they are however marked; thus the model never assigns syllabic prominence annotation to a filled pause). For the SVM classifier, we converted categorical attributes to a vector of binary attributes (e.g. A, B, C is represented by A = (1, 0, 0), B = (0, 1, 0), C = (0, 0, 1)). We also evaluated the performance of a CRF model that included the result of the RF classifier as an additional attribute. The results are presented in Table 14.2.

Algorithm	Precision	Recall	F1 measure
Random Forest	86.3%	70.4%	77.5%
SVM	83.2%	74.1%	78.4%
CRF	83.7%	82.5%	83.1%
CRF + RF	85.4%	84.4%	84.9%

TABLE 14.2: Performance of different systems for the automatic detection of prominence (full set of features, 5-fold cross-validation)

We observe that the CRF model offers an improved F-measure compared to the classifiers. Since series of several non-prominent syllables alternate with prominent ones, the model's ability to capture context relationships is useful. It has to be noted that context relationships are represented in two ways in this model: by the relative features (calculated over different context windows), and by the CRF context attributes themselves.

It may not be practical, however, to include lexical and part-of-speech information in an automatic annotator, since the detection of prosodic prominence may precede or be performed independently of POS tagging. Therefore,

in the next set of experiments, we calculated the improvement in accuracy gained by including such information in a discrete CRF model.

In a third set of experiments, we studied the effects of the size of the training corpus in the performance of different methods. We compared the accuracy (percentage of correctly labelled syllables over all syllables) of three algorithms, for different sizes of the training data-set, and with / without lexical and POS information. Figure 14.1 shows the evolution of the system's accuracy as the training data-set grows from 1,000 syllables to 60,000 syllables. The algorithms compared are: Random Forest based only on acoustic features; discrete CRF based only on acoustic features; and discrete CRF based on acoustic features and the part-of-speech tags (of tokens corresponding to each syllable).

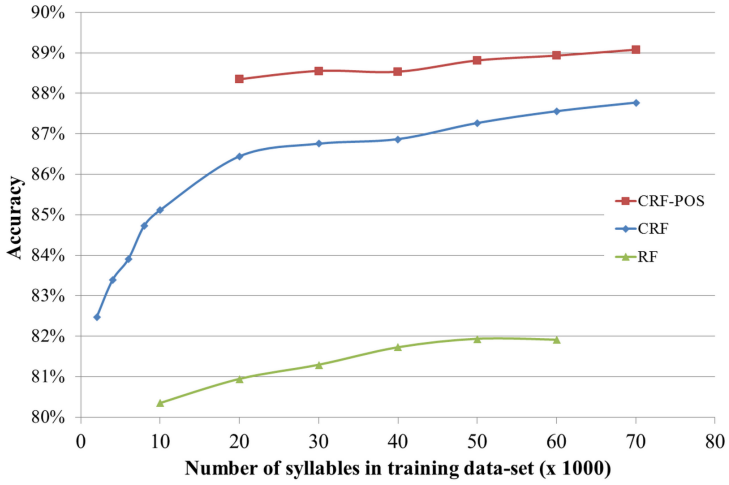


FIGURE 14.1: Effects of the training dataset size on the accuracy of different methods

We note that the CRF models were systematically more accurate than the RF method, when based on the same acoustic features. The CRF models seem to stabilize at a corpus size of approximately 20k syllables; beyond this point, accuracy does improved by training on larger datasets, but at a much slower rate. Adding access to the lexical and part-of-speech information to the CRF model increases its accuracy by 2%. This increase is probably due to the fact that lexical and part-of-speech information is capturing the influence of top-down expectations in the perception of prosodic prominence (cf. Goldman, Auchlin, Roekhaut, Simon, & Avanzi, 2010; Lacheret, Simon, et al., 2013).

Among the classification techniques tested, the Random Forest-based ensemble learning gave best results. Overall, however, the CRF-based sequence labelling approach gave the best results; the combination of classifiers with CRF labellers marginally improved performance. Including lexical and part-

of-speech information may be used to model the top-down expectations of native listeners with regards to prominence; however, it is possible to create a robust detection system based only on acoustic features.

15

PROSODIC BOUNDARIES

In this chapter we summarise our work on the perception of prosodic boundaries in French. Stress, prominence and prosodic boundaries play a central role in defining the prosodic structure and arriving at a phonological description of any language (Mertens, 2014, p. 21). For the purposes of this thesis, we are particularly interested in the annotation of prosodic boundaries, as one of our main hypotheses is that speech under cognitive load will be produced with frequent mismatches between syntactic and prosodic structure. We have performed a corpus-based study on the acoustic and syntactic correlates of prosodic boundaries as annotated by experts in the LOCAS-F corpus. In a perceptual experiment, the expert annotation was compared to the on-line perception prosodic boundaries by naïve listeners, significantly validating the expert annotation of major prosodic boundaries. Finally, we have developed an automatic annotator of prosodic boundaries, which operates in tandem with the Promise annotator of prosodic prominence.

15.1 RELATED PREVIOUS WORK

The prosodic segmentation of an utterance, as expressed by the prosodic boundary cues, is central to discourse comprehension: it has been shown that prosodic boundaries facilitate comprehension, by indicating the intended segmentation to the listener (Cutler, 1997; Frazier et al., 2006; D. Watson & Gibson, 2005); see also Chapter 6 in the present thesis. However the factors contributing to the perception of prosodic segmentation are not completely understood. Phonological theories differ in the number of prosodic segmentation levels (and consequently on the number of prosodic boundary strengths). Consequently there is no consensus on an "objective" method of segmentation of utterances into prosodic units.

Although most models on French prosody admit at least three degrees of prosodic boundaries and a hierarchy of three levels of units (Mertens, 1993; Rossi, 1999; Di Cristo, 1999), most large-scale corpus annotations are limited to 1 or 2 degrees (e.g. in the C-ORAL-ROM corpus; Cresti and Moneglia, 2005). Similarly to the perception of prosodic prominence, there is evidence that listeners perceive prosodic boundaries as a gradual phenomenon and in

relative terms, i.e. they perceive a boundary as stronger or as weaker than the previous one. On the other hand, a functional approach, such as the one adopted by phonological theories, discretises boundary perception and uses a small number of prosodic boundary strengths that define a hierarchy of prosodic units.

M. Wagner and Watson (2010) suggest that silent pauses, duration, fo movement and phonation type are the most salient cues to prosodic boundaries; those cues are known to be language-specific to some extent. In French, since the primary (final) accent is located on the last syllable of a prosodic unit, it co-occurs with the prosodic boundaries (cf. Di Cristo, 2011). However, this does not mean that French listeners cannot distinguish between prominence and prosodic phrasing, as shown experimentally by Astésano, Bertrand, Espesser, and Nguyen (2012).

15.2 CORPUS STUDY: ACOUSTIC AND SYNTACTIC CORRELATES OF PROSODIC BOUNDARIES

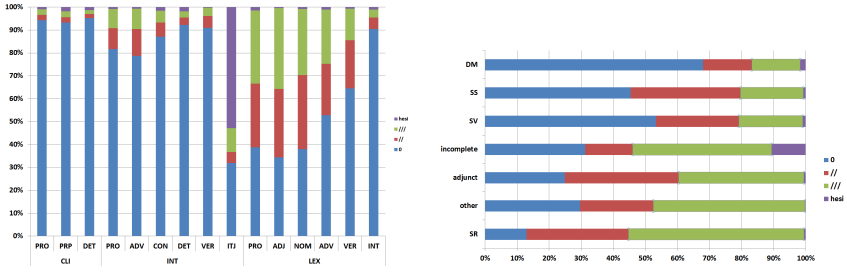
We analysed the properties of the prosodic boundaries annotated as such by experts in the LOCAS-F corpus (this study was presented in Christodoulides & Simon, 2015). We enhanced the available annotations in the corpus by applying DisMo and Prosogram (Mertens, 2004) and storing the information produced by these tools in the Praaline database of the corpus, so that it can be correlated with the expert prosodic annotation.

The prosodic annotation in LOCAS-F was performed by two experience transcribers. Each word was marked as being followed by a strong PB (///), an intermediate PB (//), or as not followed by any boundary (o). The annotators used the code "hesi" to indicate that they perceive the speaker was hesitating: this includes filled pauses (e.g. "euh") and drawls. A function was also attributed to each PB, based on the shape of the corresponding intonation contour. Four types of contours were used: C (continuation), T (final prosody), S (suspense) and F (focus). This annotation was primarily based on the annotators' perception; however they did have visual access to the pitch contour as displayed in Praat. In cases of disagreement, the annotators listened to the relevant section once again and agreed on the final PB and contour label.

On average, there is a prosodic boundary at the end of 27.7% of the tokens in the corpus; 14.5% are of intermediate strength (//) and 13.1% are strong (///), while 2.9% are marked as hesitations (hesi). The distribution of prosodic boundary types presents significant variation across genres: there is a positive correlation between the degree of preparation and the number of strong prosodic boundaries (///), with the exception of radio news bulletins (that have a high speech rate and few pauses). The number of expert-annotated hesitations (hesi) increases in the more spontaneous speaking styles. The number of strong prosodic boundaries exceeds the number of intermedi-

ate prosodic boundaries in three speaking styles (ACA: scientific conference, HOM: sermon and POL: political public address) creating the perception of over-segmented speech flow.

Regarding the syntactic correlates of prosodic boundaries (Figure 15.1a), we observed that prosodic boundaries occur mostly on lexical words (LEX includes adjectives ADJ, nouns NOM, adverbs ADV, verbs VER and clitic pronouns), while less than 3% of PBs would occur on a clitic word (CLI). An intermediate category (INT, including interrogative or relative pronouns PRO, negation particles ADV, conjunctions CON, determiners DET and auxiliary verbs VER) co-occurs with a PB in < 10% of potential positions. The relationship between prosodic boundaries and syntactical sequences is shown in Figure 15.1b. We observe that sequences may be rather long and complex syntactic units. In approximately 50% of the cases, verb sequences (SV) and subject sequences (SS) do not bear a PB, because they are followed by another sequence that is prosodically grouped with them. Conversely, the majority of dependent sequences (SR), Adjuncts and Others (sequences without verb) are followed by strong (///) prosodic boundary. Incomplete sequences are found in-between these two tendencies.



(a) PBs and part-of-speech tags

(b) PBs and syntactic sequences

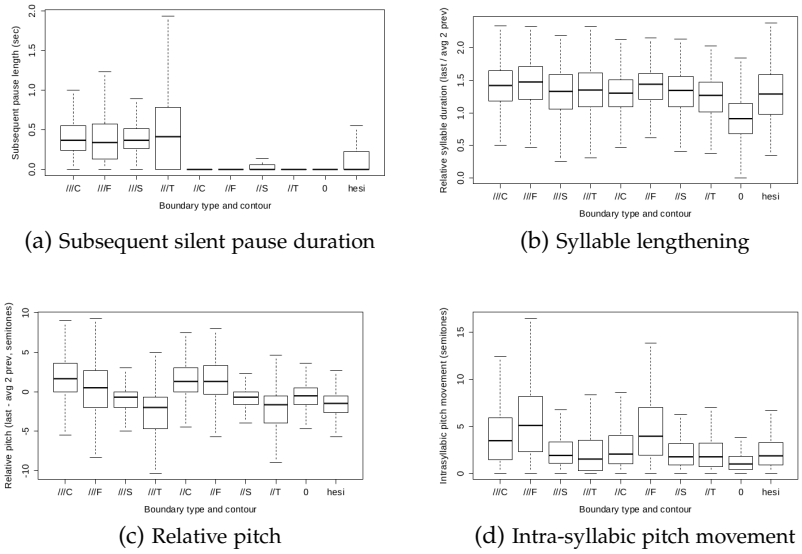
FIGURE 15.1: Relationship between prosodic boundaries and the POS tag of the token on which it occurs (left); the syntactic functional sequence on which it occurs (right) in the LOCAS-F corpus

To establish the acoustic correlates of prosodic boundaries, we calculated the following measures for each potential prosodic boundary site (i.e. for the last syllable of each multi-word unit):

- the duration of a subsequent silent pause, excluding the pauses at turn-taking;
- relative duration: defined as the duration of the last syllable divided by the average duration of the previous 2, 3, 4 or 5 syllables;
- relative pitch: defined as the difference between the pitch (in semitones) of the last syllable and the average pitch of the previous 2, 3, 4 or 5 syllables;

- intra-syllabic pitch movement (in semitones) as calculated by Prosogram

We observed that a local analysis (window of 2 syllables) is sufficient: the results presented here remain valid for windows of 3...5 syllables. The distributions of the four prosodic measures for each prosodic boundary type and associated contour can be seen in Figure 15.2.



(a) Subsequent silent pause duration

(b) Syllable lengthening

(c) Relative pitch

(d) Intra-syllabic pitch movement

FIGURE 15.2: Distribution of the four acoustic measures by prosodic boundary type and contour

We observed that the presence or absence of a silent pause is the main feature that distinguishes between strong (///) and intermediate (//) prosodic boundaries. Strong PBs (///) are almost always followed by a silent pause, while this is rare for intermediate PBs (//); hesitations are occasionally followed by a pause. Syllable lengthening occurs on PBs regardless of their associated contour, and it is more pronounced in cases of Focus. Taking non-PB syllables as the baseline, PBs associated with the C (continuation) and F (focus) show a rising intonation, while T (final) PBs exhibit a falling intonation, followed by S (suspense) PBs, while hesitations are more similar to T (final) PBs. Focus (F) PBs have the most dynamic (pronounced) intra-syllabic pitch movement, followed by C (continuation). Strong continuation (///C) prosodic boundaries are clearly marked with both inter- and intra-syllabic pitch movements, whereas intermediate continuation (//C) prosodic boundaries are only marked with relative pitch differences.

15.3 EXPERIMENTAL STUDY: THE ON-LINE PERCEPTION OF PROSODIC BOUNDARIES BY NAÏVE LISTENERS

In the previous section, the acoustic and syntactic correlates of prosodic boundaries were derived from an expert annotation. In a recent study (Simon & Christodoulides, 2016, 2016, accepted) we investigated the on-line perception of prosodic boundaries by naïve listeners, i.e. listeners who had no prior training in phonology, nor the possibility to listen to the same speech sample twice. Our objective was to test whether the perception of prosodic boundaries by naïve listeners in real time differs from expert annotation, and to what extent, as well as to examine the impact of acoustic and syntactical cues on this perception.

We selected 4 groups of 12 stimuli extracted from the LOCAS-F corpus, with an average duration of 29.9 seconds (min: 5.1, max 39.9); each group was balanced across 4 criteria (articulation rate, silent pause ratio, melodicity, filled pauses to number of syllables ratio). We produced corresponding manipulated speech stimuli, in order to mask lexical content, while retaining the temporal, syllabic and intonation structure. Phonemes were randomly replaced with another phoneme from the same group (plosives, fricatives, nasals, liquids, glides, vowels, nasal vowels), ensuring that resulting diphones exist in French. Phone duration was kept intact, while the intonation contour was approximated (with 10 points defining a pitch target). The manipulated stimuli were then synthesised using the MBROLA TTS system. The resulting stimuli sound similar to a pseudo-language.

Participants listened to the short samples of speech and were instructed to press a key whenever they perceived the end of a "group of words" (this instruction was deliberately vague, in order to avoid biasing subjects towards a syntax-based analysis). Participants could only listen to each sample once and the collection of responses was done in real time, in order to be as close as possible to natural conditions of speech perception and comprehension. The experimental sequence ran as follows: participant identification, working memory capacity test, tonal acuity test, baseline response time test (participants were asked to press the key as soon as they heard a pure tone); training; segmentation of natural stimuli; segmentation of manipulated stimuli; repetition of the baseline response time test. The experiment was presented using OpenSesame (Mathot, Schreij, & Theeuwes, 2012). In total, the responses of 125 participants were analysed; participants were students of the faculty of modern languages or psychology at the Université catholique de Louvain and participated in the experiment for course credit.

The main procedure for analysing the raw data is visualised in Figure 15.3. For each subject, we calculated a mean response time (RT) from their responses to the pure tones. These values were subtracted from their responses in order to centre them with respect to a potential location of a prosodic boundary and to reduce variability induced by individual motor skill differ-

ences. A moving average (window size: 250 ms, step: 20 ms) of the number of responses was calculated and the local maxima of this value were considered as the PPB sites. In order to group subject responses correlated with a PPB, we followed the following algorithm: starting from a local maximum which is considered as the centre of a group of responses, a response is attributed to the group if its distance from the previous response is less than 300 ms, and at the same time its distance from the centre does not exceed 500 ms. We are thus attempting to detect clusters of responses triggered by the same cues. These responses are subsequently treated as a group: each group is a perceived prosodic boundary (PPB). These PPBs were correlated with the nearest final syllable of a token (PPB sites falling within a silent pause were attributed to the previous final syllable).

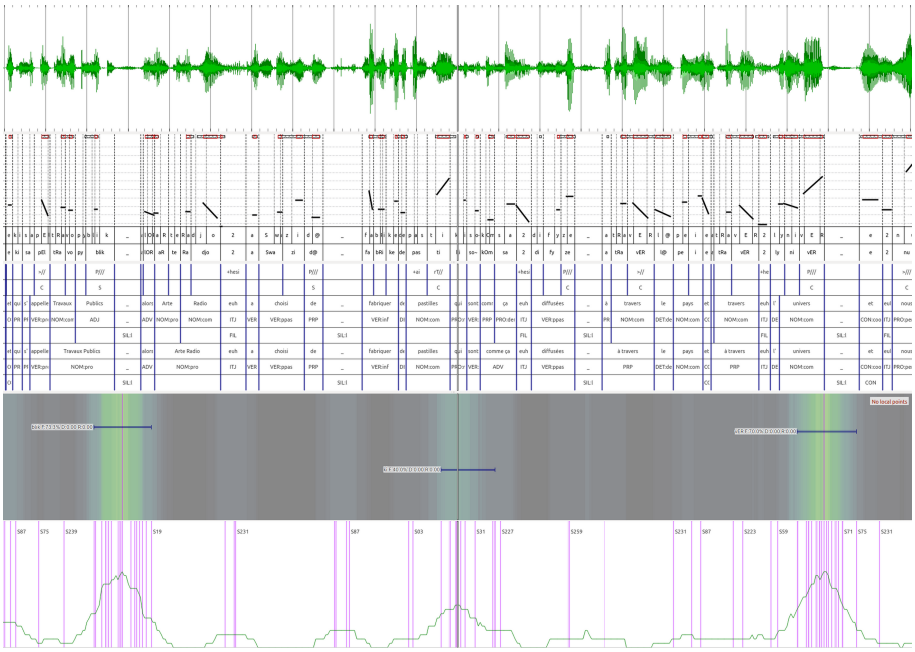


FIGURE 15.3: Visualisation of the results of the experiments on perception of prosodic boundaries by naïve listeners in real time. The waveform (A) is displayed along with its Prosogram (B) and transcription, POS and syntactic annotation (C). Subject responses (centred) are shown in panel E, along with a moving average of the number of responses. Local maxima are selected as the PPB locations; panel D displays the extent of each group of responses considered as part of the same perceived prosodic boundary (PPB).

For each PPB we calculate three measures: the boundary force is the proportion (%) of participants who registered a response at this PPB site; the boundary delay is the arithmetic mean of the temporal difference between the

syllabic nucleus and each subject response; and the boundary dispersion is the standard deviation of the aforementioned temporal differences (response times). In total, 25537 responses were grouped into 1239 PPBs. Each sample was annotated by approximately 30 participants.

The results confirmed our hypothesis that due to the tasks constraints, naïve listeners would perceive a smaller number of prosodic boundaries than those annotated by the experts. In the natural speech condition (NS), 85% (434 out of 508) of PBs annotated as strong (///) by the experts were perceived by the naïve listeners, while only 17% (89 out of 533) PBs annotated as intermediate (//) by the experts were perceived by the naïve listeners in real time. The corresponding figures for the manipulated speech (MS) condition were 93% (470 out of 508) and 9% (49 out of 533). Subjects perceived 61 PPBs in NS and 21 PPBs in MS on syllables where the experts had not annotated a prosodic boundary.

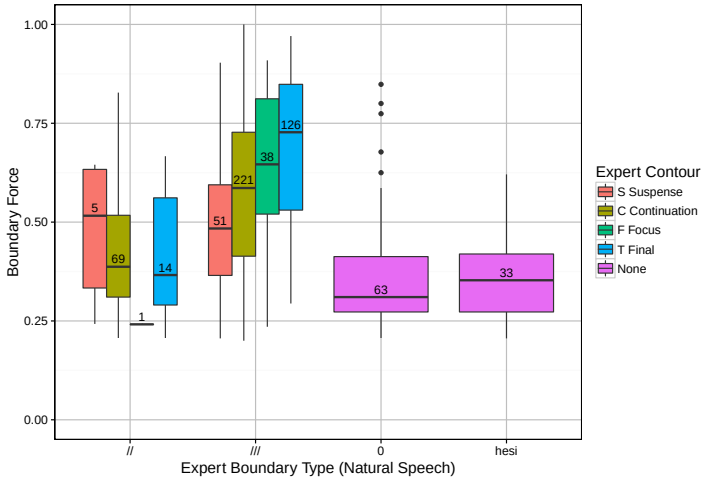


FIGURE 15.4: Comparison of the expert annotation with the perceived boundary force by naïve listeners

The results validated the experts' annotation of prosodic boundary strength. Figure 15.4 shows the distribution of boundary force, for each combination of the experts' annotation of boundary strength and contour. Boundaries annotated as strong (///) by the experts were systematically perceived by more participants (i.e. the boundary force is greater) than boundaries annotated as intermediate (//) by the experts. Focusing on PBs both annotated by experts and perceived by naïve listeners, the mean boundary force for strong PBs is 61% in the NS condition and 62% in the MS condition; whereas the mean boundary force for intermediate PBs is 42% in both conditions (in both cases the difference of means is significant; Hedges g = 0,98 in NS and 1,07 in MS). Examining all PBs annotated by the experts, irrespective of whether these

were perceived by the naïve listeners, we find that, in the NS condition, the mean boundary force of the 508 strong PBs is 52%, and the mean boundary force of the 533 intermediate boundaries is 7% (the difference of means is significant; Cohen δ = 1,94); in the MS condition, the mean boundary force of strong PBs is 57% and the mean boundary force of intermediate PBs is 4% (the difference of means is significant; Cohen δ = 2,74). We also observed that the falling-contour (T) PPBs had the highest mean perceived force; and that level-contour (C) PPBs had a similar distribution of perceived force, regardless of whether the experts annotated them as weak or strong.

An analysis of the acoustic and syntactic correlates of PPBs using linear models and linear regression trees showed that in the NS condition, the most important cue for the perception of a prosodic boundary is the presence of a subsequent silent pause, followed by the strength of a syntactic boundary (end of clause, followed by the end of a sequence), and final lengthening. In the MS condition, the strongest cue was the presence of a subsequent silent pause, followed by final lengthening and pitches movement. A qualitative study of cases where a boundary was perceived in one condition but not in the other showed that syntactic closure may trigger the perception of a prosodic boundary. Conversely, lack of syntactic closure, or the presence of a filled pause, may impede the perception of a prosodic boundary (the results of the qualitative study are detailed in Simon & Christodoulides, 2016, accepted).

15.4 AUTOMATIC ANNOTATION OF PROSODIC BOUNDARIES

Based on the corpus study outlined in section 5.2 and the experimental evidence suggesting that naïve listeners indeed perceive prosodic boundaries in real time at the same sites on which the annotators of the LOCAS-F corpus have indicated the presence of a strong prosodic boundary, we have developed an automatic annotator of prosodic boundary strength, trained on the corpus data. The annotator uses the same algorithm as the Promise system outlined in Chapter 4: the task of boundary annotation is framed as a sequence labelling problem. The statistical model used is based on Conditional Random Fields, and the acoustic properties taken into consideration are the same as those that Promise uses to model prosodic prominence of syllables. Based on the results of a 5-fold cross-validation, the annotator achieves a 92% precision in identifying strong prosodic boundaries and a 68% precision in identifying intermediate prosodic boundaries. In future development of the system, we will explore the addition of chunking information from the DisMo annotator to capture into the statistical model the effects of syntax that were experimentally observed.

16

TEMPORAL MEASURES

In this chapter we describe the main measures that are related to the temporal organisation of speech. In the first section, we review findings on speech pauses, such as their function, how the distribution of silent pause length varies across speaking styles and methodological questions relating to the identification and statistical analysis of silent pause length. In the second section, we review methods on quantifying the perceived speech rate and related measures.

16.1 SPEECH PAUSES

The temporal organisation of speech can be studied by segmenting the speech signal into measurable components, as a sequence of articulated intervals and pauses. Zellner (1994) proposes two classifications of pauses: a physical / linguistic classification and a psychological / psycholinguistic classification.

Based on the linguistic classification, "normal speech flow is considered to be interrupted by a physical pause whenever a brief silence can be observed in the acoustic signal, i.e. a segment with no significant amplitude" (Zellner, 1994, p. 42). These physical pauses are either intra-segmental pauses or inter-lexical pauses. Intra-segmental pauses are caused by articulatory constraints, such as the occlusion of the vocal tract: the articulation of plosives, for example, creates a silence of 50-100 ms. Intra-lexical pauses are those that appear between two words and provide an initial segmentation of speech to facilitate its perception (Cutler, 1997).

A psycholinguistic classification reflects the observation that perceived pauses do not correspond to physical pauses. This is a manifestation of a more general property of human sensory system: the perception threshold is higher than the actual physical stimulus (Zellner, 1994, p. 43). Some pauses are more easily perceived than others, and these pauses seem to have a functional role, such as marking focus (cf. silent pauses as a correlate of prosodic prominence; Chapter 4), marking prosodic boundaries and coinciding with syntactic boundaries (cf. findings in Chapter 5 on prosodic boundaries). Using this approach, pauses include both silent pauses and filled pauses. Silent pauses correspond to a silent interval in the signal and may be produced along

with inspiration, silent expiration, swallowing or a laryngophonatory reflex. Filled pauses correspond to a voiced interval in the signal that is perceived as a pause. Zellner (1994) proposes to include hesitation-related lengthening (drawls), "non-syntactical" repetitions (i.e. without grammatical or emphatic function) and false starts within a broad category of "filled pauses", arguing that all these phenomena affect the time course of speech production.

As we have already described in Chapter 13, we will adopt a more analytical taxonomy of disfluencies, with separate categories for the above-mentioned phenomena. In this thesis, filled pauses are defined as voiced intervals, belonging to a language-specific inventory of vowels or syllables (in the case of French: ə, œ, ø, ɛ̃ and œm; cf. Candea, 2000, p. 24), that can be either produced autonomously or as epenthetic vowels. Autonomous filled pauses are either autonomous fillers (i.e. surrounded by silent pauses), or fillers perceived to be articulated independently of the previous and following word, while epenthetic vowels are attached to the end of the previous word (co-articulation).

Goldman-Eisler (1968, 1972) was the first to posit that pauses are a reflection of cognitive activity. Her hypothesis predicts the commonly observed differences in the pause patterns between spontaneous and read speech (more pauses are produced in spontaneous speech). F. Grosjean and Lane (1976) have shown that listeners integrate pauses in their perception of speaking rate. Krivokapić (2007) studied the effects of effects of prosodic structure and phrase length on pause duration, and found that more complex syntactic structures lead to longer pauses, both pre- and post- boundary. Longer pauses around topic changes than in the middle of sentences (Goldman-Eisler, 1968) but these longer pauses disappear when they have time to plan for topic changes in advance (Greene & Lindsey, 1989). F. Grosjean and Collins (1979) analysed the breathing patterns of speakers reading a text at different speaking rates and found that the duration and frequency of breathing pauses depend on speech rate and their syntactic position. Non-breathing pauses follow the same pattern as breathing pauses, but are always shorter and tend to occur primarily at minor constituent breaks. As the speech rate increases, shorter pauses disappear and only breathing pauses persist at the major breaks.

16.2 METHODOLOGICAL CHALLENGES IN ANALYSING SPEECH PAUSES

The statistical analysis of silent and filled pause length poses a number of methodological challenges, as the typical distribution of pause durations is positively skewed; therefore the use of methods that rest upon hypotheses of normality is not appropriate (e.g. comparing the arithmetic means of silent pause durations in two conditions by ANOVA; cf. Oehmen, Kirsner, and Fay (2010)).

An alternative method is to study the distribution of the logarithm of pause durations. In log-time, pause durations form two normal distributions.

Kirsner, Dunn, and Hird (2005) hypothesis that the first component distribution (short pauses) correspond to articulatory processes and the second component (medium-length pauses) corresponds to cognitive processes, including discourse segmentation. However, this method must be applied with caution, after establishing that the original pause length distribution is indeed bimodal: Heldner and Edlund (2010, pp. 561-562) describe an analysis of gaps (inter-speaker pauses in dialogue) in which "bimodality was clearly an artifact of the [logarithmic] transformation".

Using an arbitrary minimum threshold value for pause length can also lead to erroneous conclusions. The lack of interest in articulatory pauses, combined with poor recording conditions, has led researchers to apply a minimum threshold to pause duration and discard pauses shorter than this threshold. The most widely-used silent pause threshold is 250 ms, first proposed by Goldman-Eisler (1968). F. Grosjean and Deschamps (1975) used a threshold of 300 ms, Candea (2000) used a threshold of 200 ms, and Duez (2001) suggests a threshold between 180 ms and 250 ms. Some studies set a maximum threshold over which a pause is characterised as a silence instead. However, Oehmen et al. (2010, p. 1) remark that "numerous pauses between 130 ms and 250 ms have been shown to have both cognitive and expressive functions". Campione and Véronis (2002) demonstrate the problems arising when using such thresholds. They compare the average durations in their read and spontaneous French corpora, using no threshold, using both a low and high threshold (200 ms and 2000 ms respectively), and a low threshold only (200 ms). They show that "although the average durations are higher in read speech, we would wrongly conclude that they are about equal if we used both thresholds, and even that pauses are longer in spontaneous speech if we used only a low threshold" (p. 4). They remark that "speakers, speech genres, etc. tend precisely to be opposed by the distribution of pauses in the extremes, which are cut off by the thresholds".

Oehmen et al. (2010) studied the reliability of human judgement in identifying pauses using multiple segmentations of four speech files by four analysts. They calculated inter-annotator agreement for short and long pauses, and variation in inter-annotator agreement. They show that while intra-annotator reliability is generally high, inter-annotator reliability is moderate and inversely proportional to signal to noise ratio (SNR). This shows that each analyst applies his own methodology for pause segmentation consistently; however poor recording conditions (low SNR) lead to significant differences in manual pause segmentation. A more detailed analysis of the disagreements revealed that there are greater differences in the placing of boundaries before long pauses that before short pauses, and that "disagreements are frequently caused by misclassification of non-speech artefacts such as a preparatory articulation and audible breathing in speech" (p. 268). It has to be noted that an incorrect segmentation of pauses influences not only the prosodic measures directly related to pauses, but also other any other measure indirectly linked

(e.g. articulation time, articulation to speech ratio, and measures calculated over articulation time).

Modelling pause duration distribution Campione and Véronis (2002) analysed five hours of read and spontaneous speech in five languages. They report a tri-modal distribution of pauses lengths, categorizing them in brief (less than 200 ms), medium (200 to 1000 ms) and long (over 1000 ms). They only found long (>1s) pauses in spontaneous speech. Most importantly, they report that pauses follow a log-normal distribution globally and for each category. Demol, Verhelst, and Verhoeve (2007) analysed a four-hour corpus of three different speaking styles and six European languages. They confirmed that the "logarithmic duration of the pauses can be well approximated by a bi-Gaussian distribution" both in slow and in fast speaking rates. Similar pausing strategies were found for all languages (Dutch, English, French, Italian, Romanian and Spanish). With a slow speaking rate, speakers pause more frequently and use a wider range of pause durations. With a fast speaking rate, speakers pause less frequently and refrain from using the longest pauses occurring in their normal speech. Goldman, Auchlin, Roekhaut, et al. (2010) studied a 40-minute French spoken corpus with 4 speaking styles (reading, narration, broadcast news and university lectures). They report a multimodal log-normal distribution and propose a mixture of log-normal distributions.

Silent pause length can be modelled as a mixture of log-normal distributions:

$$f(x) = \sum_{i=1}^{N} \pi_i \Lambda_i \left(\mu_i, \sigma_i^2, x\right)$$

where each component distribution Λ_i is Gaussian with mean μ_i and standard deviation σ_i. Its weight in the mixture model is π_i and silent pause durations are log-transformed. We can identify whether two or three component distributions better model the observed silent pause lengths by using the Bayesian Information Criterion. The parameters of these log-normal distributions are calculated using the Expectation-Maximisation algorithm. This methodology was used by Little, Oehmen, Dunn, Hird, and Kirsner (2013) and Goldman, Auchlin, Roekhaut, et al. (2010) and will be adopted for the present thesis.

16.3 SPEECH RATE

Listeners have an intuition about whether someone is speaking "slow" or "fast". Nevertheless, the quantification of speech rate in a perceptually - informed way that can be applied across languages is an open research question. A first approach to measuring speech rate is to count units of a specific type over some period of time. Early studies in prosody use words per minute, usually averaged over an utterance. Syllables per time unit (minute or second) and phones per time unit are also standard speech rate measures.

These measures pose three methodological questions: Which is the most appropriate unit: the word, the syllable, or the phone? Over which time span should the rate be calculated? The possibilities include inter-pause intervals (runs), rhythmic groups, intonation groups, or syntax-based units such as phrases. And how can pauses be integrated in such a model? Moreover, these methods of estimating speech rate provide an average value that cannot represent fluctuations within a given time span, e.g. accelerations and decelerations within an inter-pause unit.

F. Grosjean and Deschamps (1972, 1975), F. Grosjean, Grosjean, and Lane (1979), in an effort to compare French and English, define the following measures:

Speaking time = Articulation time + Silent pause time [sec]
Speaking rate = Number of syllables / Speaking time [syllables / sec]
Articulation rate = Number of syllables / Articulation time [syllables / sec]
AT/ST ratio = Articulation time / Speaking time [%]

They conclude that the temporal organisation of speech can be described with three primary variables, namely the articulation rate (in syllables per second), the number of pauses (or alternatively, the length of inter-pause units) and the average pause length; and several secondary variables, namely the number and duration of non-silent pauses, repetitions and false starts. These basic measures are still in common use and we will adopt them in this thesis.

However, these measures have inherent shortcomings, and it is be desirable to create a better model for speech rate estimation. Firstly, measuring articulation rate in syllables per second ignores the variation of syllable length due to different syllabic structures (that vary greatly across languages). Secondly, as we saw in the previous section, the distribution of pause durations cannot be adequately described by an arithmetic mean and that using a mixture of two Gaussian distributions is more appropriate. And finally, the perceived global speech rate is a function of articulation rate and the distribution of pauses F. Grosjean and Lane (1976).

Pfitzinger (1996) compared two automatic methods for speech rate estimation: syllable detection and phone segmentation. His study is motivated by the need to improve automatic speech recognition, for which "speech rate is one of the most important prosodic cues, because it modifies acoustic cues (e.g. transitions), phones and even words" (Crystal (1990), in Pfitzinger (1996, p. 421)). He measured speech rate for read sentences that did not contain pauses, using a smoothing method (p. 423) and compared the automatically estimated speech rate with rates calculated using manual segmentation of phones and syllables. He found that the phone-based and the syllable-based speech rates have a "correlation coefficient r = 0.711 [which] is quite high, but it shows clearly that the information content of both methods is not identical [...] consequently it is not allowed to describe neither the rate of phones nor the rate of syllables".

Based on these observations, Pfitzinger (1998) proposed that speech rate be estimated using a linear combination of syllable rate and phone rate. He conducted a perception experiment, in which listeners were asked to arrange stimuli (625 ms each, amplitude-equalised) on a computer screen. Participants were asked to place 144 stimuli along a rate scale according to the speech rate and to finally check all labels for the correct order (p. 1088). Three stimuli were pre-anchored as representing slow, normal and fast speech rate, to guarantee comparability. The perceptual local speech rate is estimated as a linear combination of phone rate and syllable rate and the weights of each are calculated by the perception results. The results show that the perceptual local speech rate correlates better with local syllable rate than with local phone rate ($r = 0.81 > r = 0.73$) and that the linear combination of both is well-correlated with perceptual local speech rate ($r = 0.88$). Therefore a better estimate of perceived local speech rate is provided by combining phone and syllable rate. Finally, all subjects appear to have a homogeneous intuition on how to assess speech rate, as the correlation coefficient between the perception results of the two participants whose values least correlate is $r = 0.96$ (p. 1090).

Zellner (1998) also argues that the characterization of speech rate with only one parameter is unsatisfactory (p. 3159) because it has been shown that changes in rate produce effects at the prosodic (Fougeron & Jun, 1998, e.g.), syllabic and articulatory levels. Analysing a corpus of 50 sentences read in French, she suggests that speakers have a number of strategies available for slowing down their speech. While F. Grosjean and Deschamps (1975), F. Grosjean, Grosjean, and Lane (1979), Barbosa (1994) and others reported that the main mechanism to slow down speech is the production of silent pauses, Zellner's results do not confirm this hypothesis. She observes, instead, that pauses already produced in normal rate are lengthened in slow rate and that the new pauses appear inside longer prosodic groups. Slow speech and fast speech have different pause patterns (p. 3160), but pauses only account for less than 8% of the time difference. The main mechanism used for slowing down speech is the lengthening of segments, accounting for almost 70% of the difference. Not all segments have the same temporal elasticity; for example, vowels and fricatives can be lengthened much more than plosives. Zellner (1998) categorises phonemes into durational classes based on their normalised average duration (based on her corpus).

Two recent studies take into account the different intrinsic properties of segments: of syllables in Obin, Rodet, and Lacheret (2008) and of phonemes in De Looze (2010). The objective of Obin et al. (2008) is to build a continuous speech rate model by smoothing inverse syllable durations. They argue that the duration of a syllable is a function of (a) its intrinsic properties, reflecting its phonological representation and (b) adjustments caused by local speech rate variations (p. 2). Each syllable by a specific speaker in the corpus is attributed a structure class based on its phones (e.g. CV, CVC, where C =

consonant and V = vowel). The mean and standard deviation of all these syllable durations are then calculated (μ_{ref} and σ_{ref}), as well as the averages and standard deviations of the syllable durations for each specific structure class (μ_c and σ_c). The normalised syllable duration is then defined as a function of its observed duration d_{obs} as follows:

$$d_{norm} = (d_{obs} - \mu_c) \frac{\sigma_{ref}}{\sigma_c} + \mu_{ref}$$

A measure called normalised speech rate is calculated by applying a smoothing technique over the normalized syllable durations within two silent pauses. Obin et al. (2008) propose to use a Hamming window for smoothing.

De Looze (2010) argues that articulation rate should be measured by calculating the difference between the observed and expected phone durations. The expected phone duration is defined for each phoneme as the mean duration of all its observed realisations (phones) by a given speaker (p. 172). She proposes an algorithm for automatically detecting speech rate variations (implemented in a set of scripts called ADoTeVa) that calculates these differences and then applies hierarchical clustering to group regions of similar speech rate. The resulting clustering indicates where significant variations of speech rate have taken place.

16.4 CORPUS STUDY: EFFECTS OF SPEAKING STYLE ON SILENT PAUSE LENGTH

In collaboration with Mathieu Avanzi, Damien Lolive, Elisabeth Delais-Roussarie and Nelly Barbot we conducted a study on the effects of different speaking styles on a number of prosodic parameters, such as articulation rate, silent pause length and prosodic phrasing. The results are presented in Avanzi, Christodoulides, Lolive, and Delais-Roussarie, Elisabeth , Nelly Barbot (2014). In this section we will summarise the findings on the effects of speaking style on silent pause length distribution. The study was conducted for a specific application: refining the design of pre-processing prosodic modules in a text-to-speech system, in order to improve the expressivity of synthesized speech.

For the purposes of the study, a corpus containing speech from four speaking styles was constituted. All speaking styles are read speech "overtly addressed to a given audience": reading of fairy tales (TAL), dictations (DIC), political speeches (POL), and reading of novels (NOV). The four speaking styles differ as to the nature of the audience (adults for POL and NOV, vs. children for DIC and TAL) and the desired impact (importance of being understood and convincing for POL and DIC; less important for NOV and TAL). 30 minutes of speech per style are analyzed. Table 16.1 details the number of speakers and the exact duration of the samples in our corpus.

Silent pause length was modelled as a mixture of log-normal distributions, as described in section 6.2:

Speaking Style	Nb. of speakers	Nb. syll.	Nb. tokens	Duration (sec.)
Tales (TAL)	6F/2M	5942	4189	1065.25
Dictation (DIC)	2F	4175	2918	893.56
Political (POL)	3F/3M	6875	4539	1362.02
Novel (NOV)	2F/2M	7496	5226	1286.97
Total	13F/7M	24488	16872	4607.81

TABLE 16.1: Corpus Composition for the study of speaking style effects on silent pause length (from Avanzi et al., 2014)

$$f(x) = \sum_{i=1}^{N} \pi_i \Lambda_i \left(\mu_i, \sigma_i^2, x\right)$$

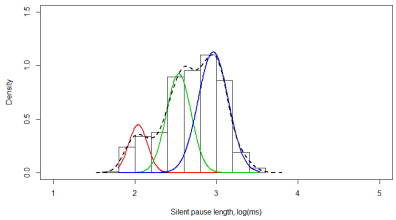
We identified whether two or three component distributions better model the observed silent pause lengths by using the Bayesian Information Criterion. After selecting the number of component distributions, their parameters are estimated using the Expectation-Maximization algorithm. The results of analyzing silent pause duration using this method can be seen in Table 16 and the following figure:

	DIC			POL			TAL		NOV	
π	13%	35%	52%	30%	39%	31%	9%	91%	12%	88%
μ (ms)	109	345	911	134	421	1024	88	555	107	605
σ (ms)	1.3	1.4	1.5	1.5	1.4	1.4	1.3	1.9	1.5	1.8

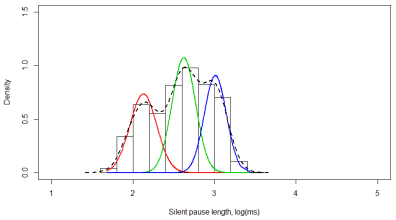
TABLE 16.2: Log-normal mixture model of silent pause length for the four speaking styles

We observe that TAL and NOV are bi-modal, whereas DIC and POL are tri-modal. We hypothesize that the long pause component distribution in DIC are the pauses the speaker makes to allow for writing time (a very specific characteristic of the dictation speaking style) and that in POL the long pauses component distribution is mainly connected to rhetorical style.

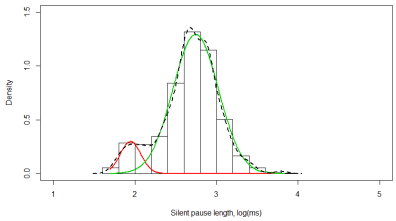
In this chapter we presented the methodological aspects of quantifying and analysing the temporal characteristics of speech. For the analysis of the experimental results collected in this thesis, we have will use the methodology developed in sections 6.2 (and applied in a corpus study in section 6.4) for silent pause, and a combination of different methods for speech and articulation rate.



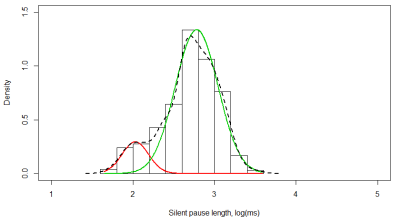
(a) Dictation



(b) Political address



(c) Fairy tales



(d) Reading of novels

FIGURE 16.1: Silent pause length distribution (log-transformed) in 4 different genres (Corpus from Avanzi, Christodoulides, Lolive, and Delais-Roussarie, Elisabeth , Nelly Barbot (2014))

17

PRAALINE: A NEW TOOL FOR SPOKEN CORPUS LINGUISTICS

In this chapter we present Praaline, an open-source software system for managing, annotating, analysing and visualising speech corpora that we developed over the course the present thesis. It attempts to address the needs of researchers working with speech corpora, who are often faced with multiple tools and formats and need to work with increasing amounts of data in a collaborative way. Praaline is based on and extends Praat (Boersma & Weenink, 2016) and Sonic Visualiser (Cannam, Landone, & Sandler, 2010), and interfaces with the R statistical language (R Core Team, 2016).

Instead of creating a system from scratch, we have chosen to focus on the integration of open-source tools that are already widely-used in the community. As a result, researchers using Praaline can benefit from the features, extensions and contributions to these tools. This design allows for the reuse of many existing tools and scripts providing automated annotation and analyses.

Praaline is written in C++ using the Qt framework (version 5, open source) for both its core functions and user interface. It is cross-platform software that runs under Windows, Linux and Mac. Recordings and corpus annotations may reside in a file system and are managed through a relational SQL database: it is possible to use SQLite for local installations or MySQL for client-server access. Praaline can import and export annotations in different formats, including Praat TextGrids, TranscriberAG (Barras, Geoffrois, Wu, & Liberman, 1998), ELAN (Brugman & Russel, 2004) and EXMARaLDA Partitur (Schmidt & Wörner, 2009) files.

An integrated user interface permits the management of corpus data and metadata, annotation using editors or automated procedures, visualisation for the purposes of data exploration or demonstration, and querying the data for analysis. Praaline can be extended with plug-ins written in C++ or Python, and also supports executing scripts written in the Praat scripting language or for the R system against the corpus data and annotations.

17.1 CORPUS MANAGEMENT

Users may construct their corpora from scratch or by importing a set of existing files (e.g. recordings and annotations). A corpus is organised in Communications (communicative situations) and Speakers. Each Communication may consist of several Recordings and Annotations. Speakers participate in Communications with specific Roles; and Annotations contain Annotation Levels and Attributes (see next section). The corpus management module of Praaline contains the main Corpus Editor, the tabular version of the Corpus Editor, and the Corpus Structure Editor. Using the Corpus Editor, the user may add, remove and modify items (Communications, Speakers, Recordings and Annotations) in the corpus. The Corpus Editor is shown in Figure 17.1: the left pane shows the corpus items organised as a tree. The user may choose to group them in a hierarchy based on the items metadata (for example, in a corpus containing different speaking styles and age groups, the user may opt to show the tree by speaking style first, then by age group). The middle and right pane are metadata editors: when a Communication is selected, the middle pane shows all metadata related to it and its associated Recordings and Annotations, and the right pane all the metadata associated with the Speakers participating in the Communication; conversely, when a Speaker is selected, the primary metadata editor (middle pane) shows the fields relating to the speaker, and the secondary metadata editor (right pane) all the Communications in which this speaker is participating. It is also possible to edit metadata in tabular form (by corpus item type).

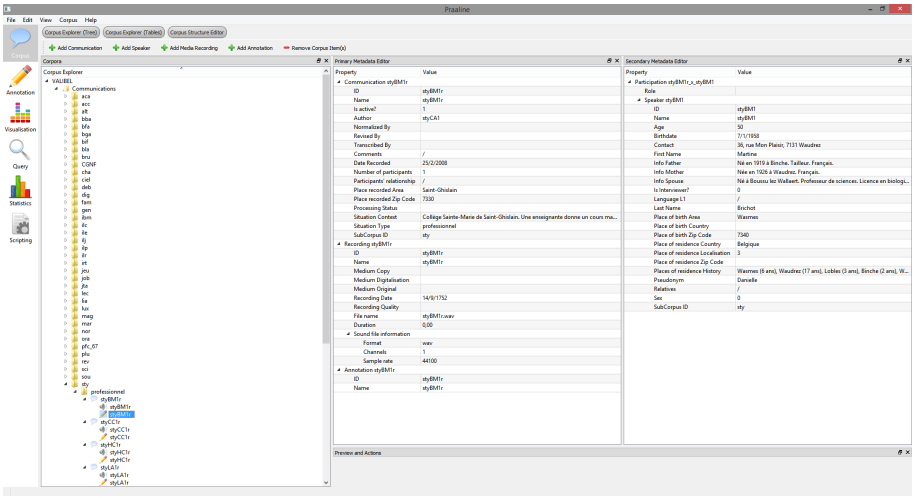


FIGURE 17.1: Praaline Corpus Editor (treeview representation of the corpus contents and metadata editors)

Unlike most other available systems, the fields available for metadata are not fixed in Praaline. The user may define a metadata structure, using an unlimited number of fields, organized in groups, for each type of corpus item (e.g. a set of fields for Communications, a set of fields for Recordings, a set of fields for Speakers, etc.). The definition of this metadata structure, along with the definition of the annotation structure (which will be explained in the next section) is performed through the Corpus Structure Editor, shown in Figure 17.2.

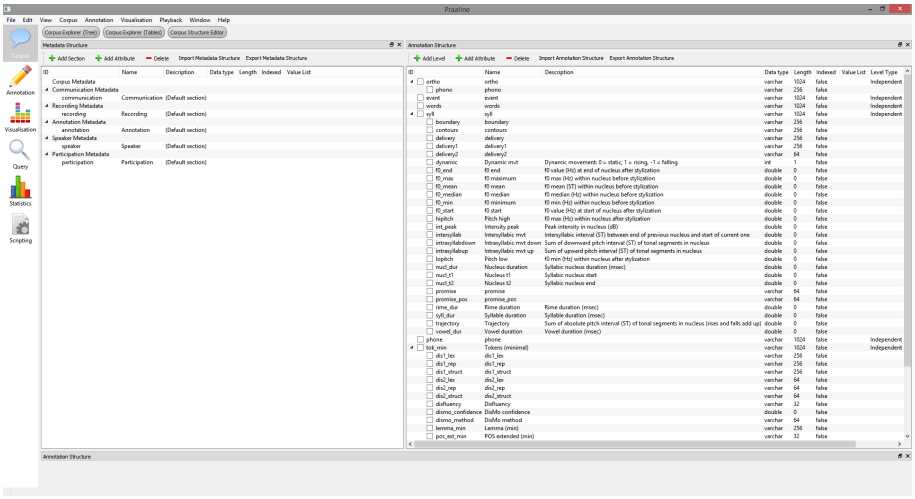


FIGURE 17.2: Praaline's Corpus Structure Editor

Praaline provides standard templates to facilitate the definition of metadata structure and annotation structure when a new corpus is created. It is possible to import data from Exmaralda and Anvil (Kipp, 2001). It is also possible to import metadata that are stored in tabular files or Excel spreadsheets, using a wizard that interactively establishes the correspondence between the file's columns and the corresponding metadata field in Praaline. By storing the corpus metadata and annotations in a network database server, and corpus media files in a network-accessible location, it is possible to set up collaborative projects.

17.2 ANNOTATION

While following the classical timeline model for annotating speech data, Praaline offers a key improvement: the option to define structural links between the annotation tiers, such as hierarchy, containment, attachment, controlled vocabularies etc. The annotation includes knowledge about the relationships between the different layers. An Annotation Level may contain any number of attributes, and relationships between Annotation Levels are encoded as part of

the corpus structure. Praaline does not impose any specific set of Annotation Levels or attributes (sample sets are included to help users). Annotation attributes have associated data types, and may be marked as optional. The annotation structure is defined as part of the corpus definition (see previous section) and standard templates are provided (e.g. a template for a speech corpus that will be transcribed, aligned and analysed with DisMo, Prosogram and Promise).

Annotations may be imported, entered manually using an editor, or obtained automatically by applying annotation plug-ins. Praaline allows the user to apply a cascade of annotation plug-ins on the entire corpus, or on subsets of it. Heterogeneous annotation utilities can be applied sequentially on the corpus. For example, a compiled plug-in for feature extraction, may be followed by a Praat script for prosodic annotation, and then by a POS tagger and an NLP parser in Python, while finally an R script is used to perform a statistical analysis. Praaline handles the data conversions needed to allow such combinations.

The Annotation Level and Annotation Attribute can represent timeline annotations, and also ensure integrity of the data. It is often the case that congruent annotation tiers are used (e.g. in Praat) to represent multiple features of the same object (e.g. a syllable, and an indication of whether it was perceived as prominent or not, or whether it is disfluent). While practical for small amounts of data, this system quickly leads to problems when corpora get larger: e.g. discrepancies in tier boundaries that should had been aligned according to the model; or data incoherence between tiers that are supposedly linked (e.g. phones-syllables). Since these relationships are explicitly captured explicitly in Praaline, it is possible to check the data integrity of a set of corpus annotations, possibly correcting them automatically.

A spreadsheet-like editor allows the user to simultaneously edit attributes belonging to several different Annotation Levels. For example, in a corpus for prosodic studies, we may discern at least three levels: phones, syllables and tokens (words). Each level may have a number of associated Attributes (e.g. syllables may be described by several automatically extracted prosodic features). The editor allows the user to view and update a selection of attributes from each level; it is essentially a timeline display, synchronised with the sound signal. It is also possible to define bookmarks in the corpus, move directly to these points and add a descriptive note; XML-based files of bookmarks can be exchanged between researchers for quick collaboration. The tabular annotation editor can be seen in Figure 17.3; its orientation is customisable: the user may select the vertical orientation (i.e. the timeline progressing from top to bottom, with annotation attributes shown in columns) or the horizontal orientation (i.e. the timeline progressing from left to right, with annotation attributes shown as rows); it is trivial to switch between the two orientations. Speech produced by different speakers is displayed on the same timeline and colours are used to encode the speaker; in this way Praaline

facilitates the annotation of multi-party dialogues: the associated recordings may be separate per speaker, or combined.

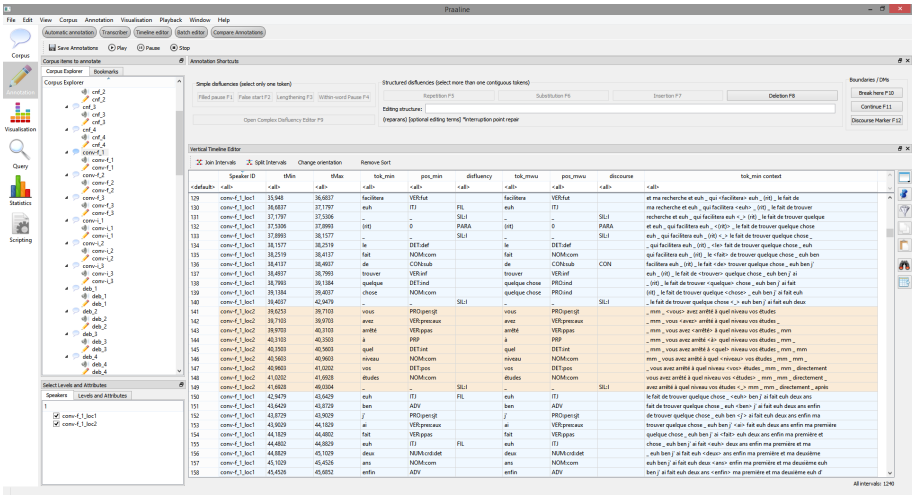


FIGURE 17.3: The tabular annotation editor in Praaline (in vertical orientation mode)

Annotation editor plugins may provide additional functionality: as an example, in the above figure, a small plugin can help the user apply the disfluency annotation protocol outlined in Chapter 3 (by selecting a series of tokens and the desired disfluency category, the plugin adds the appropriate codes). Praaline also includes a Batch Editor, which allows the user to find all distinct tags used in a specified annotation level/attribute combination and modify them in one operation; this feature is handy to correct annotation errors resulting from typos that inevitably arise as a corpus gets bigger. Furthermore, Praaline provides a way to compare annotations, visualise the differences and reconcile them; this can be useful in cases where different annotators have worked on the same data independently, and later need to merge their annotations. The Annotation Level – Annotation Attribute data model translates directly into a relational SQL database. Each Annotation Level is a table, and Annotation Attributes are columns. Praaline uses this system to provide querying functionality (see below). In addition to the possibilities offered by the user interface and through scripting, an advanced user can directly query Praaline's SQL database. It is important to note that the schema is dynamic and adapted to each corpus definition (with the exception of system tables that are always present). Finally, it is envisaged that this corpus metadata and annotation database can be linked to web interface to provide outside users with limited access to the corpus.

17.3 DATA VISUALISATION

The visualisation module of Praaline is based on Sonic Visualiser (Cannam et al., 2010). Visualisations can display waveforms, spectrograms, melodic spectrograms, any combination of annotation levels and tiers, numerical data (points, curves, histograms, colour-coded regions etc.). Plug-ins may add visualisations: for example, we have adapted Prosogram (Mertens, 2004) do display prosodic analysis information. These elements can be combined to present annotations in a format appropriate for each type of investigation. For example, a dialogue involving multiple speakers can be visualised in speaker turns.

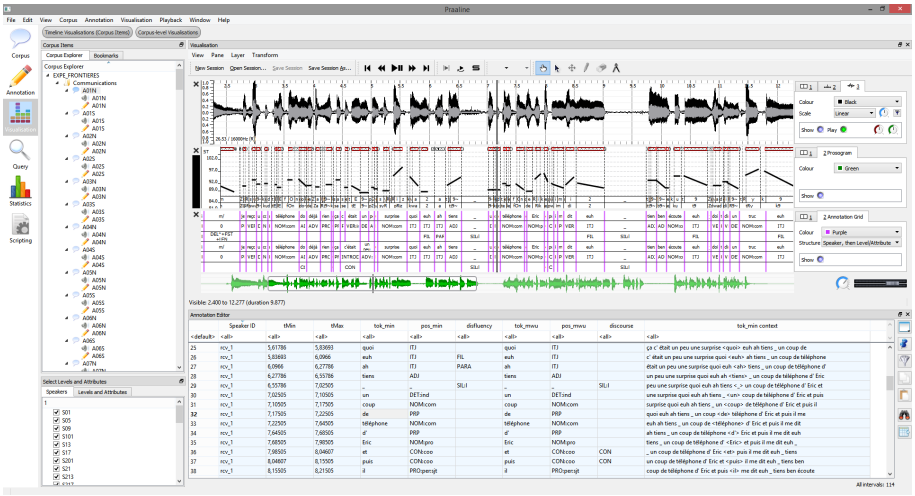


FIGURE 17.4: Simultaneous display of a visualization and a tabular annotation editor in Praaline

In Figure 17.4, a data visualisation is displayed simultaneously with a tabular annotation editor. The visualisation shows the waveform, the corresponding Prosogram, and multiple annotations: the phones level, the syllable level, the minimal token level with the corresponding part-of-speech annotation, and the multi-word unit level. The annotation editor and the visualisation remain synchronised; the user may start playing the sound and annotate at the same time.

By combining Panes and Layers, the user may create simple or complex visualisations, corresponding to the research question at hand and the level of detail required. In Figure 17.5, the visualisation combines a waveform, a spectrogram, a Prosogram, a pane displaying annotation levels (the DisMo annotation) and a pane with two layers: a colour-coded segments layer and a time regions layer (in this example, displaying the results of the experiment described in chapter 15).

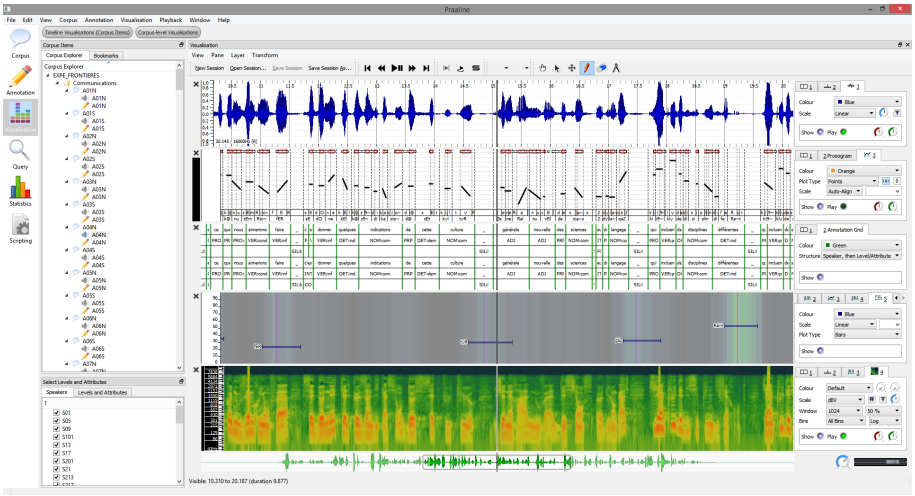


FIGURE 17.5: Praaline visualization with a spectrogram and additional panes

17.4 QUERIES AND CONCORDANCES

Corpus annotations are stored in a relational database, the schema of which is dynamically constructed based on the annotation structure definition. Annotation levels correspond to database tables and annotation tiers to columns. The relationships between the different levels are also encoded. Praaline simplifies the conversion of structured annotation into two-dimensional tables suitable for statistical analysis. SQL queries can be used to select and summarise a subset of the corpus.

It is possible to construct a Dataset by an interactive query editor. For each attribute, functions can be applied to calculate aggregate measures (e.g. sum, mean, standard deviation, etc.); a filter can be used to limit the returned values; and a normalisation transformation (e.g. z-score over all samples of the same sub-corpus) may be applied. In this way, researchers may more easily explore and analyse the information in the corpus, in an interactive way and without necessarily resorting to scripts.

Furthermore, concordances can be extracted using Praaline, based on simple value filters or regular expressions (search term, left and right context). The results of such queries are objects that can be further processed, using the statistical analysis module, or exported for use with other software. The concordance builds and executes the necessary series of SQL queries dynamically, based on user input. A prototype version of the concordance was presented in Barreca and Christodoulides (2014). The results of a simple concordance are shown in Figure 17.6.

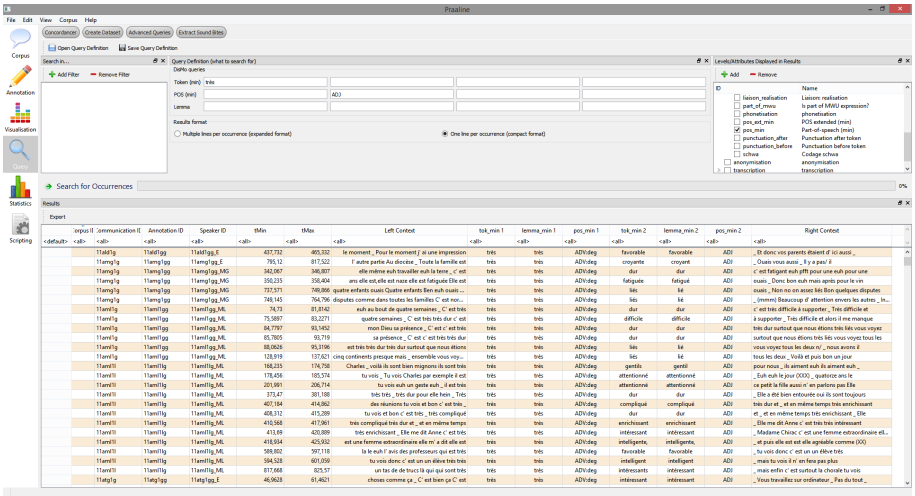


FIGURE 17.6: Praaline's concordancer

17.5 STATISTICAL ANALYSIS AND EXTENSIBILITY

Praaline interfaces with the R statistical environment, through the Rcpp package (Eddelbuettel & François, 2011). Corpus annotations (as well as the results of corpus queries) can be exposed to R as data frames, allowing for the use of R commands, scripts and extensions to analyse the data. Praaline provides a two-way link between the corpus and R: the results of analyses performed using R can be posted back into the corpus database, by adding to or updating an existing annotation level, or by creating a new annotation level.

Praaline can be extended with plug-ins, written in C++ using a simple Application Programming Interface (API) and compiled. This method is suitable for plug-ins adding substantial functionality to the system. Praaline is also scriptable with Python, by providing bindings to its core functionality. Praat scripts can be executed, in which case the corpus data are available as Praat objects.

Currently available plug-ins include the following: an adapted version of Prosogram; an automatic rule-based syllabifier, based on the increasing sonority principle (a list of allowed syllable onsets for each language is required); a plug-in version of the DisMo morphosyntactic annotator; a plug-in version of the Promise annotator for prosodic prominence and boundaries; a plug-in for calculating similarity and convergence measures in dialogue, based on the methodology of De Looze and Rauzy (2011).

Praaline is currently under active development, and is made available to the research community under the GPL licence.

Part III

STUDIES

18

STUDY 1: THE COGNITIVE LOAD SPEECH WITH EGG AND EYE-TRACKING DATABASE

In this study, we elicited speech from participants performing a series of simple tasks, namely a Stroop naming task and a Reading Span task. The experimental design was inspired from Yap (2012) and our study is essentially a replication for French. The "Cognitive Load Speech and EGG" (CSLE) database, described in Yap (2012), is a collection of synchronised recordings of speech and electroglottographic data, collected while subjects were performing a Stroop test and a Reading Span task in English. We have closely followed the experimental protocol of the CSLE, and the tasks have been adapted to French; furthermore, our database contains additional eye-tracking data. In the following sections we present the design of the "Cognitive Load Speech, EGG and Eye-Tracking Data for French" (CLSE^2-FR) database. The objective of this study is to investigate whether there are changes in vocal fold vibration patterns, detectable through EGG measures, as a result of increasing difficulty in performing these standardised cognitive tasks.

18.1 EXPERIMENTAL DESIGN

The experimental sequence comprised of the following four phases: reading a story; Stroop test with time pressure; reading span test; and Stroop test with a dual task load. Each participant sat through all phases in one session of approximately 50 minutes. The experiments were presented using OpenSesame (Mathot et al., 2012).

In the Story Reading phase, the participant was asked to read two texts, typically used in studies in French phonology. The first text is titled "La bise et le soleil", a 120-word simple text used in phonetic studies. The second text is the 300-word fictitious newspaper article "Le premier-ministre ira-t-il à Beaulieu" ("Will the prime-minister visit Beaulieu?"), created by the project Phonologie du Français Contemporain (PFC), and specifically designed to contain words and word combinations known to illustrate regional phonetic variation in French. The participants were asked to read the texts at their own pace, from a computer screen. The texts were presented using the Calibri font, at 28 pts, in white colour on a black background; the participants used the

space bar to move to the next screen. The purpose of this recording was to collect neutral read speech for each participant, in a baseline condition, i.e. without inducing cognitive load.

In the next phase, the participants performed a Stroop test (Stroop, 1935), under three conditions of increasing cognitive load. In the low-load condition, the participants were asked to name colour words displayed in matching colours on the screen; the participant pressed the space-bar to move to the next word (no time pressure). In the medium-load condition, the participants were asked to name colour words displayed in mismatched colours on the screen; the participant pressed the space-bar to move to the next word (no time pressure). In the high-load condition, the participants were asked to name colour words displayed in mismatched colours on the screen; each word was displayed for 800 ms, then moving to the next word (time pressure). For each condition, there was a practice sequence of 10 colour words, followed by 3 trials of 20 colour words each: in total, each participant produced 60 colour words per condition. The colour names used were: blanc, bleu, brun, gris, vert, orange, rose, violet, rouge, jaune. Table 18.1 shows the red-green-blue (RGB) values used for each colour. The colour words were presented in random order.

Colour name French	RGB value	Colour name English
blanc	(255, 255, 255)	white
bleu	(0, 0, 255)	blue
brun	(139, 69, 19)	brown
gris	(169, 169, 169)	grey
vert	(0, 128, 0)	green
orange	(255, 140, 0)	orange
rose	(255, 192, 203)	pink
violet	(128, 0, 128)	purple
rouge	(255, 0, 0)	red
jaune	(255, 255, 0)	yellow

TABLE 18.1: Colour names and Red-Green-Blue (RGB) values used in the Stroop tests)

The next phase consisted of an automated Reading Span test, where participants were required to read aloud a sentence, verify its plausibility, and memorise a letter that briefly appeared on the screen. After a certain number of sentence/letter pairs has been presented, the participant was asked to recall the letters seen, in order, and enter the string of letters using an on-screen form. Letters appeared on the screen for 800 ms. Participants were required to read aloud the sentences; they were encouraged to read aloud the letters

as well, but it was not required. The number of sentence/letter pairs ranged from to 2 to 5, and the participants did not know beforehand how many sentence/letter pairs they would see in a trial. There were 21 sets of trials in the experiment, grouped as follows:

5 sets with 2 sentences/letters 10 utterances
5 sets with 3 sentences/letters 15 utterances
5 sets with 4 sentences/letters 20 utterances
6 sets with 5 sentences/letters 30 utterances

The sentences used for the reading span test were adapted from the French version of the Daneman and Carpenter (1980) reading span test, presented in Desmette, Hupet, Schelstraete, and van der Linden (1995). The adaptation consisted of changing one word in the sentence to produce a corresponding implausible (not logical) sentence; an effort was made to retain approximately the same number of syllables. The sentences used in the Reading Span test can be found in the Annex. Each participant produced 75 utterances.

The fourth and final phase of the experiment was a Stroop test with a constant moderate time pressure, and a dual task in the high load condition. There were three conditions: in the low load condition, the word name and the display colour was congruent; in the medium and high load conditions, the word name and the display colour was incongruent. In all conditions, each word appeared for 1000 ms, before moving to the next word. In the high load condition, the participants were required to perform an additional tone counting task. A tone was played through headphones every two seconds (approximately every two colour words), at random time intervals. The tone was either a low-pitch tone (1000 Hz) or a high-pitch tone (2000 Hz). A random-number generator was used to decide whether to play a tone, and subsequently whether the tone would be low-pitch or high-pitch. The subjects were instructed to count only the high pitched tones, and give the number at the end of each trial. They were not allowed to use their fingers for counting. For each condition, there was a practice sequence of 10 colour words, followed by 3 trials of 20 colour words each: in total, each participant produced 60 colour words per condition. The colours used were the same as in the Stroop/Time Pressure task (Table 18.1).

18.2 PARTICIPANTS

We collected usable data from 9 participants, after excluding one non-native speaker of French, one person with colour blindness and one participant for which the EGG signal was too weak. All participants were university students recruited through the participant pool of the Faculty of Psychology at Université catholique de Louvain. All participants reported French as their first mother tongue (one participant is a bilingual French-Dutch speaker and

one participant is a bilingual French-Kinyarwanda speaker). The mean age of participants was 22.7 years old (standard deviation: 2.4), 2 participants are male and 7 participants are female. All participants reported knowledge of one second language (English: 5, Dutch: 2, Occitan: 1, Spanish: 1), 7 participants reported knowledge of a second L2 (English: 2, Dutch: 1, German: 1, Spanish: 3) and 3 participants reported knowledge of a third L2 (Dutch: 2, German: 1). One participant had musical training. The demographic data was collected by means of a questionnaire, given to both participants at the end of the experiment. The subjective task difficulty ratings were also collected using the same questionnaire.

FIGURE 18.1: One participant and the experimental setup of Study 1

18.3 DATA COLLECTION

Speech produced by the participant was recorded throughout the experiment using a DPA 4066-B34 omnidirectional headset microphone, connected to a Zoom Z24 multi-track recorder, through an external microphone preamplifier. The recordings took place in a quiet but not acoustically treated room at the university premised in Louvain-la-Neuve. Electroglottographic data was obtained using the VoceVista portable electroglottograph, model 5070A; the EGG signal output was routed to the Zoom Z24 recorder. The distractor stimuli (during the Stroop/Dual task session) were presented to the subject using noise-cancelling Philips FX4M headphones. A 4-track recording of the entire experimental session was obtained. Track 1 is a recording of the participant speech; track 2 is the recording of the EGG signal; track 3 contains a time-synchronised recording of the distractors played into the participant's headphones; and track 4 contains a signal to synchronise the audio recording with

the videos colleced with the eye-tracker (a pulse each time the eye-tracker video recording starts or stops).

The recording was performed using a sampling rate of 44.1 kHz and 16-bit resolution. In addition to the audio recordings, eye tracker data was collected for the driver only, using the portable head-mounted Pupil eye-tracker (Kassner, Patera, & Bulling, 2014). In this study, we used the second-generation model, recording both eyes at a sampling rate of 60-80 Hz (time-stamped frames collected at a variable framerate). Stimuli were presented using a Windows PC laptop, and the eye-tracker data were collected on a Mac notebook. Figure 18.1 shows one participant and the experimental setup used in this study.

18.4 SUBJECTIVE RATINGS OF TASK DIFFICULTY

At the end of the experiment, participants were asked to fill in a questionnaire that included a section in which they provided a subjective evaluation of the difficulty of different tasks, using a 7-point Likert scale. The tasks were described as shown in Table 18.2.

Task code	Task name
ReadingBise	Lecture : La bise et le soleil
ReadingBeaulieu	Lecture : Beaulieu
StroopA_Cong	Couleurs - avancement par moi-même - même encre
StroopA_Inco	Couleurs - avancement par moi-même - encre différent
StroopA_Time	Couleurs - avancement automatique - encre différent
RSpan2or3	Phrases + mémorisation des lettres 2 et 3 lettres
RSpan4or5	Phrases + mémorisation des lettres 4 et 5 lettres
RSpan6	Phrases + mémorisation des lettres 6 lettres
StroopB_Cong	Couleurs - avancement rapide - même encre
StroopB_Inco	Couleurs - avancement rapide - encre différent
StroopB_Dual	Couleurs - compter les tons en même temps

TABLE 18.2: Tasks used in Study 1

The aggregated results of the summary ratings are shown in Figure 18.2. These ratings validate the choice of tasks as a means of inducing low, medium and high-level cognitive load on the participants. It is interesting to note that the Stroop/Dual task was globally rated as more difficult than the Stroop/-Time pressure task; and that even the low-load condition of the Reading Span task was rated approximately as difficult as the medium-load conditions of the two Stroop tasks (i.e. the incongruent colours condition).

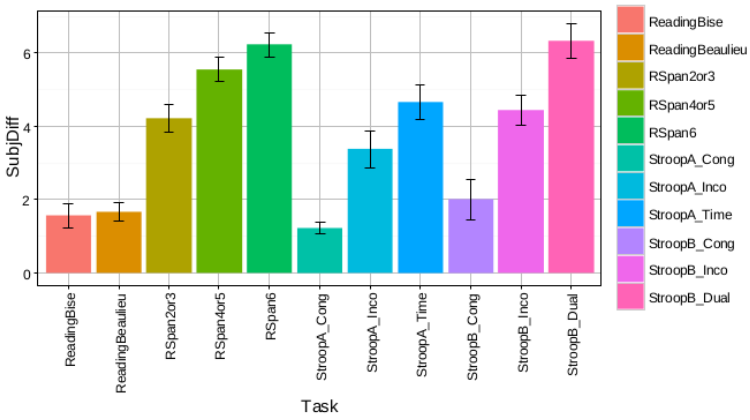


FIGURE 18.2: Subjective ratings of task difficulty for Study 1 (means and standard error bars)

18.5 DATA ANALYSIS AND RESULTS

Our analysis focuses on the vocal fold vibration parameters as approximated by the EGG signal. Figure 25 shows a typical electroglottographic (EGG) signal and its first derivative (DEGG). The electroglottograph works by applying a low-voltage, low alternating current across two electrodes placed on the neck, at the level of the thyroid cartilage. The EGG device detects and processes changes in the electrical resistance between the two electrodes, to produce a signal where lower values indicate lower vocal fold contact area. The minima in EGG waveforms appear then the vocal folds are open. The vibratory patterns of the vocal folds are more prominent in the DEGG signal, which is the first derivative of the EGG signal (see Figure 18.3).

The maximum positive peak in the DEGG signal occurs when the rate of closing of the vocal folds is at its maximum. The maximum negative peak of the DEGG signal occurs when the rate of opening of the vocal folds is at its maximum (Yap, 2012, p. 74). The pitch period can be calculated as the time span between two DEGG positive peaks. The time during which the vocal folds are closed can be calculated as the duration between the positive DEGG peak and the next negative DEGG peak; and the time during which the vocal folds are open can be calculated as the duration between this negative DEGG peak and the next positive DEGG peak. If top is the time interval between a glottal opening instant and the next glottal closure instant and To is the pitch period, the Opening Quotient (OQ) is defined as:

$$OQ = \frac{t_{op}}{T_0}$$

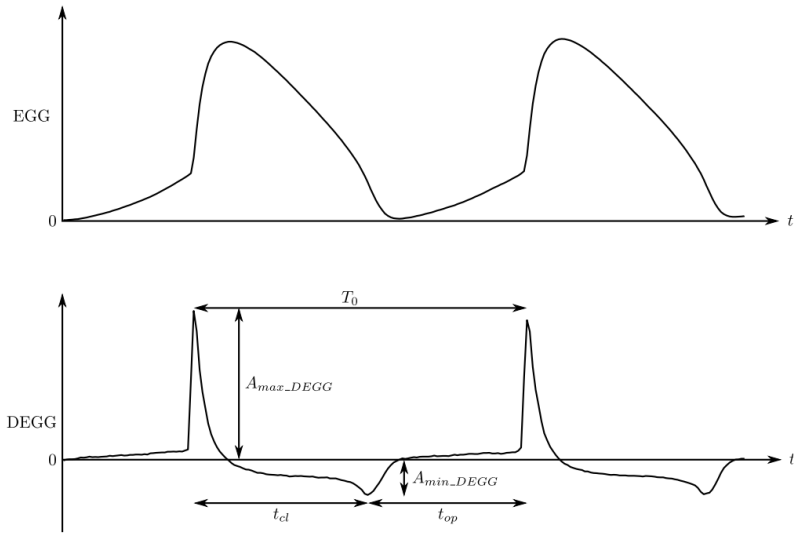


FIGURE 18.3: EGG, DEGG signals and parameters calculated from them (From Yap, 2012)

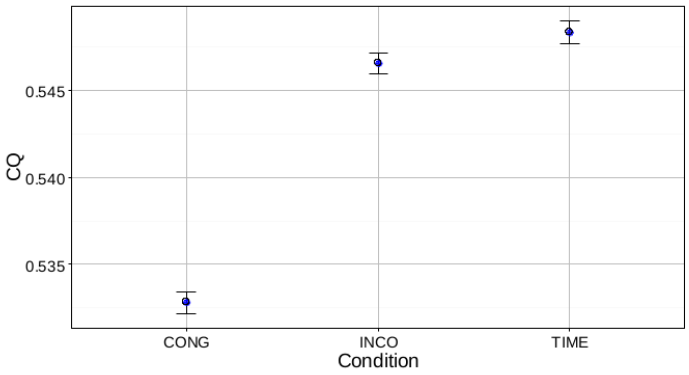
and the Closing Quotient is defined as:

$$CQ = 1 - OQ$$

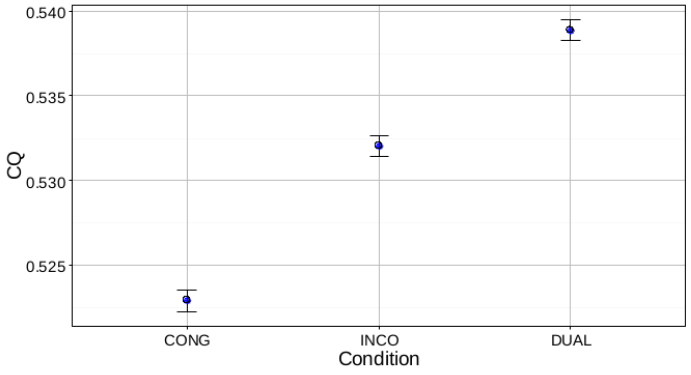
In other words, a higher CQ value indicates that the vocal folds stay closed for a longer proportion of the pitch period.

We analysed the data collected using the EGGWorks software (Tehrani, 2009), version 11.07.13. Smoothing was used (30-point) and each glottal pulse produced one measurement value. Figure 18.4a shows the mean value and 95% confidence interval of the Closed Quotient, calculated from the EGG signal, across the three conditions in the Stroop/Time task and Figure 18.4b shows the mean and 95% confidence interval of the Closed Quotient, across the three conditions in the Stroop/Dual task. The CONG condition (congruent colours) is the low-load condition, the INCO (incongruent colours) is the medium-load condition and the TIME (time pressure) or DUAL (dual task) conditions are the high-load conditions.

Tables 18.3 and 18.4 show the mean Closed Quotient value (and its standard deviation) per condition, for the Stroop/Time and the Stroop/Dual tasks respectively. In the Stroop/Time task Mann-Whitney U tests indicate that mean CQ is significantly higher in the INCO condition compared to the CONG condition ($p < 0.001$), and in the TIME condition compared to the INCO condition ($p < 0.001$). In the Stroop/Dual task, Mann-Whitney U tests indicate that mean CQ is significantly higher in the INCO condition com-



(a) Stroop / Time task



(b) Stroop / Dual task

FIGURE 18.4: Mean and 95% confidence intervals of the CQ EGG measure across conditions in the Stroop tasks

pared to the CONG condition , and in the DUAL condition compared to the INCO condition ($p < 0.001$).

Our study on French (limited to the Stroop test data) confirms Yap's (2012) findings for English. The the proportion of time the vocal folds remain closed (as estimated by the CQ EGG measure) increases as cognitive load increases. This indicates that the glottal source features are sensitive to cognitive load, and that they can be used for the automatic classification of speech (at the articulation level) under varying levels of cognitive load.

Condition	N	Mean CQ	Std Dev
CONG	158370	0.5327	0.1247
INCO	170711	0.5465	0.1278
TIME	153497	0.5483	0.1328

TABLE 18.3: Tasks used in Study 1

Condition	N	Mean CQ	Std Dev
CONG	155185	0.5229	0.1264
INCO	165040	0.5321	0.1246
DUAL	184288	0.5389	0.1369

TABLE 18.4: Tasks used in Study 1

19

STUDY 2: QUESTION-ANSWERING AND READING COMPREHENSION MONOLOGUE

The objective of this study was to collect monologue speech produced under increasing levels of cognitive load. The subjects were asked to perform tasks that necessitate memorisation of recently presented information. Unlike Study 1, we have attempted to create an experimental setting in which the subjects produced both controlled and spontaneous speech, and utterances long enough to permit prosodic analysis. This study was partially inspired by the "Reading Comprehension" corpus described in Yin et al. (2007). In their study, speakers had to answer reading comprehension questions regarding a short paragraph they had just read, and different levels of cognitive load were induced by varying the difficulty of the text. We expanded this design to include a session where speech was elicited by a sequence of short texts and related multiple-choice questions, in addition to the task described in Yin et al. (2007).

19.1 EXPERIMENTAL DESIGN AND MATERIALS

The experimental sequence comprised of four phases:

1. answering multiple-choice question on the basis of a short text;
2. reading, summarising and answering open-ended reading comprehension questions about a narrative text;
3. reading, summarising and answering open-ended reading comprehension questions about an argumentative text; and
4. a short interview.

In the first phase, the subjects were presented with a short text on a computer screen, and were asked to read the text aloud at their own pace. They could move on to the next screen by pressing the space bar; the short texts consisted of 2 to 3 screens. Immediately after reading the short text, the subjects were presented with one multiple-choice question, related to the text, with four choices. They were asked to read the question aloud and to select

one of the four answers, while explaining their reasoning: the subjects were invited to give the arguments that led them to choose one of the four options. The subjects then moved to the next text-question pair, without receiving feedback. There were 10 text-question pairs, in a fixed order of presentation.

In the second and third phases, the subjects were presented with a text on the computer screen and were asked to read it aloud, at their own pace (they used the spacebar to move to the next screen; each screen contained one or two sentences, and no sentence span more than one screen). They were then invited to answer the following three comprehension questions:

- Give a short summary of the story in at least five whole sentences.
- What was the most interesting point in this story?
- Describe at least two other points highlighted in this story.

The text used in the second phase of the experiment was the fictitious newspaper article "Le premier-ministre ira-t-il à Beaulieu" ("Will the prime-minister visit Beaulieu?"), created by the project Phonologie du Français Contemporain (PFC). In the third phase of the experiment, the text used is a presentation about the contemporary financial crisis, extracted from the transcription of the interpreter's speech in Study 3 (in which a professional interpreter was asked to render in French a presentation on the financial crisis by a German economist). The first text is essentially narrative and contains everyday vocabulary, while the second text is mainly argumentative and contains technical terminology (economy and finance). It was hypothesised that the participants would find the second text more difficult to understand and to memorise; all participants reported that indeed that was the case, when asked in the final interview.

In the third phase of the experiment (reading comprehension; text on the financial crisis), an additional load was placed on the participants using distractors. While producing speech, i.e. while answering the three reading comprehension questions, the participants wore headphones and heard a series of three numbers, at random intervals. Immediately after completing their answer to the question, they were asked to recite the numbers, in the order they heard them.

The fourth and final stage of the experiment consisted of a short unscripted, spontaneous interview with the experimenter, in which the subjects were asked to evaluate the difficulty of the tasks performed, and give feedback on the strategies they had used during the experiment's phases. The experimental sequence's duration was approximately 1 hour and participants proceeded through all phases in one sitting, with short breaks between phases. All materials used in this study (texts, questions, distractors) are reproduced in the Annex.

19.2 PARTICIPANTS

The participants to this study were 11 university students (ages: 22-23 years old, and one participant 32 years old), 4 male and 7 female. They were all students at the Faculty of Modern Languages at Université catholique de Louvain, enrolled in one of the following three Masters: Linguistics, Romance Languages, or Romance & Germanic Languages. They participated in the experiment for course credit, as part of their course Phonologie & Prosodie during the academic year 2014-2015, and were not otherwise remunerated.

19.3 DATA COLLECTION AND ANALYSIS

The experiment took place in a quiet, but not acoustically treated, office at the university premises in Louvain-la-Neuve. Speech was recorded using a DPA 4066-B34 omnidirectional headset microphone, connected to a Boss BR-800 digital multi-track recorder. Distractors were presented using Sennheiser HD-239 headphones. Eye-tracking data was simultaneously recorded using the portable head-mounted Pupil eye-tracker (Kassner et al., 2014); in this study we used the first generation model that supports a 30 Hz sampling rate and only records the right eye. A synchronisation signal was sent to the multi-track recorder, to synchronise the eye-tracker with the speech recording. Speech was recorded at a sampling rate of 44.1 kHz with 24-bit resolution.

Initial processing of the speech recordings was performed using Audacity, where the signal was normalised (peak amplitude: -1dB), and in Praat, where the beginning and end of each experimental phase was marked. Subsequently, the long per-speaker recordings were split into short wave files (one file per task and speaker) and a corpus containing these (short) files was created in Praaline.

The transcription and speech alignment were performed iteratively as follows. Initially, CMU Sphinx (http://cmusphinx.sourceforge.net) was used to build a speaker-dependent acoustic model based on the samples where the text was known (i.e. parts of the recording where the speaker was reading a text from the screen). The speaker-dependent acoustic model was created by applying MLLR (Maximum Likelihood Linear Regression) adaptation on the basic Sphinx model for French. Using these models, the recordings were split into utterances and automatically transcribed. The automatic transcriptions of the Text reading (first) and Comprehension Questions (second) phase were corrected manually. A modified version of the EasyAlign script was used to perform a phone-level alignment, based on the automatic phonetic transcription of the output. Finally, the Prosogram script was applied to the entire corpus, using automatic syllable detection.

19.4 RESULTS

Each participant produced the following speech samples:

- 10 samples: reading a short text and the corresponding comprehension question
- 10 samples: answering the comprehension question and presenting their arguments for the answer chosen
- 2 samples: reading the longer texts (Beaulieu and the text on the economy)
- 2 samples: summarising the longer texts
- 2 samples: describing the main ideas in each of the two longer texts
- 1 sample: spontaneous speech in the final interview

Responding to the comprehension questions, providing the summaries and recalling the main ideas of the texts are the tasks that impose a higher load on working memory.

The global prosodic measures calculated for each sample, based on the automatic processing of the corpus with the Prosogram script, are shown in Table 19.1.

The following observations can be made with respect to temporal prosodic variables. The percentage of silent pause time was higher in tasks involving higher CL, as can be seen in Figure 19.1.

We further analysed silent pause durations as a mixture of three log-normal distributions while the participant was reading the short texts (Figure 19.2a) and while they were answering the reading comprehension question (Figure 19.2b). We observe that the third (long pause) component distribution is stronger in the higher cognitive load condition.

With respect to the speech rate variability, we calculated the mean, standard deviation, standard error and 95% confidence interval of the stylised syllabic nuclei durations, for each type of task (condition); the results are presented Table 19.2 and Figure 19.3.

We confirm our hypothesis that tasks implying a higher cognitive load will produce speech articulated with higher variability: this observation is mainly explained by the prevalence of filled paused and hesitation-related lengthening (drawls). An F-test between low-load tasks (TEXTS and READ) and high-load tasks (RESPONSE, SUMMARY, IDEAS) indicates that the variances are not equal: the 95% CI of the ratio of variances is (0.378, 0.394) ($F = 0.38634$, $p < 0.0001$).

Finally, with respect to the pitch-related features, we have calculated the mean and 95% confidence intervals of the mean pitch, and of the pitch trajectory (z-score, normalized per sample), per task type. The results are presented

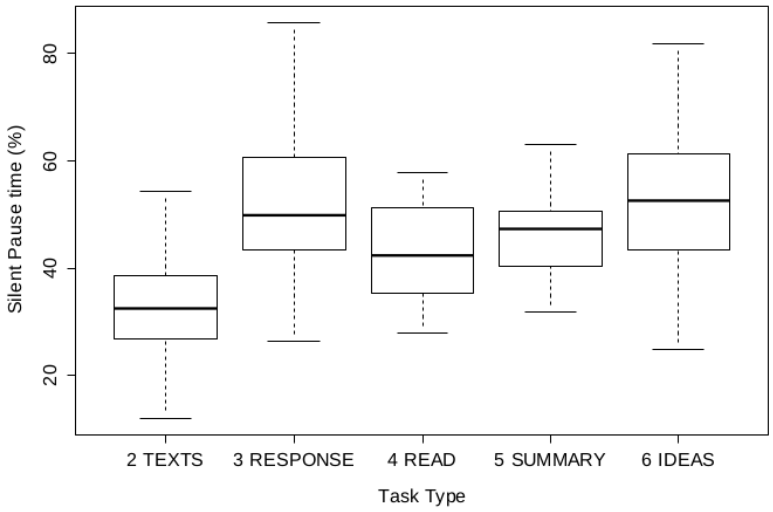
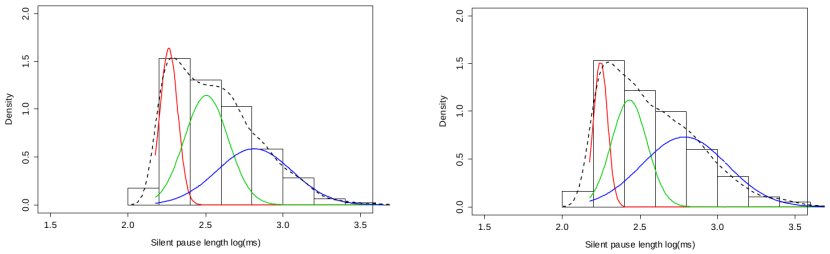


FIGURE 19.1: Proportion of silent pause time (%) per task type

in Figure 19.4 and Figure 19.5 respectively. We do not detect an effect of the task type on the mean pitch of the speaker. We partially confirm our hypothesis that the higher CL tasks will result in less expressive speech (i.e. in a lower pitch trajectory), with the exception of the responses to the comprehension questions related to the short texts.



(a) Reading (b) Answering comprehension questions

FIGURE 19.2: Silent pause length modelled as a mixture of log-normal distributions while reading the short text (text) and while answering the reading comprehension questions (right)

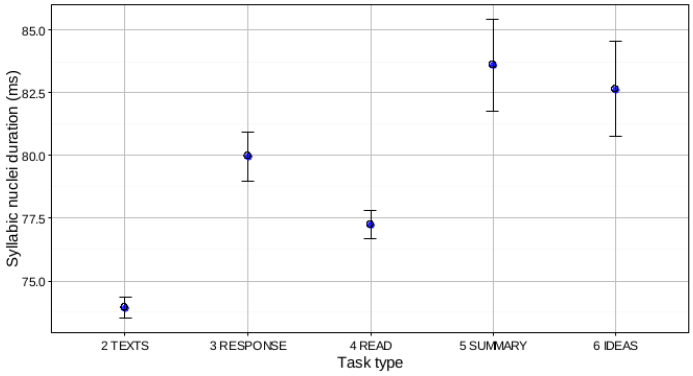


FIGURE 19.3: Syllabic nuclei duration: mean and 95% confidence interval, per task type

Measure	Description
SpeechTime	Total speech time (in seconds): internucleus time + intranucleus time + pause time
PhonTime	Phonation time (in seconds) without pauses = internucleus time + intranucleus time
PropPhon	Proportion (%) of estimated phonation time (internucleus time + intranucleus time) to speech time
PropPause	Proportion (%) of estimated pause time (when internucleus time >= 0.3) to speech time
SpeechRate	Estimated speech rate (in syll/s) = nr of nuclei / phonation time
MeanOfST	Mean of pitch values, where values are min and max pitch in ST for each syllable
StdevOfST	Standard deviation of pitch values, where values are min and max pitch in ST for each syllable
PitchRange	Estimated pitch range (in ST) (2%-98% percentiles of data in nuclei without discontinuities)
Gliss	Proportion (%) of syllables with large pitch movement (abs(distance) >= 4ST)
Rises	Proportion (%) of syllables with pitch rise (>= 4ST)
Falls	Proportion (%) of syllables with pitch fall (<= -4ST)
NuclDur	Sum of durations for nuclei for this speaker
InterNuclDur	Sum of durations between successive nuclei for a speaker
TrajIntra	Pitch trajectory (sum of absolute intervals) within syllabic nuclei, divided by duration (in ST/s)
TrajInter	Pitch trajectory (sum of absolute intervals) between syllabic nuclei (except pauses or speaker turns), divided by duration (in ST/s)
TrajPhon	Sum of TrajIntra and TrajInter, divided by phonation time (in ST/s)
TrajIntraZ	Like TrajIntra, but for pitch trajectory in standard deviation units on ST scale (z-score) (in sd/s)
TrajInterZ	Like TrajInter, but for pitch trajectory in standard deviation units on ST scale (z-score) (in sd/s)
TrajPhonZ	Like TrajPhon, but for pitch trajectory in standard deviation units on ST scale (z-score) (in sd/s)

TABLE 19.1: Global prosodic measures calculated by Prosogram

Task type	N	Nucl Dur (ms)	Std Dev	Std Error	95% CI
TEXTS	37655	73,93	40,58	0,21	0,41
RESPONSE	15773	79,96	62,54	0,50	0,98
READ	18751	77,24	39,47	0,29	0,56
SUMMARY	4668	83,59	63,73	0,93	1,83
IDEAS	5435	82,64	71,48	0,97	1,90

TABLE 19.2: Syllabic nuclei duration mean and variability measures, per task type

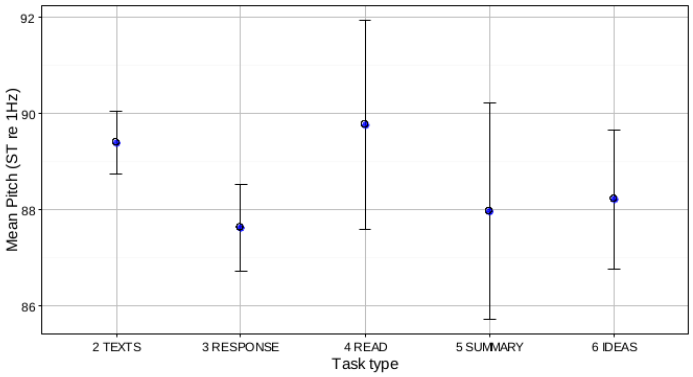


FIGURE 19.4: Mean pitch in semitones (relative to 1Hz), mean and 95% confidence interval, per task type

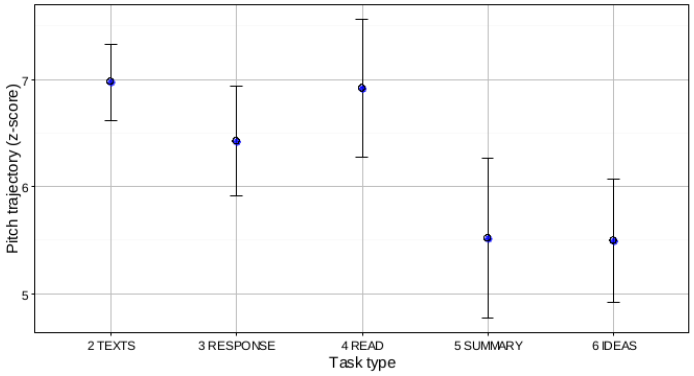


FIGURE 19.5: Pitch trajectory (z-score), mean and 95% confidence interval, per task type

20

STUDY 3: SIMULTANEOUS INTERPRETING AS A FORM OF LANGUAGE PROCESSING UNDER COGNITIVE LOAD

This study explores the relationship between prosodic features specific to simultaneous interpreting and the listeners' perception of the fluency and accuracy of interpreting, as well as their comprehension of the source speech. It was conducted in collaboration with Cédric Lenglet at the University of Mons, and is presented in detail in the PhD thesis Lenglet (2015). This chapter is largely base on the joint publication Christodoulides and Lenglet (2014).

Two groups of participants (47 subject experts and 40 non-experts) listened to a 20-minute lecture in German, along with its interpretation into French under two conditions (the actual interpretation, or a read-aloud rendition of the same text by the same interpreter) and answered comprehension and rating questions. The prosodic features of the two conditions were analysed, and differences regarding the temporal organisation of speech, disfluencies, pitch register and the interface between prosody and syntax emerged. Our results suggest that interpreting-specific prosodic features affect the perception of fluency, which in turn affects the perception of accuracy. However the impact on listeners who enjoy relevant contextual knowledge is less pronounced. In this study, we focus on the prosodic profile of the simultaneous interpretation as compared to the prosodic profile of the read speech by the same speaker. We present the data on the evaluation of fluency and comprehension for purposes of completeness.

Simultaneous conference interpreters facilitate multilingual communication in political, technical and other meetings. Typically, they work in soundproof booths, translating speech in real-time so that the participants can follow the debate in their language, without interruption. They are expected to "communicate the speaker's intended messages as accurately, faithfully, and completely as possible (...) and to be clear and lively in [their] delivery" (Association Internationale des Interprètes de Conférence, 1999/2004). Since prosody conveys important information in human communication (e.g. information status, focus, intent, emotion), the interpreter is expected to decode such information from the prosodic structure of the source language and encode it in parallel constructions in the target language. This "translation" of prosodic information, along with the linguistic information, from

SL to TL inspired studies on the correspondence, or alignment, of prosodic patterns across languages.

Beyond these conscious choices made by the interpreter, however, the prosodic features of SI are influenced by two constraints: the reformulation process of translation and the high cognitive load induced by the task itself (Seeber & Kerzel, 2012). SI strategies such as stalling (i.e. waiting for enough input before producing a translation or committing to a syntactic structure) and anticipation (i.e. predicting part of the speaker's input) (Gile, 2009, p. 201) affect the temporal structure of interpreters' speech. Speech produced during simultaneous interpretation is expected to exhibit the prosodic characteristics of speech produced under high cognitive load.

Surveys among users of simultaneous interpreting suggest that they consider accuracy (or fidelity) to be a crucial quality criterion, whereas prosodic features, such as intonation or accent, are not deemed paramount (Collados Aís, Macarena Pradas Macías, Stévaux, & García Becerra, 2007). However, many users cannot assess the interpreters' accuracy because of their lack of knowledge of the source language. Instead, the users' perception of SI quality may depend on the prosodic features of the interpreters' speech, including intonation, hesitations and pauses. Moreover, the interpreters' liveliness appears to influence the listeners' understanding of the speech content (Holub & Rennert, 2011). In other words, the core objective of SI, namely to "produce the same effect on [the listeners] as the original [speech] does on the speaker's audience" (Déjean Le Féal, 1990), might depend not only on what the interpreters say, but also on how they say it.

The link between quality perception and prosody is all the more important in SI, as previous research indicates that simultaneous interpretation has a distinctive prosodic profile. It results from the interplay of the choices and constraints outlined above, as a speaking style determined both by the situational context and by individual characteristics. A particular prosodic profile for SI has been observed in studies for at least the following language combinations: Hebrew to/from English, with a large number of "low-rise non-final pitch movements" (Shlesinger, 1994, p. 231); English to German, with "long pauses [and a] high proportion of final pitch movements that indicate a continuation" (Ahrens, 2005, p. 72); and English to French, with less numerous and longer silent pauses, and a narrower pitch range compared to the source speech (Christodoulides, 2013).

Fluency has been regularly used as a quality criterion in expectation surveys on simultaneous interpretation, although it is a polysemous concept. In ordinary language, fluency refers to general (often foreign) language proficiency, whereas a more technical definition associates fluency with speech flow and absence of disfluencies such as pauses, hesitations and repetitions (Chambers, 1997, p. 537). Experimental research has shown that these temporal features do not only influence the perception of the interpreter's fluency, but also that of its intonation (Collados Aís et al., 2007, p. 67). Listeners

asked to rate "fluency" seem to blend temporal features with intonation (e.g. pitch variation). Consequently, it seems appropriate to merge these parameters in a perceptual study.

Does the particular prosody of simultaneous interpretation have an impact on its perception? To date, most studies on prosody and quality in SI are based on carefully doctored speeches. Consequently, their findings cannot be linked directly to the perception of the authentic prosody of SI. Shlesinger (1994) conducted a small-scale experiment on the impact of authentic SI prosody on 15 listeners' understanding of speech content, comparing excerpts of speeches produced under two different conditions: read aloud from a script and interpreted simultaneously. In a listening test, the subjects' scores in the "read-aloud" condition were approximately twice as good. She concluded that SI intonation affected meaning and perception, but she argued that this effect would be counterbalanced in the case of authentic conference participants with the relevant contextual knowledge (Shlesinger, 1994, p. 234). Our experiment aimed to answer the following research questions: Does simultaneous interpreting have particular prosodic features? If yes, do these features influence the listeners' objective and subjective understanding of the speech content, and their perception of the interpreter's fluency and accuracy?

20.1 EXPERIMENTAL DESIGN

Our goal was to create a situation as close to authentic SI as possible. The original speech is an abridged presentation on investment strategy by a German fund manager; its duration is 20 minutes. German was selected as the source language in order to increase the likelihood that French-speaking listeners rely on the interpreter only. A professional conference interpreter (male, French native speaker, 6 years of experience) interpreted the German presentation into French in a state-of-the-art interpreting booth. The recording of this interpretation was transcribed; punctuation was added at syntactically - complete clause boundaries; discourse markers and interjections were included in the transcription (only filled pauses, e.g. 'euh', were omitted). The same interpreter read the transcript, after rehearsing it, and was recorded in a booth. We thus obtained two different prosodic profiles by the same speaker: under authentic SI conditions; and prepared reading, without the cognitive constraints of SI. The two versions were synchronized to the video of the original presentation using Praat, Audacity and AviDemux.

The experimental design is a conference simulation adapted from Holub and Rennert (2011). The subjects watch the video of the German presentation and listen to an interpretation into French. An interpreter pretends to work in a booth at the back of the room. This creates the impression of a live interpretation, whereas actually, the subjects are listening to one of the recordings, according to the experimental condition they were assigned to.

The subjects were 87 French-speaking university students: 47 students of economics and 40 translation students. Students in economics were chosen because of their specialized knowledge and their greater availability than professional economists. Translation students were chosen in order to control the influence of prior thematic knowledge. The subjects were matched for academic performance (based on grade records) to control memory and prior knowledge. The resulting pairs were randomly distributed between two experimental conditions: interpreted and read-aloud speech.

We use a listening comprehension test and an assessment questionnaire, which we both pretested extensively. The listening test consists of 3 multiple-choice and 4 half-open questions and assesses the comprehensibility of the speech with a listening score (interval scale). In the assessment questionnaire, the subjects are asked to rate on a 7-point ordinal scale how fluent the interpreter's delivery was (Fluency), how well they think they understood the lecture (Subjective Comprehension) and how accurately they reckon the interpreter rendered the speech (Accuracy).

20.2 LINGUISTIC AND PROSODIC ANALYSIS OF THE TWO CONDITIONS

The two recordings (SI and Read) were orthographically transcribed in Praat. We obtained a phonetic transcription as well as an automatic segmentation of words, syllables, phones and pauses, automatically using EasyAlign; the alignment was corrected manually. A "delivery" tier was added to annotate articulation-related (schwa, creaky voice, liaison and elision) and paralinguistic phenomena (audible breath, noises). Part-of-speech tagging and multi-word unit detection were obtained automatically using DisMo and subsequently verified manually. We applied an annotation scheme for disfluencies described in 13, covering: i) single-token disfluencies: filled pauses, hesitation-related lengthening, lexical false starts and intra-word pauses; ii) structured disfluencies: repetitions (of one or more words), deletions, substitutions, insertions, and complex combinations of the above.

To process our data we used Praaline, a toolkit that interfaces with Praat and runs a cascade of scripts and/or external analysis tools, each of which may add features to an annotation level (e.g. syllables, words etc.), stored in a relational database. We applied Prosogram's two-step algorithm for pitch stylisation: for each syllable, vocalic nuclei are detected based on intensity and voicing, and then the fo curve on the nucleus is stylised into a static or dynamic tone, based on a perceptual glissando approach. Syllabic prominence was estimated with ProsoProm (Goldman, Avanzi, Auchlin, & Simon, 2012), Analor (Avanzi, Lacheret, & Victorri, 2008), and a manual perceptual annotation was also performed. Segmentation into accentual and intonational phrases was performed by an expert annotator (taking into account all prominence scores); furthermore, perceptually-motivated prosodic boundaries were calculated based on the approach proposed in Mertens and Simon

(2013). Several aggregate measures were calculated using ProsoReport (Goldman, Auchlin, & Simon, 2009).

In order to study the interface between prosody and syntax, a three-level syntactic annotation was added. First, an annotation into minimal chunks based on the phrasal tag-set of the French Treebank (as described in Chapter 12) was added manually. We also applied the model for syntactic annotation into functional sequences and dependency clauses detailed in Degand and Simon (2009) and obtained segmentation into Basic Discourse Units whenever the major prosodic boundaries and the dependency clause boundaries coincide. In total, the two recordings are 42-minutes long (1256 seconds each), and contain 8760 syllables, 1335 silent pauses, and 6143 tokens (words).

20.3 EVALUATION OF AUTOMATIC TOOLS

As a corollary study, we evaluate the performance of the above-mentioned automatic tools. Results for prominence detection are shown in Table 20.1. There was a fair agreement between the human annotator and the tools, and between the tools themselves. The inter-annotator agreement was consistently lower for the SI condition.

Both conditions	ProsoProm vs. Analor	ProsoProm vs. Manual	Analor vs. Manual
Precision	97.1%	81.9%	59.3%
Recall	39.9%	49.1%	86.4%
Correct	77.4%	84.4%	81.6%
F-measure	56.6%	61.4%	70.3%
Cohen's kappa	0.447	0.524	0.576
Interpreting κ	0.394	0.456	0.561
Reading κ	0.478	0.568	0.581

TABLE 20.1: Evaluation of prominent syllable detection in Study 3

A comparison between the (manually corrected) segmentation into accentual and intonation phrases (AP/IPs), and the perceptually-motivated prosodic boundaries (PBs) is presented in Table 20.2. As expected, the perceptual prosodic boundaries as coarser than the hierarchical segmentation into AP/IPs (which, for French, is based on the prominence of the last syllable of an each unit).

The precision of the POS annotation (DisMo) was 96.3 %, while the syntactic annotation was performed manually.

20.4 GLOBAL PROSODIC FEATURES

A selection of global prosodic features is shown in the Table 20.3.

Sylls with PBs vs. IP/APs	IP boundary		AP boundary	
	Yes	No	Yes	No
Without PB	427	6773	1262	5938
Minor PB	244	264	381	127
Intermediate PB	55	72	73	54
Major PB	921	1	922	0

TABLE 20.2: Comparison of prosodic boundary detection in Study 3

Measure	SI	Read
Articulation ratio (%)	72.6	62.7
Articulation rate (syll/s)	4.91	5.27
Speech rate (syll/s)	3.59	3.33
Speech segments (runs)	493	858
...with average length (syll)	9.2	4.9
Var. coefficient of vowel duration	0.089	0.022
Var. coefficient of syllable duration	0.079	0.042
Median pitch (Hz)	127	152
Pitch range (semitones)	7.8	14.2
Pitch trajectory (semitones/s)	15.13	22.47

TABLE 20.3: Global prosodic measures for Study 3 (Simultaneous Interpreting vs. Read Speech)

We note a higher articulation rate (syllables per second excluding pauses) under the Reading condition. The interpreter made more silent pauses in the Reading condition. This is reflected in the lower articulation ratio, and the lower speech rate (sylls/s including pauses). Speech segments (continuous stretches of speech separated by silent pauses >250 ms) are considerably more numerous and shorter under the Reading condition than in SI. These measures indicate that the interpreter over-segmented his speech in the Reading condition (short utterances and extensive use of pauses). The observed difference between the variance coefficients of vowel duration and of syllable duration indicate that under the SI condition, the interpreter accelerated and decelerated his articulation more frequently than under the Reading condition. Finally, pitch range and pitch trajectory are smaller under the SI condition, compared to Reading; this indicates that the latter was a livelier rendition of the text.

20.5 SILENT PAUSES AND DISFLUENCIES

A Mann-Whitney U test on average silent pause length indicates that it is longer under the SI condition ($p < 0.001$). We modelled silent pause length as

a mixture of log-normal distributions, following the methodology described in Section II.

$$f(x) = \sum_{i=1}^{N} \pi_i \Lambda_i \left(\mu_i, \sigma_i^2, x\right)$$

Three component distributions are identified (using a Bayesian Information Criterion), and their parameters estimated using the Expectation-Maximisation algorithm (Table 20.4 and Figure 20.1). Cut-off values t (local maxima of the model's uncertainty function) are used as thresholds to categorise pauses as 'short', 'medium' or 'long' (instead of ex-nihilo fixed thresholds).

Pause type	SI			Read		
	π	μ	t	π	μ	t
Short	44%	195	283	39%	136	203
Medium	32%	568	1037	48%	581	602
Long	24%	1570		13%	1221	

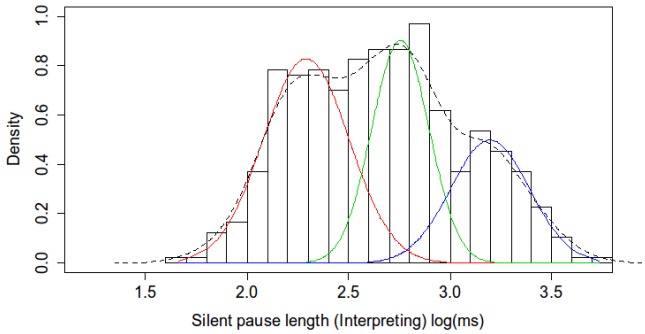
TABLE 20.4: Log-normal mixture model of silent pause length (in ms), $1.3 < \sigma < 1.8$, N = 3.

Regarding disfluencies, under the SI condition the interpreter produced 272 filled pauses vs. only 8 under the Reading condition. Other types of disfluencies were almost inexistent in Reading. Under the SI condition false starts, repetitions and deletions (in order of frequency) were observed. In total, 9.8% of the tokens were disfluent in SI, compared to 0.4% in Reading.

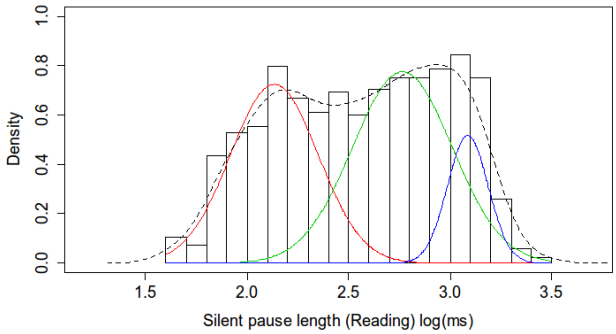
20.6 RESULTS RELATED TO THE PROSODY-SYNTAX INTERFACE

Basic Discourse Units (BDUs) in Degand & Simon (2009) are proposed as "the segments that speakers and listeners use to interpret the discourse they are engaged in". Based on the observation that listeners use both prosody and syntax as cues to information structure, BDUs are defined as segments that run between the points where major prosodic boundaries and dependency clause boundaries coincide. In a congruent BDU, one intonation unit (IU) contains one dependency unit (DU); in intonation-bound BDUs, one IU contains several DUs; in syntax-bound BDUs, one DU packs several IUs. Regulative BDUs contain only discourse markers or adjuncts. A mixed-boundary BDU contains more than one DU and more than one IU, and is the product of a lack of synchrony between prosodic and syntactic boundaries. Table 20.5 shows the distribution of BDUs of different types under the two conditions.

We observe that in the SI condition, there was frequently a mismatch between prosodic and syntactic boundaries. These are typically cases in which the interpreter constructs a phrase incrementally, pausing inside syntactic units. Figure 20.2 shows how the three different types of silent pauses are



(a) Simultaneous Interpreting



(b) Reading

FIGURE 20.1: Density plots of log (silent pause length) and component distributions for the two conditions

distributed between and within syntactic units (chunks and functional sequences) and BDUs.

We observe that in the SI condition, medium-length pauses within constituents and within BDUs occur more frequently than in the Reading condition (+15% and +30% respectively). This is in line with the high percentage of mixed-boundary BDUs observed. The high proportion of syntax-bound BDUs with a relatively short duration under the Reading condition is another indication of over-segmentation.

BDU type	SI		Read	
	%	Avg dur (s)	%	Avg dur (s)
Congruent	21.3	2.57	20.2	2.04
Regulative	21.9	1.24	24.4	0.85
Intonation-bound	5.3	6.43	0.4	1.98
Syntax-bound	30.2	8.20	55.0	5.60
Mixed-boundary	21.3	11.79	0	

TABLE 20.5: Number and average duration of BDUs in Study 3

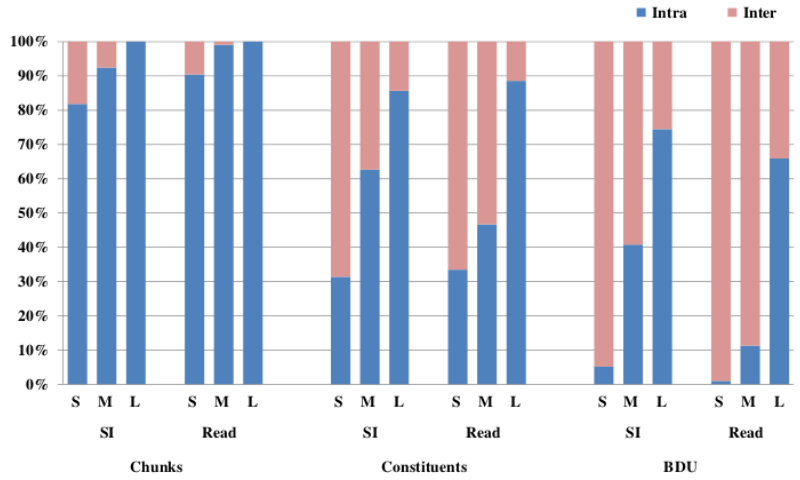


FIGURE 20.2: Silent pauses of different types (S: short, M: medium, L: long), within (intra) and between (inter) syntactic units

20.7 PERCEPTION OF QUALITY AND FLUENCY

The questionnaire data were coded and processed with IBM SPSS Statistics. Missing observations were excluded. The mean Listening Score, which measures Objective Comprehension (the sum of the correct answers in comprehension questions; the maximum score possible is 17), and the median Accuracy, Fluency and Subjective Comprehension ratings (1: best), broken down by subject groups and experimental conditions are shown in Tables 20.6 and 20.7.

The translation students' median ratings of fluency and subjective comprehension are better in the Reading condition. Across all groups, there is a moderate correlation between the experimental condition and the fluency ratings, which turns out to be significant in a Spearman correlation test (Translation: r = –0.506; p = 0.001; Economics: r = –0.323; p = 0.027; Translation + Economics:

Group	SI	Read	Both
Translation students (TRAN)	8.47	9.39	8.94
Economics students (ECON)	7.95	8.39	8.18
Both groups	8.18	8.83	8.52

TABLE 20.6: Mean Listening Score per group (The maximum score is 17. Higher scores indicate better comprehension)

Group	Rating	SI	Read	Both
TRAN	Accuracy	2	2	2
	Fluency	3	2	2
	Subjective comprehension	5	4.5	5
ECON	Accuracy	3	3	3
	Fluency	3	3	3
	Subjective comprehension	3	3	3
Both	Accuracy	2	2	2
	Fluency	3	2	3
	Subjective comprehension	4	4	4

TABLE 20.7: Median quality and subjective comprehension ratings per group (lower scores are better, 1 is best)

r = –0.393; p < 0.001; two-tailed). In other words, the subjects who listened to the read-aloud speech tended to rate fluency better.

Concerning the fluency ratings without regard to the experimental condition, there is a moderate and significant correlation between fluency ratings and subjective comprehension among the students of economics (r = 0.480; p = 0.001, two-tailed). There is a slightly stronger significant correlation between fluency and accuracy ratings across all groups (Translation: r = 0.490; p = 0.002; Economics: r = 0.496; p = 0.001; Translation + Economics: r = 0.520; p < 0.001; two-tailed).

With respect to our first research question, our findings confirm previous studies regarding the particular prosodic characteristics of SI. In the SI condition, the interpreter produced long silent pauses, frequent filled pauses and several reformulation-related disfluencies. The articulation rate was more variable (i.e. more accelerations and decelerations), and the pitch range and trajectory were both narrower in SI, indicating that the same person rendered the text in a more lively fashion, when freed from the cognitive constraints of interpreting. The main effect of SI was observed in the prosody-syntax interface, with often mismatched prosodic and syntactic boundaries, and more intra-unit pauses. The combination of these prosodic features has had an effect on the perceived fluency rating. Previous research has shown that "some

disfluencies may be considered felicitous by listeners" when used for communicative purposes (cf. Part I).

With respect to our second research question, the results lend additional support to the claim that the perception of the interpreters' accuracy is linked to that of their fluency, thus confirming previous experimental findings. The differences in listening scores (objective comprehension) between the experimental conditions are less pronounced among students of economics. This seems to support Shlesinger's claim that the prosody of interpreting has less impact on the listeners who enjoy relevant contextual knowledge. One explanation could be that the translation students processed the speech at a more superficial level and hence, were more affected by perturbations of the prosodic structure of the speech. The students of economics could use their prior knowledge to process the speech content at a deeper level and make inferences to compensate for disturbing prosodic variations. Admittedly, the higher mean listening score of translation students is unexpected. We hypothesize that these students benefited from their capacity to capture the gist of speeches in their notes thanks to an elaborate note-taking technique they develop in introductory courses to conference interpreting. In a future study, perceptual and prosodic data could be correlated to test the effect of each prosodic factor on perceived quality and fluency.

21

STUDY 4: COLLABORATIVE DIALOGUE USING A DRIVING SIMULATOR

The objective of this study was to collect both monologue and dialogue speech, produced under conditions that will tax the attentional resources of the speaker. A dual-task paradigm was used to add a continuous attentional load. Furthermore, we sought to create a realistic communicative situation that would encourage participants to produce long stretches of speech. The goal of this Study is to investigate the temporal characteristics of these speech samples.

We elicited both monologue and collaborative dialogue speech between pairs of participants, while one of them was using a driving simulator. The communicative situation is more realistic than the ones created in Studies 1 and 2, as reflected in the fact that participants produced significantly more speech material. The study uses the dual task paradigm to induce cognitive load: one of the two participants ("the driver") is constantly performing a secondary task which demands attention and co-ordination, in the driving simulator. The other participant ("the passenger") is not engaging in any secondary task. During the same experimental session, participants switched roles: a first recording was performed with the first participant in the role of the driver, and a second recording was performed with the second participant in the role of the driver (a design that facilitates within-subject comparisons).

21.1 EXPERIMENTAL DESIGN

The experimental sequence for a given pair of participants ran in three phases:

1. Syntactically Unpredictable Sentences (SUS) speech perception test,
2. Radio News collaborative dialogue task, and
3. Taboo task.

Each task was repeated in two conditions: a "slow" and a "fast" driving condition, by changing the configuration of the task performed on the driving simulator (see below). The objective of this manipulation was to induce

higher attentional load on the participant by increasing the difficulty of the secondary task (driving).

21.1.1 *Perception of Syntactically Unpredictable Sentences*

In the SUS speech perception test participants listened to short sentences presented over their headphones, and were asked to repeat them as faithfully as possible. The sentences were selected from the French Syntactically Unpredictable Sentences corpus (Boula de Mareüil, d'Alessandro, et al., 2006; Raake & Katz, 2006). The SUS corpus contains sentences following one of the following syntactic forms:

Adverb det. Noun1 Verb-t-pron. det. Noun2 Adjective ?
Determiner Noun1 Adjective Verb determiner Noun2
Determiner Noun1 Verb preposition determiner Noun2

The content words are singular, monosyllabic (unless a final schwa was uttered) and have a high frequency of use according to the BRULEX lexicon; prepositions and determiners are also monosyllabic; adjectives normally placed before a noun in French were avoided. The choice of words is such that the sentences are definitely meaningless: it is thus not possible to use the context or logical induction to infer a word that was not perceived correctly. The SUS list was read by a professional male speaker in a soundproof booth, and the recordings were sampled at 16 kHz (16 bits, mono) in the Wave format. The SUS Phrase Audio corpus was further refined by optimising "for homogeneity in terms of phoneme-distribution as compared to average French, and for word occurrence frequency of the employed monosyllabic keywords as derived from French language databases" (Raake & Katz, 2006, p. 2028). Twenty lists of 12 sentences each are publicly available by the Groupe Audio-Acoustique of the LIMSI research centre, as part of the SUS calibrated audio corpus. The sentence lists used in the experiment can be found in the Annex.

These sentences were mixed with multi-talker babble noise, to create the stimuli used in the speech perception experiment. The signal-to-noise ratio ranged from odB (no noise) to -20 dB, in -2 dB intervals. A small computer programme was written to control the presentation of stimuli (see Figure 21.1). The experimenter would listen to the sentence as repeated by the participant; if more than half of the content words were correctly repeated, the experimenter decreased the SNR (i.e. the next stimulus would be presented in louder babble noise), otherwise the experimented increased the SNR. Using this adaptive procedure, it is possible to estimate the speech reception threshold, defined as the SNR where there is 50% intelligibility (see Raake & Katz, 2006).

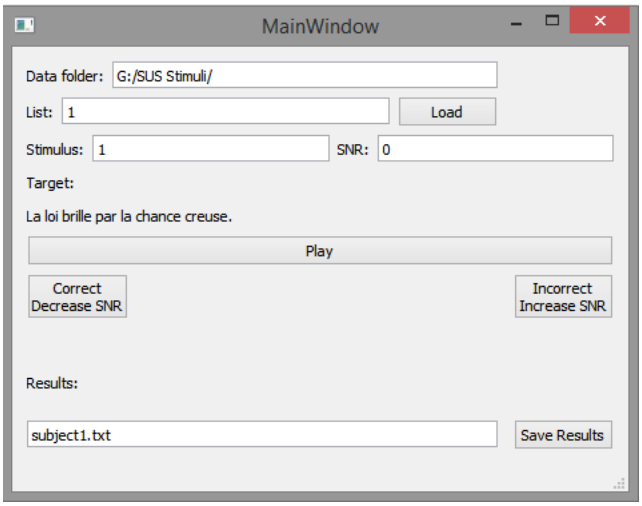


FIGURE 21.1: Experiment control software for the SUS perception test

21.1.2 *Recitation and Collaborative Dialogue in the Radio News Task*

The second and main phase of the experiment consisted of the Radio News collaborative task. In this task, the driver listened to four consecutive radio news items, extracted from the corpus described in Hupin and Simon (2007). The news items were recorded during 2006-2008 from the Belgian radio stations La Première and Bel-RTL. Only the driver listened to the radio news items, through their headphones. A list of four questions related to the content of the news items was simultaneously presented to the passenger's computer screen. After the playback of the radio news items was finished, the driver was asked to give a summary of the news, containing as much information as possible (in order to help the passenger to answer the questions, but without knowledge of the questions at this point). After the driver completed the summary, the passenger would attempt to answer the comprehension questions in sequence, and giving their answers aloud. If there was missing information (as was frequently the case) the driver and the passenger engaged in dialogue in order to find the best answer. After each set of four comprehension questions, a series of four open-ended questions was presented to the passenger, one at a time, as a means to stimulate discussion on a subject related to the news items. The passenger posed the question to the driver, and the two participants exchanged views on the subject. The questions were selected to stimulate debate about topical, societal issues, and to encourage participants to express their personal opinion. Two sets of four items were presented per driving simulator condition (see below) to each pair of participants; therefore, for each pair of participants we recorded 4 driver

summaries (2 per condition), 4 driver-passenger exchanges on the comprehension questions (2 per condition) and 8 short dialogues (4 per condition). The comprehension questions, as well as the questions used to stimulate dialogue can be found in the Annex.

21.1.3 *The Taboo Task*

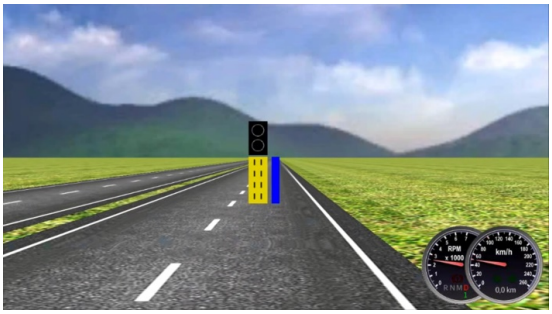
In the third and final phase of the experiment, the participants engaged in fast-paced collaborative dialogue in a game of Taboo. A target word was presented on the computer screen of the passenger (invisible to the driver) along with a list of "forbidden" words. The task of the passenger was to help the driver guess the target word, giving clues but without using the forbidden words. The objective was to have the driver guess as many words as possible in one minute. The passenger had the possibility to skip a word. The presentation of the next word (when the driver had correctly guessed or when a forbidden word was used) was controlled by the experimenter. We recorded one game per driver simulator condition.

After completing the three phases, driver and passenger exchanged roles, and the experiment was repeated. Different SUS sentences, radio news items and related questions, as well as the Taboo words were used (i.e. participants were confronted with new material, regardless of whether they were taking the role of the driver in the first run or in the second run). At the end of the experiment, participants were asked to fill in a questionnaire with basic demographic information and a subjective rating of the perceived difficulty of each task.

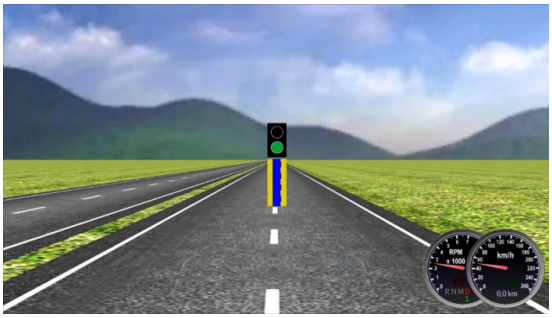
21.2 DRIVING SIMULATOR AND THE CONTINUOUS TRACKING AND REACTION TASK

The driving simulator used was the OpenDS Pro, version 3.5, a Java-based open source system (the Pro version used is commercially licensed). The participant was controlling the simulator by means of the Thrustmaster T500 RS steering wheel and foot pedals. The driving task used was the ConTRe (Continuous Tracking and Reaction) task (see Figure 21.2).

The driving task is described as follows in Mahr, Feld, Mehdi Moniri, and Math (2012): "The driver's primary task in the simulator is comprised of actions required for normal driving: operating the brake and acceleration pedals, as well as turning the steering wheel. System feedback, however, differs from normal driving. In the ConTRe task, the car moves autonomously with a constant speed through a predefined route on a unidirectional straight road consisting of two lanes. Neither operating the acceleration or brake pedal, nor changing the direction of the steering wheel does have an effect on speed or direction of the vehicle. Accordingly, motion rather feels like a video clip. Steering, braking and using the gas pedal do not actually control the car, but



(a) Cylinders



(b) Traffic lights

FIGURE 21.2: ConTRe task screenshots

instead manipulate a moving cylinder which is rendered in front of the car. On the road ahead, the driver perceives two such cylinders, which continuously move at a constant longitudinal distance (20 meters) in front of the car. The two cylinders differ only in colour: one is blue and the other one is yellow. The latter is called the reference cylinder, as it moves autonomously cording to an algorithm. The movement direction and the movement speed of the reference cylinder are neither controlled nor predictable by the user, except that the cylinder never exceeds the roadsides. In contrast, the driver controls the lateral position of the blue cylinder with the help of the steering wheel, trying to keep it overlapping with the reference cylinder as well as possible. As the user turns the steering wheel, the controllable cylinder moves to the left or to the right, depending on the direction of the steering wheel and its angular velocity (i.e. the steering wheel controls the cylinder's lateral acceleration). Effectively, this corresponds to a task where the user has to follow a curvy road or the exact lateral position of a lead vehicle, although it is more strictly controlled and thus with less user-dependent variability. Furthermore, there is a traffic light placed on top of the reference cylinder containing two lights: The lower one can be lighted green, whereas the top light shines red when it is switched on. Either none or only one of these lights appears at a time. The

red light requires an immediate brake reaction with the brake pedal, whereas green indicates that an immediate acceleration with the gas pedal is expected. As soon as the driver reacts correctly, the light is turned off". (pp. 2-3).

The task's parameters were configured to provide two levels of task difficulty. In the EASY condition, the speed of the lateral movement of the reference cylinder was 0.4 m/s, while in the DIFFICULT condition, the speed was 1.0 m/s; the maximum speed of the controllable cylinder was 4 m/s. These values were selected after a pilot study involving 10 participants (cf. Engonopoulos, Sayeed, and Demberg (2013) for a study using three levels of difficulty). The subjective ratings (see below) confirmed that the subjects indeed found the DIFF condition more challenging. A pair of participants is shown in Figure 21.3 (left photo: the driver; right photo: the passenger).

(a) Driver (b) Passenger

FIGURE 21.3: Participants in the driving simulation collaborative dialogue study

21.3 PARTICIPANTS

A pilot study was run with 10 participants (all female, all university students). The pilot study subjects participated in the experiment for course credit, as part of their course Phonologie & Prosodie during the academic year 2015-2016, and were not otherwise remunerated.

The main study, reported here, was conducted with 28 participants, in 14 pairs (6 male participants and 22 female). The average age of the main study participants was 22.3 years (standard deviation: 3.2; min: 19; max: 36). Most participants were university students, and were recruited through the participant pool of the Faculty of Psychology at Université catholique de Louvain. All participants reported French as their mother tongue; 23 participants were born and raised in Belgium and 5 participants in Metropolitan France. 26 participants reported knowledge of one or more second languages: 6 had one L2, 20 had two L2s, and 6 had three L2s (First L2: English: 20, Dutch: 5, Italian: 1. Second L2: English: 6, Dutch: Dutch: 9, Spanish: 3, German: 1,

Italian: 1. Third L2: Spanish: 3; Dutch: 1; German: 1; Italian: 1). None of the participants reported any auditory problems and 9 participants had musical training. Regarding their driving experience, 20 participants reported having a driver's licence (obtained on average 2.93 years before the experiment) and 16 participants reported driving regularly. With regards to the relationship between the participants, 16 did not know each other before the experiment; 10 participants were friends; and 2 were in an intimate relationship.

The demographic data was collected by means of a questionnaire, given to both participants at the end of the experiment (both recording sessions). The subjective task difficulty ratings were also collected using the same questionnaire. One participant did not complete the questionnaire.

21.4 DATA COLLECTION

Data from multiple sources was collected throughout the experiment. Both participants were wearing a Sennheiser ME3 head-worn microphone and Philips FX4M headphones. They could listen to each other, and they were relatively isolated from outside noise. Their speech was recorded throughout the experiment. The microphones were connected to a Focusrite 18i8 audio interface (through two microphone preamplifiers), which was used to control the interconnection between the driver's microphone and the passenger's headphones and vice versa; the audio signal was then routed to a Zoom Z24 multi-track recorder. The stimuli were presented using a dedicated laptop PC running Windows; the stimuli were presented through the special interface developed for the SUS phase of the experiment, and using Winamp for the Radio News phase of the experiment. The audio signal of the stimuli presented to the participants was also routed to the Zoom multi-track recorder. Two audio synchronisation signals, from the eye-tracker and the driving simulator, were also connected to the Zoom recorder; finally the internal microphone of the recorder was used to capture ambient noise in the room and to serve as a backup recording. Using this configuration, we collected an 8-track time-synchronised recording of the driver's and passenger's speech, the stimuli presented to the driver and passenger, the events of the eye tracker (start and stop) and of the driving simulator (start, stop, red light, green light, coded as short pure tones with different frequencies) and of any interaction with the experimenter. The audio recordings were performed using a sampling rate of 44.1 kHz and 16-bit resolution.

In addition to the audio recordings, eye tracker data was collected for the driver only, using the portable head-mounted Pupil eye-tracker (Kassner et al., 2014). In this study, we used the second-generation model, recording both eyes at a sampling rate of 60-80 Hz. The OpenDS driving simulator system, which was running on a dedicated laptop Windows PC, also recorded driving behaviour data: the precise time at which the driver pressed the acceleration or brake pedal in response to the traffic lights (and a calculated response

time), as well as steering wheel position and the deviation of the controllable cylinder from the reference cylinder. The driving simulator data were stored in a MySQL database.

21.5 SUBJECTIVE RATINGS OF TASK DIFFICULTY

This section presents the analysis of the subjective ratings of task difficulty, as reported by the participants. The questionnaire used a 7-point Likert scale for task difficulty rating. The tasks were described as presented in Table 21.1.

Task code	Task name
SUS_DrvSlow	Répétition des phrases, en tant que conducteur - conduite lente
SUS_DrvFast	Répétition des phrases, en tant que conducteur - conduite rapide
RListen_DrvSlow	Écouter les nouvelles à la radio, en tant que conducteur - conduite lente
RListen_DrvFast	Écouter les nouvelles à la radio, en tant que conducteur - conduite rapide
RSummary_Drv	Résumer les nouvelles, en tant que conducteur
RDialogue_DrvSlow	Discussion libre, en tant que conducteur - conduite lente
RDialogue_DrvFast	Discussion libre, en tant que conducteur - conduite rapide
Taboo_DrvSlow	Taboo : deviner les mots, en tant que conducteur - conduite lente
Taboo_DrvFast	Taboo : deviner les mots, en tant que conducteur - conduite rapide
SUS_Pass	Répétition des phrases, en tant que passager
RQuestions_Pass	Poser les questions sur les nouvelles de la radio, en tant que passager
RAnswers_Pass	Répondre aux questions sur les nouvelles, en tant que passager
RDialogue_Pass	Discussion libre, en tant que passager
Taboo_Pass	Taboo : faire deviner / décrire les mots, en tant que passager

TABLE 21.1: Tasks performed by participants in Study 4

The results of the subjective ratings are shown in Table 21.2 and Figure 21.4. The most difficult tasks were, as expected, listening to the radio news (comprehension) and summarising them (production). Tasks performed under the fast driving condition were systematically rated more difficult than task performed under the slow driving condition. The passenger's participation in the Radio News and Taboo tasks was rated as the easiest tasks; however, repeating the SUS sentences as a passenger was rated as more difficult than participating in the free exchange dialogue or playing the Taboo game while driving in the slow condition.

Task / Condition	Mean	Std Dev
RListen_DrvFast	5,11	1,52
RSummary_Drv	5,11	1,42
RListen_DrvSlow	4,30	1,30
SUS_DrvFast	4,15	1,10
Taboo_DrvFast	3,37	1,22
RDialogue_DrvFast	3,26	1,35
SUS_DrvSlow	3,19	1,16
SUS_Pass	3,11	1,45
Taboo_DrvSlow	2,81	1,06
RDialogue_DrvSlow	2,74	1,07
Taboo_Pass	2,38	1,18
RAnswers_Pass	1,74	0,97
RQuestions_Pass	1,26	0,52
RDialogue_Pass	1,22	0,42

TABLE 21.2: Summary of subjective ratings of task difficulty

21.6 DATA ANNOTATION AND ANALYSIS

The speech recordings of the driver and passenger were split in sections depending on the task perrformed and the driving simulator condition, using the Ardour audio editor. The following sections were separated: SUS/Driver and SUS/Passenger, EASY and DIFF conditions; Radio News part 1 and part 2 (8 + 8 questions) Driver and Passenger, EASY and DIFF conditions; Taboo game, EASY and DIFF conditions. The total duration of speech recorded (i.e. including both recording channels, the Driver and the Passenger) was 47.6 hours. The SUS task sub-corpus contains 555.6 minutes (9.2 hours) of recordings; the Radio News task sub-corpus contains 1915.9 minutes (31.9 hours) of recordings; and the Taboo Game sub-corpus contains 385 minutes (6.4 hours) of recordings. The total number of section recordings is 442. Due to the large amount of data collected, we focus our analysis only on the Radio News task sub-corpus.

The Radio News task sub-corpus was further processed and its recordings split into corpus samples, using the following breakdown: Summary by the driver; Exchange between the driver and the passenger on the basis of the comprehension questions; Dialogue between the driver and the passenger (each question/topic is a different corpus sample). The sub-corpus contains 657 sections (in 1314 audio files since the driver and passenger speech were recorded in different channels). The total duration of the sub-corpus, after removing the regions of recordings where the driver was listening to the radio news stimuli (and therefore there is no speech to analyse) is 1424.5 minutes (23.7 hours).

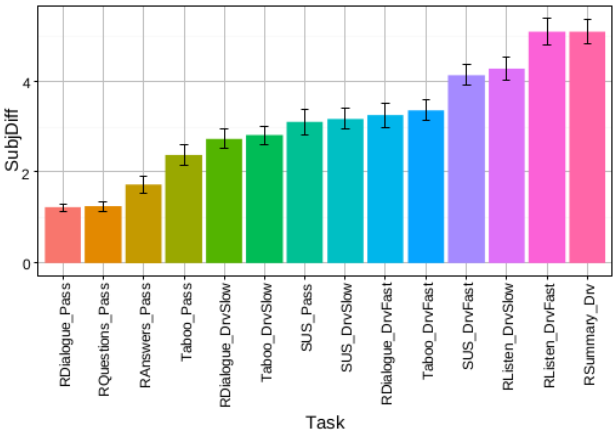


FIGURE 21.4: Subjective difficulty ratings of tasks in this study (mean and standard error)

Samples were selected for manual transcription. The selection was representative of all tasks performed under both conditions (e.g. we selected 8 question-based dialogues for each pair of participants, 4 in the EASY and 4 in the DIFF driving conditions, out of the original 16 dialogues), giving priority to dialogues longer than one minute. This selection of samples was then transcribed under Praat, and the transcriptions imported in Praaline. The final corpus analysed is 8.7 hours long and its contents are presented in Table 21.3.

Condition / Task	Driver or Passenger		Total (both)	
	Count	Duration (min)	Count	Duration (min)
DIFF	116	150,46	176	265,79
Summary	28	35,14	28	35,14
Exchange	27	31,80	54	63,61
Dialogue	61	83,52	122	167,04
EASY	114	144,34	172	256,22
Summary	28	32,46	28	32,46
Exchange	30	35,88	60	71,76
Dialogue	56	76,00	112	152,00
Grand Total	230	294,80 (4,9h)	348	522,01 (8,7 h)

TABLE 21.3: Radio News sub-corpus (manually transcribed and analysed) contents

A cascade of automatic analysis tools was applied: the orthographic transcriptions were converted into phonetic transcriptions and aligned with the speech signal at the phone and syllable level using EasyAlign. A morpho-

syntactic analysis of the corpus was performed using DisMo. The Prosogram script was applied to the entire corpus (using the automatic annotation functionality in Praaline), to obtain detailed prosodic information on each syllable, based on pitch stylisation. The corpus contains 54.002 tokens and 73.381 syllables.

We focus on the temporal characteristics of the speech of the participants, while performing different tasks and under different conditions (driver in the EASY or DIFF driving simulator, passenger). To this effect, we have developped a Praaline plug-in that calculates several temporal measures, per corpus sample and per speaker. The measures per corpus sample (per dialogue) are presented in Table 21.4.

Measure name	Description
TimeTotalSample	Total sample time
TimeSingleSpeaker	Single-speaker time
TimeOverlap	Overlap time
TimeGap	Gap time
RatioSingleSpeaker	Ratio: Single-speaker time / Total sample time (%)
RatioOverlap	Ratio: Overlap time / Total sample time
RatioGap	Ratio: Gap time / Total sample time
GapDurations_Median	Gap duration (median)
GapDurations_Q1	Gap duration (1st quartile)
GapDurations_Q3	Gap duration (3rd quartile)
TurnChangesCount	Number of turn changes
TurnChangesCount_Gap	Number of turn changes, without overlap
TurnChangesCount_Overlap	Number of turn changes, with overlap
TurnChangesRate	Turn changes per minute
TurnChangesRate_Gap	Turn changes per minute, without overlap
TurnChangesRate_Overlap	Turn changes per minute, with overlap

TABLE 21.4: Temporal measures calculated per dialogue (corpus sample)

The temporal measures calculated per sample and speaker are presented in Table 21.5.

We first turn our attention to articulation ratio and silent pause ratio: as can be seen on Figure 21.5, the task associated with the highest cognitive load (S: summary) is associated with the highest silent pause ratio, and the lowest articulation ratio. Among the dialogue tasks, exchanging information on the comprehension questions (E: exchange) has a higher articulation ratio and lower silent pause ratio compared to the free dialogue (Q: questions).

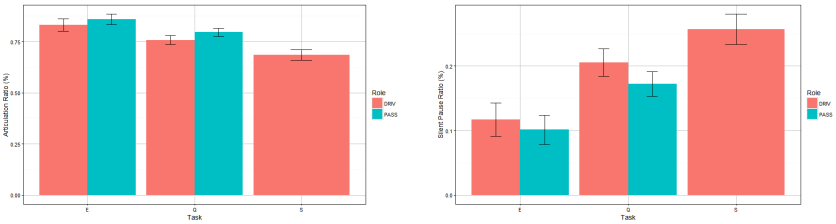
A more detailed analysis of the duration of silent pauses, as a mixture of 2 log-normal distributions is presented in Figure 21.6 and Table 21.6. It can be observed that the long silent pause component distribution has a higher weight factor across tasks (the relevant figures are in boldface); whereas speech produced by the passenger (under no cognitive load) has a more balanced distribution between short and long silent pauses. However, there is no

Measure name	Description
TimeSpeech	Speech time: Articulation + Silent Pause + Filled Pause time (s)
TimeArticulation	Articulation time (s)
TimeArticulation_Alone	Articulation, speaking alone time (s)
TimeArticulation_Overlap	Articulation, overlap time (s)
TimeArticulation_Overlap_Continue	... without turn change, time (s)
TimeArticulation_Overlap_TurnChange	... with turn change, time (s)
TimeSilentPause	Silent pause time (s)
TimeFilledPause	Filled pause time (s)
RatioArticulation	Articulation ratio (%)
RatioArticulation_Alone	Articulation, speaking alone ratio (%)
RatioArticulation_Overlap	Articulation, overlap ratio (%)
RatioArticulation_Overlap_Continue	... without turn change, ratio (%)
RatioArticulation_Overlap_TurnChange	... with turn change, ratio (%)
RatioSilentPause	Silent pause ratio (%)
RatioFilledPause	Filled pause ratio (%)
NumTokens	Number of tokens
NumSyllables	Number of syllables (articulated)
NumSilentPauses	Number of silent pauses
NumFilledPauses	Number of filled pauses
SpeechRate	All articulated syllables (excluding SIL and FIL) / Speech time
SilentPauseRate	Number of silent pauses / Speech time
FilledPauseRate	Number of filled pauses / Speech time
ArticulationRate	All articulated syllables (excluding SIL and FIL) / Articulation time
PauseDur_SIL_Median	Silent pause duration (median)
PauseDur_SIL_Q1	Silent pause duration (1st quartile)
PauseDur_SIL_Q3	Silent pause duration (3rd quartile)
PauseDur_FIL_Median	Filled pause duration (median)
PauseDur_FIL_Q1	Filled pause duration (1st quartile)
PauseDur_FIL_Q3	Filled pause duration (3rd quartile)
TurnDuration_Time_Mean	Mean turn duration (s)
TurnDuration_Syll_Mean	Mean number of syllables in turn
TurnDuration_Token_Mean	Mean number of tokens in turn

TABLE 21.5: Temporal measures calculated per speaker and dialogue

clear correlation between the short/long pause balance and the task that the driver is executing. The task of driving is altering the silent pause distribution coarsely, with higher cognitive load associated with a higher proportion of long pauses in speech. Furthermore, a Wilcoxon rank sum test indicates a significant difference in mean silent pause duration produced by the Driver and the Passenger ($p < 0.001$).

Regarding filled pauses, Figure 21.7 presents two measures calculated per task type (S: summary, E: exchange and Q: dialogue on questions) and speaker



(a) Articulation Ratio (b) Silent Pause Ratio

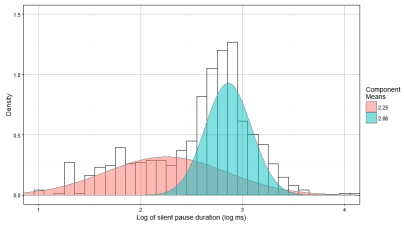
FIGURE 21.5: Articulation Ratio (left) and Silent Pause Ratio (right) per Task and Speaker Role (means and 95% CI)

Task / Condition / Speaker Role	Component weight λ		Mean value μ		Std deviation σ	
	Comp1	Comp2	Comp1	Comp2	Comp1	Comp2
S EASY	25,6%	74,4%	1,824	2,828	0,349	0,288
S DIFF	29,7%	70,3%	1,844	2,838	0,352	0,314
E EASY Driver	20,6%	79,4%	1,743	2,767	0,262	0,325
E DIFF Driver	27,8%	72,2%	1,666	2,748	0,282	0,309
E EASY Passenger	57,3%	42,7%	2,285	2,781	0,620	0,229
E DIFF Passenger	44,5%	55,5%	2,241	2,698	0,557	0,275
Q EASY Driver	17,5%	82,5%	1,610	2,714	0,255	0,340
Q DIFF Driver	20,8%	79,2%	1,723	2,726	0,355	0,341
Q EASY Passenger	60,7%	39,3%	2,362	2,730	0,587	0,269
Q DIFF Passenger	43,2%	56,8%	2,265	2,768	0,583	0,250

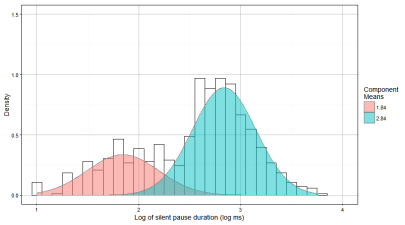
TABLE 21.6: Analysis of silent pause length as a mixture of 2 component log-normal distributions, per task, driving condition and speaker role

role (Driver vs. Passenger): the filled pause ratio, i.e. the percentage of speech time covered by filled pauses, and the filled pause rate, i.e. the number of filled pauses per second. A Wilcoxon rank sum test indicates that filled pause ratio is significantly higher for the driver compared to the passenger ($p < 0.0001$), while filled pause rate is not significantly different for the two speaker roles. In other words, the additional attentional requirements of the secondary task (driving) do not lead to more filled pauses, but to globally longer filled pauses (cf. Figure 21.8: median filled pause length per task and speaker role). This finding is consistent with the previously mentioned observations on silent pauses and articulation ratio.

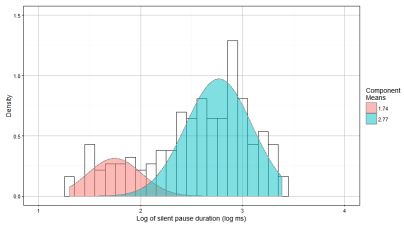
Finally, with respect to dialogue dynamics, Figure 21.9 presents the mean turn duration (in seconds) per task type and speaker role for the two tasks



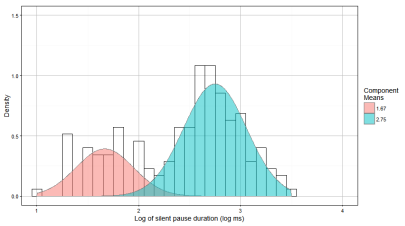
(a) Summary (Driver) Easy



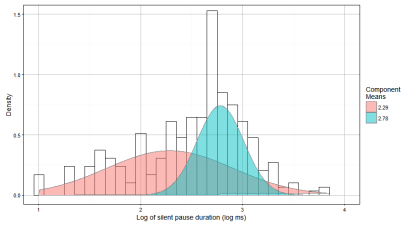
(b) Summary (Driver) Diff



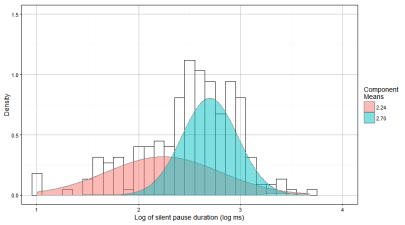
(c) Exchange Driver Easy



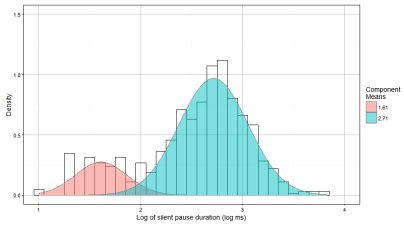
(d) Exchange Driver Diff



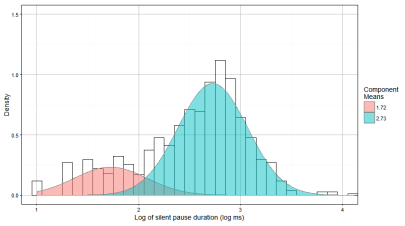
(e) Exchange Passenger Easy



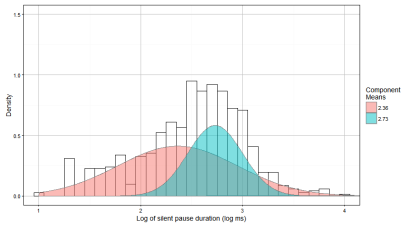
(f) Exchange Passenger Diff



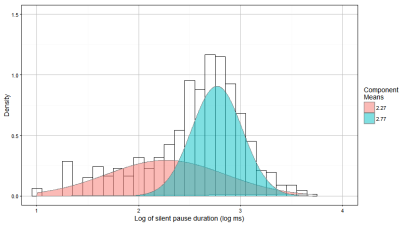
(g) Dialogue Driver Easy



(h) Dialogue Driver Diff

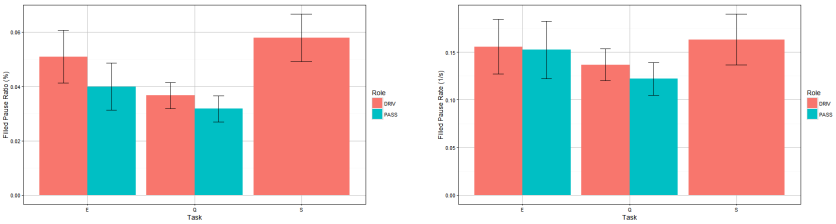


(i) Dialogue Passenger Easy



(j) Dialogue Passenger Diff

FIGURE 21.6: Analysis of silent pause durations as a mixture of 2 log-normal distributions, per task, driving condition and speaker role



(a) Filled Pause Ratio (b) Filled Pause Rate

FIGURE 21.7: Filled pause ratio (left) and filled pause rate (right) per task and speaker role

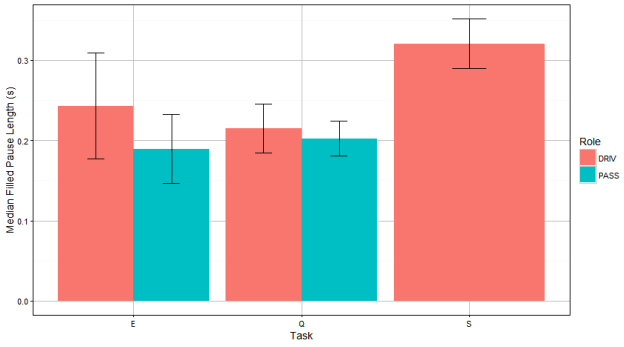


FIGURE 21.8: Median filled pause duration per task and speaker role

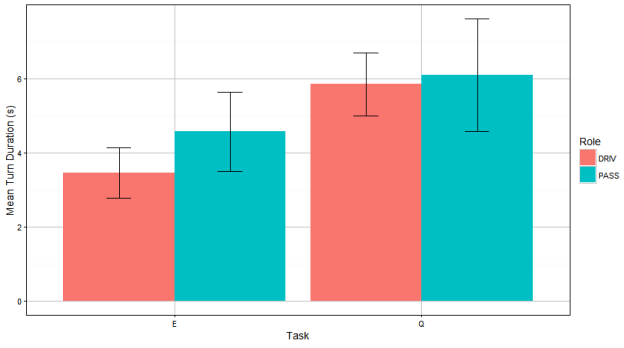


FIGURE 21.9: Mean turn duration for the two dialogue tasks, per speaker role

involving dialogue (E: exchange of factual information on the radio news and Q: open-ended dialogue based on questions). The difference in means is statistically significant between the two task types (Wilcoxon rank sum test,

$p < 0.0001$) but not between driver and passenger in each of the two tasks: this measure is affected by the communicative context with the E task being more dynamic as expected.

22

CONCLUSION

This chapter presents a summary of the main findings from the experimental studies (section 22.1), a list of the main contributions of this thesis (section 22.2) and a discussion on the limitations of our work and perspectives for future research (22.3).

22.1 SUMMARY OF MAIN FINDINGS

We have presented a series of studies on speech produced under different levels of cognitive load. As explained in Section I of the thesis, cognitive load is a multifaceted construct that refers to the demands placed on working memory and attention. Consequently, we have used a variety of tasks, each focusing on a different aspect of cognitive load. Study 1 used standardized cognitive psychology tests (the Stroop test and the Reading Span test); Study 2 demanded a memorisation effort on the part of the participants; Study 3 used simultaneous interpretation as a form of extreme language processing where language comprehension and production are competing for the same attentional resources; and Study 4 used a dual task paradigm to limit the attentional resources available to one of two participants engaging in concurrent language tasks (sentence repetition, language comprehension and memorisation and collaborative dialogue). We studied the speech produced under all these conditions, focusing on phonetic (Study 1) and prosodic (Studies 2, 3, 4) aspects, including the temporal organisation of speech, intonation and phrasing and the relationship between syntactic and prosodic boundaries.

With respect to phonetic measures of speech under cognitive load, our study on French confirmed Yap's (2012) findings (on English): the proportion of time the vocal folds remain closed, as estimated by the Closed Quotient EGG measure increases as cognitive load increases. Glottal source features are sensitive to cognitive load.

With respect to the temporal organisation of speech, all three studies confirmed our hypothesis that an increase of cognitive load leads to a different frequency distribution of silent pause length: silent pauses are more numerous and longer in duration (also expressed as a decrease in articulation ratio). Findings of filled pauses are less clear: while filled pause ratio (proportion

of speech time spent on filled pauses) increases, some speakers do not produce more filled pauses but rather produce longer filled pauses or drawls (hesitation-related lengthening). In high cognitive load conditions, we observed a more variable articulation rate (Studies 2 and 3), i.e. the speakers produced more accelerations and decelerations.

The findings on silent and filled pauses must be interpreted in conjunction with the findings on the prosody-syntax interface. Study 3 indicates that a very high cognitive load condition is associated with frequent mismatches between prosodic and syntactic boundaries: in these cases, silent and filled pauses are inserted inside minor syntactical units (e.g. chunks). We confirm our hypothesis that cognitive load incurs an increase in the number of occurrences of major prosodic boundaries inside minor syntactic units (chunks), i.e. in positions where there are normally not expected. Silent and filled pauses are placed incongruently with the syntactic structure.

With the exception of pitch movements used to denote major prosodic boundaries, we do not detect a systematic change in the mean pitch of speakers under cognitive load. However, Studies 2 and 3 indicate high CL tasks result in less expressive speech (as measured by mean pitch variance and pitch trajectory).

In view of the above findings, we argue that the prosodic phenomena observed under higher levels of cognitive load can be attributed to reduced capacity for speech planning, leading to mismatches between the syntactical and prosodic structure of utterances. The less expressive nature of speech under cognitive load is also potentially due to a reduction in the available capacity for audience design (adapting one's speech to the interlocutor), as more basic aspects of speech planning are prioritised. More research is certainly needed in this direction, involving a wide variety of speech elicitation techniques and several different languages.

22.2 CONTRIBUTIONS

The main contributions of the present thesis are the following:

1. The construction of 3 new French spoken language corpora of speech produced under cognitive load, using a variety of techniques to induce cognitive load, with high-quality recordings of speech and additional performance and physio-psychological data. More specifically:
    - The Cognitive Load Speech with EGG and Eye-Tracking Database (CLSE^2-FR), consisting of recordings of 9 speakers (approximately 50 minutes per participant). Each participant engages in standardised tasks (Stroop test with a dual task, Stroop test with time pressure, Reading span test). This French corpus is designed to be comparable to the Australian English CLSE corpus (Yap, 2012). It is

accompanied by electroglottographic and eye-tracker recordings, and performance data.

- The Question-Answering and Reading Comprehension Corpus, consisting of recordings of 11 speakers (approximately 8 hour of speech in total). Elicited speech includes reading, answering short comprehension questions (argumentative) and summarising. This French corpus is comparable to (Yin et al., 2007) and is accompanied by performance data.
- The Driving Simulator Corpus, consisting of approximately 31 hours of speech, 8h of which are already transcribed and analysed. It contains multi-modal recordings of 28 participants (speech, eye-tracker videos, driving simulator data) engaging in four different tasks (repetition of syntactically unpredictable sentences, listening to and summarising radio news, collaborative dialogue exchanging information and debating, and quick interaction game).

It is hoped that these corpora will contribute to further research in the field of French speech produced under cognitive load.

2. Improvement of methods and automatic tools available for the management, annotation and analysis of spoken language corpora, including:
   - Improvement of the DisMo automatic annotator for the morpho-syntactic analysis of spoken corpora and the automatic annotation of disfluencies.
   - A detailed annotation protocol for disfluencies in spoken corpora, tested and validated on a 7-hour corpus (CPROM-PFC).
   - Promise, a new automatic annotation tool for detecting perceived syllabic prosodic prominence and perceived prosodic boundaries in French spoken corpora.
   - Praaline, a new integrated tool for working with spoken corpora.
   - A new tool to extract and analyse temporal prosodic measures in annotated multi-party dialogue corpora.
3. Findings regarding the temporal organisation of speech, EGG measures and pitch measures from the experimental studies carried out in French confirm the results reported for English in the literature (see also the section on Perspectives).
4. Converging findings from different studies indicate that cognitive load is associated with a mismatch in the packing of prosodic and syntactic units. This is more interesting from a linguistic analysis point of view, despite the fact that is currently difficult to envisage an automatic classification system based on these observations.

22.3 LIMITATIONS AND PERSPECTIVES

The studies presented in this thesis have, of course, their limitations:

- Study 1 should be completed with more data from a larger number of participants, to reach a corpus size equivalent to the corresponding English corpus.
- In Study 2, we observed that the communicative situation did not encourage some participants to produce stretches of speech long enough for meaningful prosodic analysis.
- In Study 3, the data analysed came from only one professional interpreter, in two speaking styles. The results are compatible with Christodoulides (2013), where we studied 6 professional interpreters; however, no control condition was recorded.
- In Study 4, the task of listening to and summarising radio news was not performed without the dual-task (a control condition). The corpus should be completed in this respect.
- Due to the large amount of data collected, we were forced to be very selective in the analyses performed: e.g. the analysis of the prosody-syntax interface could be generally applied across all studies (with the exception of recordings of isolated words).

Given the data collected, the findings and the shortcomings outlined above, we can propose the following perspectives for further research:

1. Using the data from Study 1, investigate the effects of cognitive load on all features that were studied on the comparable English corpus.
2. Using data from Study 1, perform automatic classification and test whether the performance of automatic classification algorithms proposed for English is consistent or differs for French.
3. Perform a series of perceptual experiments, to test whether listeners are aware of the fact that the speaker is under cognitive load, on the basis of stimuli. Research questions include the relative importance of each of the prosodic features found to be affected under CL (e.g. pauses, highly variable speech rate, mismatch of boundaries etc.).
4. In the context of our research on simultaneous interpreting as a form of speech under cognitive load, construct a corpus where professional interpreters are recorded in: (a) a real-world SI task, (b) a controlled SI task, (c) reading and (d) spontaneous speech (4 speaking styles per speaker). Such a corpus would allow the study of individual differences, along with CL.

5. The eye-tracking data from Studies 1 and 4 can be processed to provide cognitive pupillometry measures (e.g. blink rate). This work is already under way.

6. Analyse additional data from Study 4 (non-transcribed dialogues, SUS repetitions and Taboo games).

7. Study 4, which proved to be producing the largest amount of speech data, may be replicated in another language, to allow for cross-linguistic comparisons across the 4 studies.

8. On-line perceptual experiments (similar to the one described in Chapter 15) can test whether cognitive load influences boundary perception (the hypothesis being that under higher CL, subjects will tend to process smaller units).

9. With respect to the automatic tools, several improvements are planned. Future work on DisMo will focus on improving the integration of disfluency detection with the chunking, investigate the performance of chunk-parsing, adding a web interface that will allow users to annotate files without downloading the tool, and improving language coverage. Future work on Promise will focus on providing a gradual perception score, and the possibility for the user to control detection sensitivity for both prosodic prominence and prosodic boundaries. Several improvements are planned for Praaline, as it is gradually released to the community.

In closing, we express the hope that this line of research on speech production and perception under cognitive load will continue to be fruitful, with further studies on French and other languages.

Part IV

APPENDICES

A

MATERIALS USED IN THE STUDIES

A.1 STUDY 1: COGNITIVE LOAD SPEECH WITH EGG AND EYE-TRACKING DATABASE

A.1.1 *Reading text*

The following text was the first one to be presented to the participants. The second text is reproduced as part of Study 2.

La bise et le soleil se disputaient, chacun assurant qu'il était le plus fort, quand ils ont vu un voyageur qui s'avançait, enveloppé dans son manteau. Ils sont tombés d'accord que celui qui arriverait le premier à faire ôter son manteau au voyageur serait regardé comme le plus fort. Alors, la bise s'est mise à souffler de toute sa force, mais plus elle soufflait, plus le voyageur serrait son manteau autour de lui et à la fin, la bise a renoncé à le lui faire ôter. Alors le soleil a commencé à briller et au bout d'un moment, le voyageur, réchauffé a ôté son manteau. Ainsi, la bise a du reconnaître que le soleil était le plus fort des deux.

A.1.2 *Reading span test*

The following sentences were presented in the reading span test trials. Logical sentences have been selected from Desmette et al. (1995), and we have constructed corresponding illogical sentences.

Un jeune garçon débordant d'énergie multipliait les sauts dans la salle de sport.
Les élèves ne contenaient plus leur fluorescence lorsqu'approcha l'heure de la leçon.
Le battement de ses deux pieds dans l'eau laissait derrière lui une légère mouche.
Il tourna la tête, comme par instinct, et inspecta les visages inquiets de ses ouvriers.
Dès que le calme fut revenu, la tortue sortit la tête de sa carapace.
Il rangea ses zèbres dans des caisses en carton qui avaient contenu des salades.
Il se fit passer pour un malheur en civil et entra avec beaucoup de facilité.
Gravement blessé, le bandit est cloné sur le trottoir, les mains crispées sur le ventre.
Il revenait du bord de la chaudière où il avait empli ses poches de petits cailloux.
Ces sacs de romance étaient précieux car ils représentaient vraiment leur seul avenir.
Le battement de ses deux pieds dans l'eau laissait derrière lui une légère mousse.
Elle attendit que la poire se lève et put alors admirer le lac et la vallée.
Ses nerfs étaient tendus car on lui avait répété que son bibliothécaire était en forme.

Le vendeur rassembla tous ses instruits et les rangea soigneusement dans leur boîte.
Le nouveau mécanicien lui avait conseillé de vérifier plus souvent le niveau d'huile.
Les élèves ne contenaient plus leur impatience lorsqu'approcha l'heure de la leçon.
Le cartable réunit les derniers dossiers et les porta dans le bureau de son chef.
Elle l'embrassa avec fureur et le laissa comme pétrifié sur le divan.
Le tonnerre grondait sauvagement sur le chantier et la foudre tomba sur le toit.
Cette cuisson était étrange, mais cela valait mieux que de dormir dans la voiture.
Assis sur un banc, le visiteur rédigeait un prêtre tout en mangeant du chocolat.
L'incompétence des directeurs est souvent à l'origine de graves problèmes.
Le navire se mit à prendre la parole car sa coque avait été percée par un rocher.
La porte s'ouvrit et le jeune officier au visage glacé entra dans la demeure.
Il fut ainsi assoupli hors de notre petit pays, au milieu d'une foule de bourgeois.
A cet instant, la vitre fut brisée par un projectile qui rebondit sur le tapis.
Il donnait l'impression d'avoir clignoté, comme s'il avait renoncé au succès.
Dès que le calme fut revenu, la tortue sortit la tête de sa dédicace.
D'une main qui ronflait, il se coiffa puis, sans détourner la tête, gagna la sortie.
Il était un peu moins de midi trente quand son sommeil sortit lentement de de l'abri.
Le ciel prit de délicates teintes pâles et le lardon glissa lentement vers l'horizon.
À ce moment, le train entra en guerre et mon ami abandonna sa lecture avec regret.
Il fut ainsi conduit hors de notre petit pays, au milieu d'une foule de bourgeois.
Le conducteur portait un casque et une combinaison de cuir mais jamais de lunettes.
Ce jeune aventurier est parti en Amérique et il a vite fait fortune.
Le lendemain, leur maison fut cernée par des individus rôtis comme des hercules.
Elle se leva avec nonchalance et dit à son manteau qu'il était un ivrogne.
Un barbier s'approcha tout en aiguisant un vieux rasoir sur le cuir de son pantalon.
Elle avait l'impression d'être constamment épiée et éprouvait un profond malaise.
Le nouveau mécanicien lui avait conseillé de tuméfier plus souvent le niveau d'huile.
Il revenait du bord de la rivière où il avait empli ses poches de petits cailloux.
Le malade se pencha à l'oreille de son voisin pour lui confier ses angoisses.
Il a passé ses vacances à chercher les labradors enfouis au plus profond de l'océan.
Le février cachait ses mains comme si une saleté s'était glissée sous ses ongles.
Il a passé ses vacances à chercher les trésors enfouis au plus profond de l'océan.
Le navire se mit à prendre l'eau car sa coque avait été percée par un rocher.
Le lendemain, il a retrouvé la clef de sa maison parmi les verres et les bouteilles.
Le raton-laveur portait un casque et une combinaison de cuir mais jamais de lunettes.
D'une main qui tremblait, il se coiffa puis, sans détourner la tête, gagna la sortie.
L'aumônier du bretzel a beaucoup parlé aux religieuses du sens de la vertu.
Le flacon contenait un liquide bleu clair et frais, au goût de chute et de bataille.
Un étranger apparut sur le râle et tendit à la fille un petit sac de jouets.
Le notaire a donné rendez-vous à son client vendredi prochain en fin de matinée.
Devant ce spectacle, même l'homme le plus brave aurait un moment de faiblesse.
Elle attendit que la brume se lève et put alors admirer le lac et la vallée.
Ce mauvais ventre reproduisait le même coucher de soleil sur toutes ses toiles.
Le carton se pencha à l'oreille de son voisin pour lui confier ses angoisses.
Elle rassembla ses maigres affaires et partit sans se retourner loin de ce village.
Ce jeune aventurier est parti en Amérique et il a vite fait le ménage.
Sur le bord du quai, un clochard rêvait au plaisir qu'il aurait de boire une prière.
Le ciel prit de délicates teintes pâles et le soleil glissa lentement vers l'horizon.

Le dîner achevé, nous rangions les chaises et notre mère berçait la vaisselle.
À peine fut-il entré chez le dentiste qu'il ressentit un pénible sentiment.
Ce mauvais peintre reproduisait le même coucher de soleil sur toutes ses toiles.
Ses nerfs étaient tendus car on lui avait répété que son adversaire était en forme.
Elle avait l'impression d'être constamment épiée et éprouvait un profond vinaigre.
Le cheval blanc hennit, allongea l'encolure et partit au galop à travers la plaine.
Après le déjeuner, il étudiait dans un tigre qu'il avait emprunté à son maître.
Il conservait des objets étranges et délabrés qui n'étaient plus d'aucune utilité.
Les alarmes qui ont perquisitionné les lieux ont trouvé de l'argent et des armes.
Le notaire a donné rendez-vous à son frigo vendredi prochain en fin de matinée.
Le maître d'hôtel frappa dans les mains et demanda au garçon d'apporter le rôti.
Gravement blessé, le bandit est tombé sur le trottoir, les mains crispées sur le ventre.
Parmi ces acteurs, il y en avait un dont la prestation était d'une grande qualité.
Le maître d'hôtel frappa dans les mains et demanda au glaçon d'apporter le rôti.
Ils leur apprenaient à pincer l'obscurité et à ne pas craindre la solitude.
La nappe était mise sur la table et les assiettes brillaient comme des apprenants.
L'homme en vêtement de combat se redressa et s'adossa tranquillement au mur.
L'homme regardait discrètement son invité et sut alors qu'il avait atteint son but.
La porte s'ouvrit et le jeune officier au collage glacé entra dans la demeure.
Sans rien dire, il s'asseyait près du feu et buvait un alcool de poires.
Le lendemain, il a empoisonné la clef de sa maison parmi les verres et les bouteilles.
Au cours des derniers mois, ce parrain sablonneux est devenu le dépotoir de l'île.
Un barbier s'approcha tout en guérissant un vieux rasoir sur le cuir de son pantalon.
L'aumônier du carmel a beaucoup parlé aux religieuses du sens de la vertu.
Il se remit à marcher et cette fois, il était certain qu'il parviendrait au sommet.
Les poubelles du restaurant étaient souvent le point de ralliement des pauvres.
Rappelé par un télégramme obscur, le baron fit le plus pénible des cauchemars.
La bigote se dépêcha de boire le vin et de mettre l'hostie dans sa bouche.
L'incompétence des directeurs est souvent à l'origine de longues manches.
Sur le dos de l'homme était attaché un dispositif ressemblant à une plage.
La jeune fille est partie en chatouillant sa voiture devant le magasin.
Pour dégager ses pieds de ce piège, il se résolut à abandonner ses chaussures.
Après avoir dévasté le village, le vieux pirate ordonna de gagner en maturité.
Les soirs d'été, nous aimions manger dans le jardin, en compagnie des baleines.
La petite baraque en planches et au toit vitré nous protégeait tous de la tempête.
Elle se leva avec nonchalance et dit à son ami qu'il était un ivrogne.
Elle regagna le balcon de son appartement pour y admirer les toits de la cité.
La nappe était mise sur la table et les assiettes brillaient comme des diamants.
La prise de vue n'était pas bonne mais cela ne diminuait en rien son éclat.

A.2 STUDY 2: QUESTION-ANSWERING AND READING COMPREHENSION MONOLOGUE

A.2.1 *Phase 1: Short Texts and Comprehension Questions*

The texts were reproduced from Hetru, E. & Bizeur, J.-L., Livre QCM de raisonnement verbal, 2012 edition, ORSEU, ISBN: 978-2-918796-09-1.

TEXTE 1

Une équipe du Centre de Développement Prélangagier de l'Université de Würzburg a étudié les mélodies formées par les cris et les pleurs de 30 bébés français et de 30 bébés allemands peu après leur naissance. Il apparaît que les nouveau-nés français insistent sur la fin des cris, comme les francophones insistent sur la fin des mots et des phrases, tandis que les petits Allemands font exactement l'inverse, copiant le phrasé germanophone. Il ne s'agit bien sûr pas d'un accent à proprement parler, puisque celui-ci concerne la façon dont les mots sont prononcés, mais cette étude confirme l'importance que revêtent les mélodies dans le processus d'apprentissage d'une langue, et montre clairement que celui-ci commence dans l'utérus, par la perception puis la reproduction des mélodies entendues.

Question : Laquelle des affirmations suivantes est correcte?

(A) L'apprentissage d'une langue commence avant la naissance
(B) Les chercheurs ont constaté que les nouveau-nés francophones insistaient sur la fin des mots
(C) Les nouveau-nés français crient plus fort que les nouveau-nés germanophones
(D) L'étude menée par les chercheurs de l'Université de Würzburg vise à mieux comprendre l'origine des accents.

TEXTE 2

À la Plagne, entre Lyon et Genève, ont été découvertes les plus grandes traces de dinosaures connues à ce jour – pouvant atteindre 1,20 à 1,50 mètres de diamètre et formant une véritable piste courant sur des centaines des mètres. Elles ont été identifiées par deux chercheurs de l'université Claude Bernard à Lyon comme les empreintes de sauropodes. Ces animaux mesuraient plus de 25 mètres de long et pouvaient peser jusqu'à une quarantaine de tonnes. Ce ne sont cependant pas les scientifiques qui ont fait cette découverte, mais deux amateurs passionnés de paléontologie, de fossiles, de volcans et de nature.

Question : Laquelle des affirmations suivantes est correcte?

(A) Les empreintes de sauropodes ont été découvertes par deux chercheurs de l'université Claude Bernard
(B) Les sauropodes pesaient en moyenne 40 tonnes
(C) Ceux qui ont découvert les traces de dinosaures ne savaient probablement pas qu'il s'agissait d'empreintes de sauropodes
(D) Les sauropodes sont les plus grands dinosaures connus à ce jour

TEXTE 3

Une équipe de chercheurs européens a repéré dans la constellation de la Balance l'exoplanète la plus légère jamais identifiée. Cette planète extrasolaire, baptisée Gliese 581-e par les astronomes, a une masse équivalente à environ deux fois celle de la Terre et se situe à 20,5 années-lumière de celle-ci. Si Gliese 581-e est probablement une planète rocheuse, se proximité avec l'étoile autour de laquelle elle orbite n'en fait pas une planète de la zone habitable définie par les astronomes. Cette zone correspond à

une région de l'espace où l'eau peut exister à l'état liquide. Le but de cette chasse aux exoplanètes est de parvenir à identifier une planète aux conditions environnementales similaires à celles de notre planète bleue et ainsi susceptible d'arbitrer la vie.
Question : Laquelle des affirmations suivantes est correcte?

(A) Il n'y a pas d'eau sur Gliese 581-e
(B) Les scientifiques ne connaissent pas de planète plus légère que Gliese 581-e
(C) Le but de la chasse aux exoplanètes est de trouver une planète sur laquelle nous pourrions habiter
(D) On ne connaît pas précisément la composition de Gliese 581-e

TEXTE 4

Les diodes électroluminescentes organiques (OLED) sont des semi-conducteurs qui émettent de la lumière lors du passage d'un courant électrique. Destinées à terme à remplacer les technologies LCD et plasma, les OLED sont fines et souples, ce qui les rend adéquates à la fabrication d'écrans enroulables. Elles pourraient même être développées pour être utilisées à des fins d'éclairage conventionnel, concurrençant ainsi les ampoules incandescentes et les tubes fluorescents. Mais jusqu'ici, les OLED les plus efficaces du point de vue énergétique atteignaient seulement 44 lumens de lumière par Watt consommée, alors que les tubes fluorescents conventionnes ont une efficacité de 60 à 70 lumens par Watt. Récemment, des chercheurs de l'université de Dresde ont réussi à créer des OLED aussi efficaces que les tubes fluorescents traditionnels. Pour ce faire, ils ont associé un concept innovant de couche d'émission à haute efficacité énergétique avec des concepts à découplage de lumière.
Question : Laquelle des affirmations suivantes est correcte?

(A) Les OLED ont une efficacité énergétique inférieure à celle des tubes fluorescents conventionnels
(B) L'efficacité énergétique d'un OLED ne dépasse pas 44 lumens par Watt.
(C) Les OLED remplacent peu à peu les technologies LCD et plasma.
(D) Jusqu'ici, les OLED ne pouvaient pas être utilisées pour l'éclairage conventionnel en raison de leur faible efficacité énergétique.

TEXTE 5

La résistance aux antibiotiques constitue un problème grandissant dans le monde. En 2007, 10% des Streptococcus pneuomoniae isolés étaient insensibles à la pénicilline dans 30 pays. La proportion de patients européens se voyant prescrire des antibiotiques par le médecin généraliste pour une infection de la partie inférieure de la trachée varie d'environ 27% au Pays-Bas à 75% au Royaume-Uni. Les études montrent que la plupart de ces prescriptions n'aident pas le patient à aller mieux ou à se rétablir plus rapidement. Elles engendrent ainsi un gaspillage de ressources, exposent inutilement les patients au risque d'effets secondaires et favorisent le développement d'organismes résistants. C'est pourquoi une meilleure standardisation des pratiques permettrait d'améliorer la qualité des soins de santé.
Question : Laquelle des affirmations suivantes est correcte?

(A) Trois quarts des britanniques se voient prescrire des antibiotiques par leur médecin généraliste
(B) Les antibiotiques sont inefficaces contre les Streptococcus pneuomoniae
(C) Les infections de la partie inférieure de la trachée sont plus fréquentes au Royaume-Uni qu'aux Pays-Bas
(D) Les antibiotiques sont le plus souvent inefficaces en cas d'infection de la partie inférieure de la trachée

Texte 6

Les victimes d'accidents, de guerres ou de viol, développent souvent un syndrome de stress post-traumatique, une affection caractérisée par d'irrépressibles flash-back. Des psychiatres ont eu l'idée de contrer ce symptôme à l'aide d'un jeu vidéo : ils ont exposé des volontaires à une série de séquences vidéo violentes puis invité une partie du groupe à faire une partie de Tétris. La semaine suivante, les chercheurs ont conduit une série d'examens destinées à mesurer l'impact des films sur les participants. Les joueurs de Tétris manifestaient moins de flash-back et obtenaient de meilleurs résultats au test utilisé par les psychiatres pour évaluer l'ampleur du syndrome de stress post-traumatique d'un patient. Les psychiatres affirment que les jeux vidéo du genre Tétris épuiseraient les ressources visuo-spatiales du cerveau, amoindrissant au passage la mémorisation de la composante visuelle associée à un évènement. La technique ne constituerait pas une réponse globale au syndrome de stress post-traumatique, mais elle pourrait s'avérer utile au cours de la prise en charge immédiate de victimes d'incidents violents.

Question : Laquelle des affirmations suivantes est correcte?

(A) Le fait de jouer à Tétris peu de temps après un incident violent réduirait la fréquence des flash-back liés au syndrome de stress post-traumatique.
(B) Tétris est un jeu conçu par des psychiatres pour diminuer certains symptômes du stress post-traumatique
(C) Le stress post-traumatique diminue les ressources visuo-spatiales du cerveau
(D) Des psychiatres ont mené une expérience destinée à mieux comprendre l'impact des jeux vidéo sur la capacité de mémorisation

Texte 7

Pour résoudre un problème, soyez détendu et ouvert d'esprit. C'est ce que démontre l'étude menée par des chercheurs de l'Université de Londres. Ils ont analysé les pulsations électriques produites par les neurones de 21 volontaires soumis à la résolution de problèmes. En se basant sur les électroencéphalogrammes récoltés, les scientifiques ont mis en évidence que les ondes gamma – de fréquence supérieure à 30 Hz – générées par le cerveau sont d'autant plus importantes que le sujet s'entête face à un problème qu'il ne peut résoudre. Bloqué dans cette impasse mentale, il est moins enclin à ouvrir son esprit pour restructurer le problème posé ou même exploiter l'indice qu'on lui communique. À l'inverse, les personnes détendues produisent des ondes alpha – fréquence comprise entre 8 et 13 Hz – et se montrent capables de trouver la solution en abordant le problème sous différentes angles.

Question : Laquelle des affirmations suivantes est correcte?

(A) Votre cerveau a tendance à produire davantage d'ondes gamma si vous ne parvenez pas à trouver la réponse à cette question et si vous ne vous décidez pas à passer à la question suivante.
(B) Une personne qui trouve une solution à un problème est détendue et ouverte d'esprit.
(C) Des chercheurs de l'Université de Londres ont montré que plus le sujet était détendu, plus son cerveau émettait des ondes de haute fréquence
(D) Selon des chercheurs de l'Université de Londres, une personne qui ne trouve pas une solution à un problème a un esprit étroit.

Texte 8

Si un porc s'aventurait devant un miroir, il foncerait probablement vers ce qu'il croit être un congénère, serait-on tenté de croire. D'après des chercheurs en anthropo-

zoologie, il n'est rien, du moins si l'on s'arme de patience. Les scientifiques ont placé huit porcs deux par deux pendant cinq heures dans un enclos contenant un miroir. Les animaux commencèrent par s'énerver devant leur reflet, brisant à chaque fois le miroir. Mais après quelques heures, ils semblaient comprendre que le miroir ne faisait que renvoyer l'image de leur environnement. Pour le vérifier, les chercheurs placèrent alors un bol rempli de nourriture dans l'enclos en prenant soin d'installer un ventilateur pour brouiller leur odorat. Le bol était dissimulé derrière un écran, mais en revanche visible dans le miroir. Tous les cochons, sauf un, se sont dirigés vers le bol réel en une vingtaine de secondes.
Question : Laquelle des affirmations suivantes est correcte?

(A) Les scientifiques ont observée comportement de huit porcs pendant cinq heures
(B) Au début de l'expérience, les porcs ne comprennent pas que le miroir reflète leur propre image
(C) Les procs n'ont pas utilisé leur vue pour se diriger vers le bol contenant de la nourriture
(D) L'un des porcs s'est dirigé vers le miroir au lieu de foncer vers le bol rempli de nourriture

TEXTE 9

La civilisation de Caral, considérée comme la plus ancienne d'Amérique, a brusquement disparu il y a 3600 ans. Installée sur la côte du Pacifique à 180 km au nord de Lima au Pérou, cette civilisation a prospéré durant 2000 ans. Outre d'impressionnantes constructions pyramidales, les sites archéologiques de Caral recèlent un vaste réseau d'irrigation agricole. Une étude suggère que la conjonction d'un tremblement de terre avec des variations climatiques induites par El Niño serait à l'origine du déclin des Carals. Il y a 3600 ans, de violentes secousses sismiques ont déstabilisé la chaîne de montagnes dominant la vallée où ils vivaient, entraînant l'accumulation de fragments de roches en contrebas. Des pluies torrentielles saisonnières engendrées par El Niño emportèrent ensuite ces débris rocheux jusqu'à l'océan. Ce dernier le rejeta sous forme de sable et de limon, le tout aboutissant à la formation d'une grande crête isolant les baies fertiles. Sous l'influence des vents, le sable aurait envahi les champs et englouti les systèmes d'irrigation, affamant ainsi les Carals.
Question : Laquelle des affirmations suivantes est correcte?

(A) La disparition de la civilisation de Caral il y a 3600 ans serait due à la famine
(B) Les Carals ont connu leur âge d'or entre 5600 et 3600 ans avant JC
(C) La vallée dans laquelle vivaient les Carals a subi un tremblement de terre il y a 3600 ans
(D) Les plus grandes pyramides d'Amérique se trouvent au Pérou

TEXTE 10

En 2008, près de 1,38 million de femmes ont été diagnostiquées avec un cancer du sein dans le monde entier, ce qui représente près d'un dixième des nouveaux cancers et approximativement le quart des cas de cancers touchant les femmes. Les taux de cancer du sein les plus élevées se trouvent principalement en Europe, soit 332 000 nouveaux cas pour l'Union européenne en 2008. La méthode de détection habituelle est la mammographie qi, selon les chercheurs, fonctionne très bien pour les femmes de plus de 50 ans et apporte des résultats avec près de 95% de précision. Mais ils soulignent que la mammographie est moins efficace chez les femmes plus jeunes pour un taux de détection de 60% pour les femmes en-dessous de 50 ans, qui représentent 20% des cas de cancers du sein.

Question : Laquelle des affirmations suivantes est correcte?
(A) Dans le monde en 2008, près d'un cancer sur 10 était un cancer du sein
(B) 20% des femmes de moins de 50 ans ont un cancer du sein
(C) La mammographie se révèle inefficace chez les femmes de moins de 50 ans
(D) Le taux de détection du cancer du sein par mammographie est supérieur à 60%

A.2.2 *Phase 2: Narrative Text*

The symbol ‖ is used to denote the passages of text presented on the same screen. Subjects pressed the space-bar key to move to the next passage.

Le Premier Ministre ira-t-il à Beaulieu? ‖1 Le village de Beaulieu est en grand émoi. ‖ Le Premier Ministre a en effet décidé de faire étape dans cette commune au cours de sa tournée de la région en fin d'année. ‖ Jusqu'ici les seuls titres de gloire de Beaulieu étaient son vin blanc sec, ses chemises en soie, un champion local de course à pied (Louis Garret), quatrième aux jeux olympiques de Berlin en 1936, et plus récemment, son usine de pâtes italiennes. ‖ Qu'est-ce qui a donc valu à Beaulieu ce grand honneur? ‖ Le hasard, tout bêtement, car le Premier Ministre, lassé des circuits habituels qui tournaient toujours autour des mêmes villes, veut découvrir ce qu'il appelle « la campagne profonde ». ‖ Le maire de Beaulieu – Marc Blanc – est en revanche très inquiet. ‖ La cote du Premier Ministre ne cesse de baisser depuis les élections. ‖ Comment, en plus, éviter les manifestations qui ont eu tendance à se multiplier lors des visites officielles ? ‖ La côte escarpée du Mont Saint-Pierre qui mène au village connaît des barrages chaque fois que les opposants de tous les bords manifestent leur colère. ‖ D'un autre côté, à chaque voyage du Premier Ministre, le gouvernement prend contact avec la préfecture la plus proche et s'assure que tout est fait pour le protéger. ‖ Or, un gros détachement de police, comme on en a vu à Jonquière, et des vérifications d'identité risquent de provoquer une explosion. ‖ Un jeune membre de l'opposition aurait déclaré: ‖ « Dans le coin, on est jaloux de notre liberté. S'il faut montrer patte blanche pour circuler, nous ne répondons pas de la réaction des gens du pays. Nous avons le soutien du village entier. » ‖ De plus, quelques articles parus dans La Dépêche du Centre, L'Express, Ouest Liberté et Le Nouvel Observateur indiqueraient que des activistes des communes voisines préparent une journée chaude au Premier Ministre. ‖ Quelques fanatiques auraient même entamé un jeûne prolongé dans l'église de Saint Martinville. ‖ Le sympathique maire de Beaulieu ne sait plus à quel saint se vouer. ‖ Il a le sentiment de se trouver dans une impasse stupide. ‖ Il s'est, en désespoir de cause, décidé à écrire au Premier Ministre pour vérifier si son village était vraiment une étape nécessaire dans la tournée prévue. ‖ Beaulieu préfère être inconnue et tranquille plutôt que de se trouver au centre d'une bataille politique dont, par la télévision, seraient témoins des millions d'électeurs.

A.2.3 *Phase 3: Argumentative Text and Distractors*

Il y a eu d'énormes évolutions sur les devises. ‖ Depuis que les Japonais sont entrés dans ce jeu, depuis qu'ils sont entrés massivement dans ce jeu, même, il y a un nouveau choc des liquidités, pourrait-on presque dire. ‖ Ou en tout cas, des signes très clairs que la position des devises faibles s'est intensifiée. ‖ Et les Japonais veulent maintenir leur devise à un niveau faible. ‖ Ces énormes flux de liquidités qui

viennent d'Asie arrivent surtout sur les marchés obligataires européens. || Certains profitent de la faiblesse de leur devise nationale. || Et ils essaient de trouver de meilleurs rendements. || Ils les trouvent surtout dans les pays industriels, pays européens, etc. || Si vous examinez les flux de devises, les mouvements, les principaux acheteurs de titres de la dette des pays, d'Italie, Espagne, Europe du Sud, se trouvent surtout en Asie et au Japon. || Toutes les grandes puissances monétaires ont une politique à peu près similaire. || Nous avons parlé des États-Unis. Les Britanniques font un peu la même chose. || La banque centrale canadienne a surpris tout le monde en laissant tomber l'augmentation de ses taux pour laisser filer sa devise. || Ce qui veut dire que toutes les grandes puissances ont pris des mesures pour ne pas trop laisser croître leur devise. || La seule qui n'a pas agi ainsi, c'est la zone euro. || Les politiciens chez nous se félicitent du resserrement de l'écart de taux. || À court terme, c'est une bonne chose, évidemment. || On pense que la crise de la zone euro est terminée. || Mais à moyen terme, les conséquences peuvent être énormes. || Si vous examinez l'évolution de l'euro par rapport aux autres devises, y compris des pays qui sont directement concurrents de l'Europe pour les exportations, vous constatez que le problème va se montrer très fort chez les entreprises dès ce trimestre ou le trimestre prochain. || Examinez un peu la situation de Fiat ou Peugeot par rapport à Mazda, Suzuki, etc., qui ont des coûts beaucoup plus bas, et vous comprendrez. || Et selon moi, cela va encore poser beaucoup de problèmes à la zone euro, malgré les signes encourageants qu'on a pu constater. || Alors, revenons aux États-Unis. Les programmes de rachat de titres de la dette des grands pays ont pour conséquence que le plus grand créancier des États-Unis, ce n'est plus la Chine ou même le Japon. || Ce sont les États-Unis eux-mêmes, leur propre banque centrale, qui détient le plus de titres de dette américains. || Dans une certaine mesure, c'est la même chose aussi au Royaume-Uni, puisque 30% se trouvent sur les comptes de la banque centrale. || Et donc, on peut peut-être remettre en cause l'indépendance des banques centrales. || Alors, voyons un peu les effets que cela peut avoir. || On dit qu'il y a des césures dans la structure, qu'il faut tout repenser, effacer tout ce qu'on a appris, qu'on ne peut plus utiliser les connaissances dont on dispose. || Et je pense vraiment, effectivement, qu'on vit dans un monde tout à fait nouveau. || Et ce, pas depuis hier, mais, en fait, depuis 2009, quand les États-Unis ont commencé à mener cette politique. || Les plus grands acteurs qui sont présents sur le marché des devises, les banques centrales, se sont engagées, se sont engagées vis-à-vis du facteur correcteur le plus important pour une économie orientée sur le marché, c'est-à-dire les taux d'intérêt. || Ils ont été manipulés. Ils ne répondent plus à la demande et à l'offre, mais sont manipulés politiquement par les banques centrales, ou sont orientés dans un sens ou dans l'autre, souvent vers le bas. || Et si vous examinez toutes les décisions des banques centrales, ces derniers temps, les seules tâches de ces banques centrales depuis 2009, on constate alors que le seul but, c'est de maintenir la capacité de paiement des États. || Le but n'est absolument plus de stabiliser les prix. Bon, c'est vrai, ce n'est pas à l'ordre du jour étant donné la faiblesse de l'économie de bien des États, pour le moment, l'inflation. || Mais on constate tout de même que c'est un grand changement. || Parce que, par exemple, le quantitative easing aux États-Unis ou bien les mesures de la banque centrale anglaise, la BCE, de la banque centrale japonaise, toutes ces mesures qui sont prises par les banques centrales, elles ont pour seul but de réduire les montagnes de dette des États. || Et nous allons continuer dans cette voie dans le monde.

DISTRACTORS: the following strings of numbers, spoken by a native French speaker, were used as distractors, while participants were answering the reading comprehension questions related to the text in Phase 3: 12_25_09, 15_49_32, 27_28_96, 56_78_32, 64_57_12, 71_41_59, 78_77_64, 84_42_21, 88_77_33, 96_63_27.

A.3 STUDY 4

A.3.1 *Phase 1: Syntactically Unpredictable Sentences*

The following syntactically unpredictable sentences were used in the perception study. The audio used for the study comes from the SUS Calibrated Audio corpus described in Boula de Mareüil et al. (2006) and in Raake & Katz (2006).

L	N	Sentence	L	N	Sentence
1	1	La loi brille par la chance creuse.	15	1	Le genre fonce sur la buée muette.
1	2	La classe gaie montre le frein.	15	2	Le bol coupe la loi qui tarde.
1	3	Quand le lien signe-t-il l'onde pleine ?	15	3	Comment le verre signe-t-il le grain rude ?
1	4	Le test clair mange la haine.	15	4	Le rite pose la femme qui brille.
1	5	L'or jaune porte le dôme.	15	5	Le sens fixe la fée qui court.
1	6	Comment la soif lance-t-elle le bol proche ?	15	6	Quand le coup plonge-t-il le drap vrai ?
1	7	Le mur siffle la buée qui vole.	15	7	La bête morte nomme la brume.
1	8	La banque dit la dinde qui plaît.	15	8	La rue entre contre le nuage droit.
1	9	La terre dresse la boîte qui rage.	15	9	Où la vache casse-t-elle la soif plate ?
1	10	Où l'œuf cite-t-il le thé doué ?	15	10	La part fuit dans le vent froid.
1	11	Le nom luit sur le bras nu.	15	11	La joie ronde ose la crainte.
1	12	Le choix tape dans la queue close.	15	12	La gare pâle baisse le crime.
2	1	La main pose le son qui souffre.	16	1	La vigne rose fonde la sœur.
2	2	Quand le hall règle-t-il le coin vide ?	16	2	Le hall brun peint la pierre.
2	3	Comment le feu tourne-t-il l'ange roué ?	16	3	La chance pue contre le deuil sec.
2	4	La sœur douce tue la pente.	16	4	Quand la tasse sert-elle la peine muette ?
2	5	Le mont pense pour la nuit grave.	16	5	Comment le son quitte-t-il la main vide ?
2	6	Le trou dort sous la gamme rouge.	16	6	Le bras parle par l'ours pâle.
2	7	Le jean drôle ouvre le poste.	16	7	La lune pointe dans la boîte seule.
2	8	Comment le fleuve prend-il le gaz digne ?	16	8	Le gaz presse l'or qui chasse.
2	9	La peine craque sans la bière brune.	16	9	La caisse droite donne la plage.
2	10	Le plat seul jette la croix.	16	10	Le lit cite la clef qui rêve.
2	11	La plaine manque le truc qui pile.	16	11	Où le gué dicte-t-il le trou fade ?
2	12	L'œil ruine la cause qui part.	16	12	La vie trace la pierre qui rit.
7	1	Le drame entre sans la plaine cuite.	17	1	Le feu tombe contre le film sale.
7	2	Le mal pense contre le mec proche.	17	2	La jupe riche manque la croix.
7	3	Où le mot loge-t-il le champ noir ?	17	3	La peine cache le hall qui naît.
7	4	Comment l'ange tire-t-il le feu doué ?	17	4	Le frère gai pique la preuve.
7	5	Comment le jean sonne-t-il le bois clair ?	17	5	Le mont rampe pour la bière morte.
7	6	Le vin lutte pour le rang bleu.	17	6	La feuille douce ruine le riz.
7	7	Le chêne blond juge la croix.	17	7	Comment l'oeil baisse-t-il le son roux ?
7	8	Le plat marque la vue qui gaffe.	17	8	Le corps règne par l'air jaune.
7	9	La pluie noble file l'oeil.	17	9	Où le linge tend-il la thèse lisse ?
7	10	Le vol doit la scène qui rit.	17	10	Le grade lave le thème qui ment.
7	11	Le parc sûr claque le ciel.	17	11	Comment la robe plonge-t-elle la phase rouge ?
7	12	Le seuil saute le mont qui songe.	17	12	Le bled prend le soir qui tremble.
8	1	Où le vent mène-t-il la pâte brave ?	18	1	Le bar tape dans la dinde grave.
8	2	Le lion tend le train qui dîne.	18	2	La ville digne dresse le frein.
8	3	L'heure pieuse trouve le pied.	18	3	La fois pousse la diète qui court.
8	4	La dame fonce par le vice tiède.	18	4	Quand le verre glisse-t-il le coup noir ?
8	5	La face vient vers le site sec.	18	5	Où le code traite-t-il le coude blond ?
8	6	Quand le prince pique-t-il la mère rose ?	18	6	Où la grange chante-t-elle la buée prête ?
8	7	Le frère tient la crainte qui râle.	18	7	Le train porte le lion qui ment.
8	8	Le livre bave contre la fois prête.	18	8	La soif moche couvre la loi.
8	9	La jupe mûre cache le nuage.	18	9	Le bol fume sans le lien roué.

8	10	La ville ose la date qui songe.
8	11	Quand la suite rend-elle la preuve moche ?
8	12	Le genre fade vise le soir.
9	1	Le nid lutte dans le lien dense.
9	2	Où l'air penche-t-il la science sage ?
9	3	La cage ouvre le toit qui plaît.
9	4	La nuée trompe l'homme qui bave.
9	5	Le bar lourd tourne la sphère.
9	6	Le fer parle vers la salle froide.
9	7	La grange molle tient le film.
9	8	Comment le biais vise-t-il le foin gras ?
9	9	Le fils crie sur le frein vrai.
9	10	Quand la classe tue-t-elle le quai louche ?
9	11	Le grade plat fonde la dinde.
9	12	La phase trouve le nez qui ruse.
10	1	Le coeur reste vers la dune muette.
10	2	La pâte chante le livre qui vient.
10	3	Le thème pousse le lieu qui gaffe.
10	4	La face pâle trace la suite.
10	5	Le fruit raide mange le bled.
10	6	Le pain songe sous le style sale.
10	7	Où la dame montre-t-elle le sac gris ?
10	8	La nuque lisse crée le pied.
10	9	Quand la douane roule-t-elle la robe sourde ?
10	10	La cour pile contre le groupe riche.
10	11	La thèse sert l'heure qui ment.
10	12	Comment la date cite-t-elle la voix brune ?
11	1	La lame part vers le prince née.
11	2	Le nid juge le fond qui dort.
11	3	Le doigt pense pour la viande gaie.
11	4	Où la pomme claque-t-elle la tête rousse ?
11	5	Comment le bain tire-t-il la cour vive ?
11	6	Où le plan pique-t-il la nuque drôle ?
11	7	Le temps nu tient le vice.
11	8	Le monde tend le site qui rit.
11	9	Le blé colle le toit qui tremble.
11	10	La fin ronde craint le fer.
11	11	Le nerf rude traîne le sac.
11	12	La mère pointe par le fils jaune.
12	1	Le mec blanc cache le pot.
12	2	Le lit creux saute le nord.
12	3	Comment le jeu marque-t-il la lune bleue ?
12	4	Le ciel rêve dans la caisse blonde.
12	5	Le bois doit le champ qui naît.
12	6	Comment le ton porte-t-il la vigne brute ?
12	7	Où le rang trouve-t-il la vie riante ?
12	8	Le seuil chante le pré qui chasse.
12	9	Le veau plein lance le mal.
12	10	Le deuil vit contre le goût cuit.
12	11	Le cas vise la pierre qui tape.
12	12	Le mot joue sur la tasse close.
13	1	La cage craque chez le fleuve dur.
13	2	Le jeu beige penche le monde.
13	3	Comment le vin file-t-il la lame sale ?
18	10	Le quai crée le drap qui pile.
18	11	La cuisse tiède lance la rue.
18	12	La classe naît vers la nuée cuite.
19	1	Quand la haine hausse-t-elle la viande louche ?
19	2	Le plomb prend le veau qui rage.
19	3	Comment le ton traîne-t-il la lame grasse ?
19	4	Le pré traite la banque qui règne.
19	5	Le fil sec claque le jeu.
19	6	Où le loup ouvre-t-il la fin mûre ?
19	7	Le pot tue le cas qui dort.
19	8	Le plan fuit pour le monde blême.
19	9	La terre passe dans le choix mince.
19	10	Le prix vrai glisse le nord.
19	11	Le goût tremble par le doigt fade.
19	12	Le ton cher juge le blé.
20	1	Comment la diète nomme-t-elle le corps lourd ?
20	2	Le code pieux hurle le lion.
20	3	La cuisse lutte dans la crainte brave.
20	4	Le nuage trace la fois qui tousse.
20	5	Quand le grain colle-t-il la tête folle ?
20	6	Le coude rose roule le crime.
20	7	Le genre marche vers la pomme froide.
20	8	Comment le nerf traîne-t-il le linge plat ?
20	9	La vache craint la joie qui entre.
20	10	La ville passe sur le train riant.
20	11	La feuille prise sert le temps.
20	12	Le riz coupe le vent qui fuit.
21	1	Le nom jette le vol qui joue.
21	2	Le goût aide le cas qui meurt.
21	3	La fin marche vers le doigt vif.
21	4	La cause rousse rend la dent.
21	5	Le pain luit sans le stade né.
21	6	L'oeuf dicte la viande qui rampe.
21	7	Comment le truc mène-t-il le poste dur ?
21	8	La vue crie dans la nuit creuse.
21	9	Quand le test monte-t-il le blé proche ?
21	10	La pluie beige lave le tort.
21	11	La queue pure souhaite le veau.
21	12	Quand le pot règle-t-il le chêne clair ?
22	1	La place bave par le vol sûr.
22	2	La pente noble roule le thé.
22	3	La gamme mange le vin qui râle.
22	4	Le bout souffre vers le soin dense.
22	5	Le drame trompe la nuit qui chasse.
22	6	Le dôme bleu presse la vue.
22	7	La scène mince veille le mur.
22	8	Le fleuve plaît pour le parc sage.
22	9	Quand le truc penche-t-il la pluie prête ?
22	10	Comment le coin peint-il le chêne tiède ?
22	11	Comment le poste montre-t-il l'onde grave ?
22	12	Le bloc fonde la cause qui pue.
23	1	La soeur drape la nuque qui passe.
23	2	Le plat nu donne le site.
23	3	Le prince baisse l'ange qui fume.

13	4	La bête grise monte le sens.	23	4	Comment le vice quitte-t-il le trou gris ?
13	5	Où l'homme veille-t-il le biais riche ?	23	5	Le soir jette le thème qui règne.
13	6	Le pré rampe sous le coin vert.	23	6	Le frère pue pour la robe brave.
13	7	Où la fée drape-t-elle le plan fluide ?	23	7	Où la mère règle-t-elle le bain mûr ?
13	8	Le foin tousse sur la gamme lente.	23	8	La cour rage sans la preuve née.
13	9	Le fond trompe le parc qui dîne.	23	9	La main raide plonge le jean.
13	10	La part doit le ton qui tombe.	23	10	La thèse sourde pose la plaine.
13	11	Le drame pur saute la pente.	23	11	Comment le bled lave-t-il le gaz vif ?
13	12	Le nord souhaite la scène qui marche.	23	12	La jupe flotte sous le sac pieux.
14	1	La femme blanche hurle le bras.	24	1	Où le corps siffle-t-il le grade louche ?
14	2	La haine râle contre la sphère pure.	24	2	La vigne grasse quitte le riz.
14	3	La salle marque l'or qui vole.	24	3	La phase monte le film qui brille.
14	4	La chance folle siffle la plage.	24	4	L'air hurle la vie qui craque.
14	5	La terre souffre sur le choix vert.	24	5	Le champ vit contre le mot drôle.
14	6	Où le plomb signe-t-il la clef riante ?	24	6	Le fils lourd souhaite le seuil.
14	7	Quand le prix nomme-t-il la boîte prise ?	24	7	Comment le bois donne-t-il la pomme rouge ?
14	8	La science luit sous le loup brut.	24	8	La tête dit le fer qui dîne.
14	9	Le rite loge la gare qui meurt.	24	9	La feuille douce dicte la lune.
14	10	Le nez coupe l'ours qui skie.	24	10	Où le temps drape-t-il le linge rude ?
14	11	Comment la banque sonne-t-elle la brume dure ?	24	11	Le nid flotte sans le deuil rond.
14	12	Le fil net dit le gué.	24	12	Le toit fonce par le nerf mort.

A.3.2 *Phase 2: Radio News Comprehension and Discussion Questions*

Comprehension Question	Corresponding Discussion Question
Pourquoi est-il choquant que le ministre se soit déplacé en hélicoptère pour aller voir un film ?	En général, est-ce que vous choisissez votre mode de déplacement en fonction des effets sur le climat ?
Quel objectif relatif à la réduction des émissions de gaz à effet de serre a-t-il été fixé ?	Dans votre vie quotidienne, quels types d'efforts faites-vous pour protéger l'environnement ?
Que veulent obtenir les agents de la poste à Nivelles ?	Est-ce que les facteurs devraient décider eux-mêmes de leur itinéraire de distribution du courrier ?
Que se passe-t-il au Québec quand il y a une grève de train ?	Êtes-vous en faveur du service minimum pour les services publics ?
Combien de personnes ont participé à l'étude sur les vêtements ?	Portez-vous tous les vêtements que vous achetez ?
Que s'est-il passé le 1er mai en France et à Hambourg ?	Le 1er mai devrait-il être une fête, ou plutôt une journée de manifestation ?
Quelles sont les provinces et régions concernées par la pénurie de professeurs ?	Faut-il attribuer plus de moyens au recrutement de professeurs ou à la construction de nouvelles classes ?
Quelles sont les revendications des syndicats qui ont appelé à cet arrêt de travail ?	Quel autre moyen de protestation les cheminots pourraient-ils utiliser ?
Où ont été amenées les personnes évacuées ?	Que pensez-vous de l'efficacité de ce type de protestation?
Quelles sont les accusations portées contre le fonctionnaire de Charleroi ?	Quel autre moyen le bourgmestre aurait-il pu utiliser pour arriver au même objectif ?
Quelle est la mission des Centres régionaux d'identification?	Quelles techniques de police scientifique connaissez-vous ?
Qui a annoncé une rencontre éventuelle entre la Chine et des représentants du Dalaï Lama ?	Quel est le problème entre la Chine et le Dalaï Lama ?

Que disent les États-Unis sur le rôle d'Israël dans cette affaire?	Trouvez-vous justifié qu'un pays puisse en bombarder un autre pour détruire des armes nucléaires?
Quelle est l'activité principale de cette société qui rencontre des difficultés financières?	Aimez-vous le cristal et l'artisanat en général?
Qui a démisionné en bloc dans la société Cumerio ?	Préféreriez-vous travailler pour une société nationale ou multi-nationale, et pourquoi?
Pourquoi ces deux cinéastes belges sont-ils heureux?	Quel est votre réalisateur préféré et pourquoi?
Quel est l'objectif du groupe de Mme Ebadi en Iran ?	Comment peut-on aider ceux qui luttent pour la défense des droits humains à l'étranger ?
Quels types de véhicules seront visés par les contrôles renforcés sur les autoroutes ?	Comment améliorer la sécurité routière ?
Quelle est la question essentielle selon le parquet dans le procès des dirigeants de la société qui a fait faillite?	Pensez-vous que les gérants de société reconnus coupables de fraudes devraient encourir des peines de prison, ou plutôt d'autres types de peines?
Quelles raisons sont évoquées pour expliquer la diminution du nombre des immigrés italiens en Belgique ?	Quelles sont selon vous les plus grandes communautés d'immigrés en Belgique, outre les Marocains et les Italiens ?
Où se trouvent actuellement les forces des militants islamistes au Liban ?	La communauté internationale devrait-elle intervenir plus activement dans les conflits au Proche-Orient?
Quels ont été les résultats de Justine Henin à Roland Garros les années précédentes?	En général, comment les anciens champions de tennis poursuivent-ils leur carrière professionnelle lorsqu'ils ne font plus de tournois?
Qu'est-ce qui sera exposé au Musée Hergé à Louvain-la-Neuve?	Aimez-vous les bandes dessinées de Tintin, et pourquoi?
Pourquoi est-ce que le magasin Cora a été fermé?	Pensez que les caisses des magasins contiennent encore beaucoup d'argent liquide, ou est-ce que tous les clients payent par carte ?
Qui a cassé le café Golden Gate et quelle était la motivation?	Pensez-vous que les équipes de football sont en partie responsables des violences commises par leurs supporters ?
Quel était le dossier posant un contentieux pour les deux communautés mentionnées dans ce reportage ?	Pensez-vous que les relations entre les communautés en Belgique s'améliorent ou pas ?
Pourquoi critique-t-on le Selor?	Pensez-vous qu'il faille sélectionner tous les fonctionnaires de manière objective, ou y a-t-il des exceptions?
Comment pourriez-vous gagner de l'argent pendant la semaine du Bonjour ?	Que pensez-vous de cette initiative (la semaine du Bonjour) ?
Combien de travailleurs sont concernés par les élections sociales ?	Faut-il limiter le nombre de mandats pour les représentants syndicaux ?
Que dénonce le syndicat des travailleurs de Proximus ?	Quel fournisseur de GSM et d'Internet utilisez-vous, et en êtes-vous satisfait ?
Qu'apportent les ONG aux sans-abris après le cyclone en Birmanie ?	Quelles ONG connaissez-vous et quelles sont leurs activités principales ?
Pourquoi est-ce que l'ONU a arrêté sa distribution de nourriture aux camps de réfugiés ?	Savez-vous combien il y a de réfugiés dans le monde, et où ils se trouvent principalement ?
Pourquoi est-il choquant que le ministre se soit déplacé en hélicoptère pour aller voir un film ?	En général, est-ce que vous choisissez votre mode de déplacement en fonction des effets sur le climat ?
Quel objectif relatif à la réduction des émissions de gaz à effet de serre a-t-il été fixé ?	Dans votre vie quotidienne, quels types d'efforts faites-vous pour protéger l'environnement ?

Que veulent obtenir les agents de la poste à Nivelles ?	Est-ce que les facteurs devraient décider eux-mêmes de leur itinéraire de distribution du courrier ?
Que se passe-t-il au Québec quand il y a une grève de train ?	Êtes-vous en faveur du service minimum pour les services publics ?
Combien de personnes ont participé à l'étude sur les vêtements ?	Portez-vous tous les vêtements que vous achetez ?
Que s'est-il passé le 1er mai en France et à Hambourg ?	Le 1er mai devrait-il être une fête, ou plutôt une journée de manifestation ?
Quelles sont les provinces et régions concernées par la pénurie de professeurs ?	Faut-il attribuer plus de moyens au recrutement de professeurs ou à la construction de nouvelles classes ?
Quelles sont les revendications des syndicats qui ont appelé à cet arrêt de travail ?	Quel autre moyen de protestation les cheminots pourraient-ils utiliser ?
Où ont été amenées les personnes évacuées ?	Que pensez-vous de l'efficacité de ce type de protestation?
Quelles sont les accusations portées contre le fonctionnaire de Charleroi ?	Quel autre moyen le bourgmestre aurait-il pu utiliser pour arriver au même objectif ?
Quelle est la mission des Centres régionaux d'identification?	Quelles techniques de police scientifique connaissez-vous ?
Qui a annoncé une rencontre éventuelle entre la Chine et des représentants du Dalaï Lama ?	Quel est le problème entre la Chine et le Dalaï Lama ?
Que disent les États-Unis sur le rôle d'Israël dans cette affaire?	Trouvez-vous justifié qu'un pays puisse en bombarder un autre pour détruire des armes nucléaires?
Quelle est l'activité principale de cette société qui rencontre des difficultés financières?	Aimez-vous le cristal et l'artisanat en général?
Qui a démisionné en bloc dans la société Cumerio ?	Préféreriez-vous travailler pour une société nationale ou multi-nationale, et pourquoi?
Pourquoi ces deux cinéastes belges sont-ils heureux?	Quel est votre réalisateur préféré et pourquoi?
Quel est l'objectif du groupe de Mme Ebadi en Iran ?	Comment peut-on aider ceux qui luttent pour la défense des droits humains à l'étranger ?
Quels types de véhicules seront visés par les contrôles renforcés sur les autoroutes ?	Comment améliorer la sécurité routière ?
Quelle est la question essentielle selon le parquet dans le procès des dirigeants de la société qui a fait faillite?	Pensez-vous que les gérants de société reconnus coupables de fraudes devraient encourir des peines de prison, ou plutôt d'autres types de peines?
Quelles raisons sont évoquées pour expliquer la diminution du nombre des immigrés italiens en Belgique ?	Quelles sont selon vous les plus grandes communautés d'immigrés en Belgique, outre les Marocains et les Italiens ?
Où se trouvent actuellement les forces des militants islamistes au Liban ?	La communauté internationale devrait-elle intervenir plus activement dans les conflits au Proche-Orient?
Quels ont été les résultats de Justine Henin à Roland Garros les années précédentes?	En général, comment les anciens champions de tennis poursuivent-ils leur carrière professionnelle lorsqu'ils ne font plus de tournois?
Qu'est-ce qui sera exposé au Musée Hergé à Louvain-la-Neuve?	Aimez-vous les bandes dessinées de Tintin, et pourquoi?
Pourquoi est-ce que le magasin Cora a été fermé?	Pensez que les caisses des magasins contiennent encore beaucoup d'argent liquide, ou est-ce que tous les clients payent par carte ?
Qui a cassé le café Golden Gate et quelle était la motivation?	Pensez-vous que les équipes de football sont en partie responsables des violences commises par leurs supporters ?

BIBLIOGRAPHY

Abeillé, A. (Ed.). (2003). *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Kluwer Academic Publishers. (Cited on pp. 103, 109).

Abney, S. (1991). Parsing by chunks. In R. Berwick, S. Abney, & C. Tenny (Eds.), *Principle-Based Parsing*. Kluwer Academic Publishers. (Cited on p. 109).

Adda, G., Adda-Decker, M., Barras, C., Boula de Mareüil, P., Habert, B., & Paroubek, P. (2007). Speech overlap and interplay with disfluencies in political interview. In *Proceedings of the International Workshop on Paralinguistic Speech* (pp. 41–46). (Cited on p. 116).

Adda-Decker, M., Habert, B., Barras, C., Adda, G., Boula de Mareüil, P., & Paroubek, P. (2004). Une étude des disfluences pour la transcription automatique de la parole et l'amélioration des modèles de langage. In *Actes des 25èmes journées d'étude sur la parole*. (Cited on p. 116).

Ahrens, B. (2005). Prosodic phenomena in simultaneous interpreting: A corpus-based analysis. *Interpreting*, *7*(1), 51–76. (Cited on p. 184).

Al Moubayed, S., Beskow, J., & Granström, B. (2009). Auditory visual prominence. *Journal on Multimodal User Interfaces*, *3*(4), 299–309. doi:10.1007/s12193-010-0054-0. (Cited on p. 127)

Allen, R. W. & Jex, H. R. (1968). An experimental investigation of compensatory and pursuit tracking displays with rate and acceleration control dynamics and a disturbance input: NASA CR-1082. (Cited on p. 77).

Anderson, A. H., Bader, M., Gurman Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thomson, H. S., & Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, *34*(4), 351–366. (Cited on p. 74).

Arnold, D. & Wagner, P. (2008). The influence of top-down expectations on the perception of syllable prominence. In *ISCA Workshop on Experimental Linguistics* (pp. 25–28). (Cited on p. 129).

Arnold, D., Wagner, P., & Baayen, R. H. (2013). Using Generalized Additive Models and Random Forests to Model Prosodic Prominence in German. In *Proceedings of Interspeech 2013* (pp. 272–276). (Cited on pp. 129, 130).

Arnold, J. E. (2010). How Speakers Refer: The Role of Accessibility. *Language and Linguistics Compass*, *4*(4), 187–203. doi:10.1111/j.1749-818X.2010.00193.x. (Cited on p. 39)

Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies Signal Theee, Um, New Information. *Journal of Psycholinguistic Research*, *32*(1), 25–36. doi:10.1023/A:1021980931292. (Cited on p. 116)

Arnold, J. E., Losongco, A., Wasow, T., & Ginstrom, R. (2000). Heaviness vs. Newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, *76*, 28–55. (Cited on p. 40).

Association Internationale des Interprètes de Conférence. (1999/2004). Practical guide for professional conference interpreters. Retrieved from http://aiic.net/page/628. (Cited on p. 183)

Astésano, C., Bertrand, R., Espesser, R., & Nguyen, N. (2012). Perception des frontières et des proéminences en français. In *Actes des Journées d'études sur la parole et conférence annuelle du Traitement Automatique des Langues Naturelles* (Vol. 1: JEP, pp. 353–360). (Cited on p. 138).

Atkinson, R. C. & Shiffrin, R. M. (1968). Human Memory: A Proposed System and its Control Processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89–195). New York: Academic Press. (Cited on p. 18).

Auer, P. (2009). On-line syntax: Thoughts on the temporality of spoken language. *Language Sciences*, *31*(1), 1–13. doi:10.1016/j.langsci.2007.10.004. (Cited on p. 57)

Avanzi, M. (2014). A Corpus-Based Approach to French Regional Prosodic Variation. *Nouveaux cahiers de linguistique française*, *31*, 309–323. (Cited on pp. 29, 72, 92, 130).

Avanzi, M., Béguelin M.-J., & Diémoz, F. (2015). De l'archive de parole au corpus de référence: Le corpus oral de français de Suisse romande (OFROM). *Corpus*, *14*, 309–342. (Cited on pp. 92, 113).

Avanzi, M., Christodoulides, G., Lolive, D., & Delais-Roussarie, Elisabeth , Nelly Barbot. (2014). Towards the Adaptation of Prosodic Models for Expressive Text-to-Speech Synthesis. In *Proceedings of Interspeech 2014* (pp. 1796–1800). ISCA. (Cited on pp. 151, 153).

Avanzi, M., Lacheret, A., & Victorri, B. (2008). ANALOR A Tool for Semi-Automatic Annotation of French Prosodic Structure. In *Proceedings of Speech Prosody 2008* (pp. 119–122). (Cited on p. 186).

Avanzi, M., Lacheret, A., & Victorri, B. (2010). A Corpus-based Learning Method for Prominence Detection in Spontaneous Speech. In *Proceedings of the Prosodic Prominence Workshop, Speech Prosody*. (Cited on pp. 94, 98, 129, 130).

Baddeley, A. D. (1976). *The psychology of memory*. Basic topics in cognition series. New York: Basic Books. (Cited on p. 18).

Baddeley, A. D. (1996). Exploring the Central Executive. *The Quarterly Journal of Experimental Psychology*, *49A*(1), 5–28. doi:10.1080/027249896392784. (Cited on p. 20)

Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. (Cited on pp. 19, 20).

Baddeley, A. D., Eysenck, M. W., & Anderson, M. C. (2015). *Memory* (Second edition). London: Psychology Press, Taylor & Francis Group. (Cited on p. 19).

Baddeley, A. D. & Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47–89). New York: Academic Press. (Cited on p. 19).

Baldwin, C. L. (2012). *Auditory cognition and human performance: Research and applications*. Boca Raton, FL: Taylor & Francis. (Cited on pp. 55, 56).

Barbosa, P. A. (1994). *Caractérisation et génération automatique de la structuration rythmique du français.* (PhD Thesis, Université Stendhal, Grenoble, France). (Cited on p. 150).

Bargh, J. A. (1994). The Four Horsemen of Automaticity: Awareness, Intention, Efficiency, and Control in Social Cognition. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (Vol. 1, pp. 1–40). Hillsdale, N.J.: Lawrence Erlbaum Associates. (Cited on pp. 41, 59).

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (1998). Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech. In *Proceedings of the 1st International Conference on Language Resources and Evaluation* (pp. 1373–1376). (Cited on p. 155).

Barreca, G. (2015). *L'acquisition de la liaison chez des apprenants italophones: Des atouts d'un corpus de natifs pour l'étude de la liaison en français langue étrangère (FLE)* (PhD Thesis, Université de Paris Ouest Nanterre). (Cited on p. 43).

Barreca, G. & Christodoulides, G. (2014). Un concordancier multi-niveaux et multimédia pour des corpus oraux. In *Actes de la 21ème Conférence Traitement Automatique du Langage Naturel (TALN)* (pp. 499–504). (Cited on p. 161).

Baumann, S. & Roth, A. (2014). Prominence and coreference – on the perceptual relevance of fo movement, duration and intensity. In N. Campbell, Gibbons, & D. Hirst (Eds.), *Proceedings of Speech Prosody 2014*. (Cited on p. 127).

Bawden, R., Botalla, M.-A., Gerdes, K., & Kahane, S. (2014). Correcting and Validating Syntactic Dependency in the Spoken French Treebank Rhapsodie. In *Proceedings of the 9th International Language Resources and Evaluation Conference*. (Cited on p. 103).

Beatty, J. (1982). Task-Evoked Pupillary Responses, Processing Load and the Structure of Processing Resources. *Psychological Bulletin*, *91*(2), 276–292. (Cited on p. 68).

Beatty, J. & Lucero-Wagoner, B. (2000). The Pupillary System. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of Psychophysiology* (pp. 142–162). Cambridge University Press. (Cited on p. 68).

Benzitoun, C., Fort, K., & Sagot, B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. In *Actes des Journées d'études*

sur la parole et conférence annuelle du Traitement Automatique des Langues Naturelles (pp. 99–112). (Cited on p. 103).

Berthold, A. (1998). *Repräsentation und Verarbeitung sprachlicher. Indikatoren für kognitive Ressourcenbeschränkungen: Representation and processing of linguistic indicators of cognitive resource limitations* (Master's thesis, University of Saarbrücken, Germany). (Cited on p. 78).

Berthold, A. & Jameson, A. (1999). Interpreting Symptoms of Cognitive Load in Speech Input. In J. Kay (Ed.), *Proceedings of the 7th International Conference on User Modeling*. Vienna, New York: Springer. (Cited on p. 78).

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., & Rauzy, S. (2008). Le CID - Corpus of Interactional Data: Annotation et exploitation multimodale de parole conversationnelle. *Traitement Automatique des Langues*, *49*(3), 105–136. (Cited on p. 73).

Bilger, M. & Estelle Campione. (2002). Propositions pour un étiquetage en séquences fonctionnelles. *Recherches sur le Français Parlé*, *17*, 117–136. (Cited on p. 96).

Blanc, O., Constant, M., Dister, A., & Watrin, P. (2008). Corpus oraux et chunking. In *Actes des 27es Journées d'étude sur la parole (JEP)*. (Cited on p. 102).

Blanche-Benveniste, C. (1996). Trois remarques sur l'ordre des mots dans la langue parlée. *Langue Française*, *111*(1), 109–117. (Cited on p. 96).

Blodgett, A. (2004). *The interaction of prosodic phrasing, verb bias, and plausibility during spoken sentence comprehension* (Unpublished doctoral dissertation, The Ohio State University, Columbus). (Cited on p. 61).

Bocci, G. & Avesani, C. (2011). Phrasal prominences do not need pitch movements: postfocal phrasal heads in Italian. In *Proceedings of Interspeech 2011* (pp. 1357–1360). (Cited on p. 127).

Bock, K. J. (1982). Toward a Cognitive Psychology of Syntax: Information Processing Contributions to Sentence Formulation. *Psychological Review*, *89*(1), 1–47. (Cited on p. 34).

Bock, K. J. (1991). A sketchbook of production problems. *Journal of Psycholinguistic Research*, *20*(3), 141–160. doi:10.1007/BF01067212. (Cited on pp. 32, 34)

Bock, K. J. & Levelt, W. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 741–779). New York: Academic Press. (Cited on p. 34).

Boersma, P. & Weenink, D. (2016). Praat: Doing phonetics by computer (Computer Program). Retrieved from http://www.praat.org/. (Cited on pp. 93, 130, 155)

Boland, H. T., Hartsuiker, R. J., Pickering, M. J., & Postma, A. (2005). Repairing inappropriately specified utterances: Revision or restart? *Psychonomic Bulletin & Review*, *12*(3), 472–477. doi:10.3758/BF03193790. (Cited on p. 49)

Bosker, H. R., Pinget, A.-F., Quene, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*(2), 159–175. doi:10.1177/0265532212455394. (Cited on p. 116)

Boula de Mareüil, P., d'Alessandro, C., Raake, A., Bailly, G., Garcia, M.-N., & Morel, M. (2006). A joint intelligibility evaluation of French text-to-speech synthesis systems: the EvaSy SUS/ACR campaign. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (pp. 2034–2037). (Cited on p. 196).

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 DiSS 2005* (pp. 27–32). (Cited on p. 116).

Bouraoui, J.-L. & Vigouroux, N. (2009). Traitement automatique de disfluences dans un corpus linguistiquement contraint. In *Actes de TALN 2009*. (Cited on p. 117).

Branca-Rosoff, S., Fleury, S., Lefeuvre, F., & Pires, M. (2012). Discours sur la ville. Présentation du Corpus de Français Parlé Parisien des années 2000 (CFPP2000). (Cited on p. 98).

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324. (Cited on p. 132)

Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology*, *42*, 361–377. (Cited on p. 66).

Brugman, H. & Russel, A. (2004). Annotating Multimedia/Multi-modal resources with ELAN. In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (pp. 2065–2068). (Cited on p. 155).

Brugos, A. & Shattuck-Hufnagel, S. (2012). A proposal for labelling prosodic disfluencies in ToBI: Poster presentation. In *Advancing Prosodic Transcription for Spoken Language Science and Technology*. (Cited on p. 117).

Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 53–61. (Cited on pp. 63, 64).

Buhmann, J., Caspers, J., van Heuven, V. J., Hoekstra, H., Martens, J.-P., & Swerts, M. (2002). Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus. In *Proceedings of LREC 2002* (pp. 779–785). (Cited on p. 129).

Byrd, D. & Krivokapić, J. (2008). Prosodic Variation: Understanding Scope, Categoricity, and Recursion in Speech Production and Perception. *Journal of the Acoustical Society of America*, *123*(5), 3423. doi:10.1121/1.2934176. (Cited on p. 51)

Campione, E. & Véronis, J. [Jean]. (2002). A Large-Scale Multilingual Study of Silent Pause Duration. In B. Bel & I. Marlien (Eds.), *Proceedings of Speech Prosody 2002* (pp. 199–202). (Cited on pp. 147, 148).

Candea, M. (2000). *Contribution à l'étude des pauses silencieuses et des phénomènes dits d'hésitation en français oral spontané: Etude sur un corpus de récits en classe de français* (PhD Thesis, Université de la Sorbonne nouvelle - Paris III). (Cited on pp. 146, 147).

Cannam, C., Landone, C., & Sandler, M. (2010). Sonic visualiser: An Open Source Application for Viewing, Analysing, and Annotating Music Audio Files. In *Proceedings of the ACM Multimedia 2010 International Conference* (p. 1467). doi:10.1145/1873951.1874248. (Cited on pp. 155, 160)

Caplan, D. & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences*, *22*(1), 77–126. doi:10.1017/S0140525X99001788. (Cited on p. 60)

Carhart, R., Johnson, C., & Goodman, J. (1975). Perceptual masking of spondees by combinations of talkers. *Journal of the Acoustical Society of America*, *58*(S1), S35. doi:10.1121/1.2002082. (Cited on p. 82)

Chambers, F. (1997). What do we mean by fluency? *System*, *25*(4), 535–544. (Cited on p. 184).

Chasaide, A. N., Yanushevskaya, I., Kane, J., & Gobl, C. (2013). The Voice Prominence Hypothesis: The Interplay of Fo and Voice Source Features in Accentuation. In *Proceedings of Interspeech 2013*. (Cited on p. 127).

Chen, S., Epps, J., Ruiz, N., & Fang, C. (2009). Eye activity as a measure of human mental effort in HCI. In *Proceedings of CHI 2009*. (Cited on p. 68).

Christodoulides, G. (2013). *The Prosody of Simultaneous Interpreting: A Corpus-Based Study* (MA Thesis, Université Catholique de Louvain, Louvain-la-Neuve, Belgium). (Cited on pp. 184, 214).

Christodoulides, G. (2014). Praaline: Integrating Tools for Speech Corpus Research. In *Proceedings of the 9th International Language Resources and Evaluation Conference* (pp. 31–34). (Cited on p. 93).

Christodoulides, G. & Avanzi, M. (2014). Phonetic and Prosodic Characteristics of Disfluencies in French Spontaneous Speech: Poster presented at The 13th Conference on Laboratory Phonology (LabPhon 2014), Tokyo, Japan. (Cited on p. 120).

Christodoulides, G. & Avanzi, M. (2015). Automatic Detection and Annotation of Disfluencies in Spoken French Corpora. In *Proceedings of Interspeech 2015* (pp. 1849–1853). (Cited on p. 122).

Christodoulides, G., Avanzi, M., & Goldman, J.-P. (2014). DisMo: A Morpho-syntactic, Disfluency and Multi-Word Unit Annotator: An Evaluation on a Corpus of French Spontaneous and Read Speech. In *Proceedings of the 9th International Language Resources and Evaluation Conference* (pp. 3902–3907). (Cited on pp. 101, 111).

Christodoulides, G. & Lenglet, C. (2014). Prosodic correlates of perceived quality and fluency in simultaneous interpreting. In N. Campbell, Gib-

bons, & D. Hirst (Eds.), *Proceedings of Speech Prosody 2014* (pp. 1002–1006). (Cited on p. 183).

Christodoulides, G. & Simon, A. C. (2015). Exploring Acoustic and Syntactic Cues to Prosodic Boundaries in French: A Multi-Genre Corpus Study. In *Proceedings of ICPhS 2015*. (Cited on pp. 51, 138).

Clark, H. H. & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111. doi:10.1016/S0010-0277(02)00017-3. (Cited on p. 116)

Clark, H. H. & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*(1), 62–81. doi:10.1016/j.jml.2003.08.004. (Cited on p. 39)

Clark, H. H. & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology*, *37*(3), 201–242. doi:10.1006/cogp.1998.0693. (Cited on p. 45)

Clavel, C., Adda, G., Cailliau, F., Garnier-Rizet, M., Cavet, A., Chapuis, G., Courcinous, S., Danesi, C., Daquo, A.-L., Deldossi, M., Guillemin-Lanne, S., Seizou, M., & Suignard, P. (2013). Spontaneous speech and opinion detection: Mining call-centre transcripts. *Language Resources and Evaluation*, *47*(4), 1089–1125. doi:10.1007/s10579-013-9224-5. (Cited on p. 118)

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. doi:10.1177/001316446002000104. (Cited on p. 131)

Cole, J. (2014). Prosody in context: A review. *Language, Cognition and Neuroscience*, *30*(1-2), 1–31. doi:10.1080/23273798.2014.963130. (Cited on p. 29)

Cole, J., Mo, Y., & Hasegawa-Johnson, M. (2010). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, *1*, 425–452. (Cited on pp. 128, 129).

Collados Aís, Á., Macarena Pradas Macías, E., Stévaux, E., & García Becerra, O. (Eds.). (2007). *La Evaluación de la Calidad en Interpretación Simultánea: Parámetros de Incidencia*. Granada, Spain: Comares. (Cited on p. 184).

Constant, M. & Sigogne, A. (2011). MWU-Aware Part-of-Speech Tagging with a CRF Model and Lexical Resources. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World* (pp. 49–56). (Cited on p. 103).

Cooke, M., Garcia Lecumberri, M. L., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *Journal of the Acoustical Society of America*, *123*(1), 414–427. doi:10.1121/1.2804952. (Cited on p. 82)

Cooper, W. E. & Paccia-Cooper, J. (1980). *Syntax and Speech*. Cambridge, MA: Harvard University Press. (Cited on p. 51).

Courtois, B. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, *87*, 11–22. (Cited on p. 111).

Courtois, B., Garrigues, M., Gross, G., Gross, M., Jung, R., Mathieu-Colas, M., Monceaux, A., Poncet-Montange Anne, Silberztein, M., & Vivès, R.

(1997). Dictionnaire électronique DELAC : les noms composés binaires: Rapport Technique du LADL 56. (Cited on p. 111).

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin*, *104*, 163–191. (Cited on pp. 21, 22).

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory* (pp. 62–101). Cambridge: Cambridge University Press. (Cited on p. 21).

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(87-185). (Cited on pp. 18, 22).

Cowan, N. (2011). Working Memory and Attention in Language Use. In J. Guendouzi, F. Loncke, & M. J. Williams (Eds.), *The handbook of psycholinguistic and cognitive processes* (pp. 75–98). New York: Psychology Press. (Cited on pp. 17, 18, 21).

Crawford, M., Brown, G. J., Cooke, M., & Green, P. (1994). Design, collection and analysis of a multi-simultaneous-speaker corpus. *Proceedings of The Institute of Acoustics*, *16*(5), 183–190. (Cited on p. 74).

Cresti, E. & Moneglia, M. (2005). *C-ORAL-ROM: Integrated reference corpora for spoken Romance languages*. Studies in corpus linguistics, 1388-0373. Amsterdam: John Benjamins. (Cited on p. 137).

Crystal, T. H. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *Journal of the Acoustical Society of America*, *88*(1), 101. doi:10.1121/1.399955. (Cited on p. 149)

Cutler, A. (1997). Prosody in the comprehension of spoken language - A literature review. *Language and Speech*, *40*(2), 141–201. (Cited on pp. 29, 50, 61, 137, 145).

Cutler, A. & Pearson, M. (1985). On the analysis of prosodic turn-taking cues. In C. John-Lewis (Ed.), *Intonation in Discourse* (pp. 139–155). London: Croom Helm. (Cited on p. 29).

Cutugno, F., Leone, E., Ludusan, B., & Origlia, A. (2012). Investigating Syllabic Prominence with Conditional Random Fields and Latent-Dynamic Conditional Random Fields. In *Proceedings of Interspeech 2012* (pp. 2402–2405). (Cited on p. 130).

Daneman, M. (1991). Working memory as a predictor of verbal fluency. *Journal of Psycholinguistic Research*, *20*(6), 445–464. doi:10.1007/BF01067637. (Cited on p. 37)

Daneman, M. & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. doi:10.1016/S0022-5371(80)90312-6. (Cited on p. 167)

Daneman, M. & Green, I. (1986). Individual differences in comprehending and producing words in context. *Journal of Memory and Language*, *25*(1), 1–18. doi:10.1016/0749-596X(86)90018-5. (Cited on p. 37)

de Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America*, *97*(1), 491. doi:10.1121/1.412275. (Cited on p. 127)

De Looze, C. (2010). *Analyse et Interprétation de l'Empan Temporel des Variations Prosodiques en Français et en Anglais* (PhD Thesis, Aix-Marseille Université). (Cited on pp. 150, 151).

De Looze, C. & Rauzy, S. (2011). Measuring speakers' similarity in speech by means of prosodic cues: methods and potential. In *Proceedings of Interspeech 2011* (pp. 1393–1396). (Cited on p. 162).

De Looze, C., Scherer, S., Vaughan, B., & Campbell, N. (2014). Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. *Speech Communication*, *58*, 11–34. doi:10.1016/j.specom.2013.10.002. (Cited on p. 72)

Degand, L. & Simon, A. C. (2009). On identifying basic discourse units in speech: theoretical and empirical issues. *Discours*, *4*. Retrieved from http://discours.revues.org/5852. (Cited on pp. 53, 96, 187)

Déjean Le Féal, K. (1990). Some thoughts on the evaluation of simultaneous interpretation. In Bowen D. & M. Bowen (Eds.), *Interpreting: Yesterday, Today, and Tomorrow* (pp. 154–160). Binghamton: State University of New York Press. (Cited on p. 184).

Delais-Roussarie, E. (2000). Vers une nouvelle approche de la structure prosodique. *Langue Française*, *126*(1), 92–112. doi:10.3406/lfr.2000.991. (Cited on p. 29)

Delais-Roussarie, E., Post, B., Avanzi, M., Buthke, C., Di Cristo, A., Feldhausen, I., Jun, S.-A., Martin, P., Meisenburg, T., Rialland, A., Sichel-Bazin, R., & Yoo, H.-Y. (2015). Intonational phonology of French: Developing a ToBI system for French. In S. Frota & P. Prieto (Eds.), *Intonation in Romance* (pp. 63–100). Oxford University Press. doi:10.1093/acprof:oso/9780199685332.003.0003. (Cited on pp. 29, 51)

Delattre, P. (1966). Les dix intonations de base du français. *The French Review*, *40*(1), 1–14. (Cited on p. 29).

Dell, G. S., Chang, F., & Griffin, Z. M. (1999). Connectionist Models of Language Production: Lexical Access and Grammatical Encoding. *Cognitive Science*, *23*(4), 517–542. (Cited on p. 35).

Dellwo, V. (2010). *Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence* (PhD Thesis, Rheinischen Friedrich-Wilhelms-Universität, Bonn, Germany). (Cited on p. 97).

Demol, M., Verhelst, W., & Verhoeve, P. (2007). The Duration of Speech Pauses in a Multilingual Environment. In *Proceedings of Interspeech 2007* (pp. 990–993). (Cited on p. 148).

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., & Zupan, B. (2013). Orange: Data Min-

ing Toolbox in Python. *Journal of Machine Learning Research*, *14*, 2349–2353. Retrieved from http://jmlr.org/papers/v14/demsar13a.html. (Cited on p. 132)

Denis, P. & Sagot, B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*, *46*(4), 721–736. doi:10.1007/s10579-012-9193-0. (Cited on p. 103)

Desmette, D., Hupet, M., Schelstraete, M.-A., & van der Linden, M. (1995). Adaptation en langue française du « Reading Span Test » de Daneman et Carpenter (1980). *L'année psychologique*, *95*(3), 459–482. doi:10.3406/psy.1995.28842. (Cited on p. 167)

Deulofeu, J., Duffort, L., Gerdes, K., Kahane, S., & Pietrandrea, P. (2010). Depends on What the French Say: Spoken Corpus Annotation with and beyond Syntactic Functions. In *Proceedings of the Fourth Linguistic Annotation Workshop* (pp. 274–281). ACL. (Cited on p. 103).

Di Cristo, A. (1999). Vers une modélisation de l'accentuation du français: Première partie. *Journal of French Language Studies*, *9*(2), 143. doi:10.1017/S0959269500004671. (Cited on pp. 30, 137)

Di Cristo, A. (2011). Une approche intégrative des relations de l'accentuation au phrasé prosodique du français. *Journal of French Language Studies*, *21*(01), 73–95. doi:10.1017/S0959269510000505. (Cited on pp. 51, 138)

D'Imperio, M. (1998). Acoustic-perceptual correlates of sentence prominence in Italian questions and statements. *Journal of the Acoustical Society of America*, *104*(3), 1779. doi:10.1121/1.424133. (Cited on p. 127)

Dister, A. (2007). *De la transcription à l'étiquetage morphosyntaxique: Le cas de la banque de données textuelle orale VALIBEL* (Unpublished PhD thesis, Université Catholique de Louvain, Louvain-la-Neuve, Belgium). (Cited on p. 102).

Duez, D. (2001). Signification des hésitations dans la parole spontanée. *Revue PArole*, (17-19), 113–138. (Cited on p. 147).

Durand, J., Laks, B., & Lyche, C. (2009). *Phonologie, variation et accents du français*. IC2 : Traité Cognition et traitement de l'information. Paris: Hermes / Lavoisier. (Cited on pp. 72, 92, 98, 113).

Dutoit, T. (1997). *An introduction to text-to-speech synthesis*. Text, speech, and language technology. Dordrecht: Kluwer Academic Publishers. (Cited on p. 25).

Dutrey, C., Rosset, S., Adda-Decker, M., Clavel, C., & Vasilescu, I. (2014). Disfluences dans la parole spontanée conversationnelle: Détection automatique utilisant des indices lexicaux et acoustiques. In *Actes des 30es Journées d'Étude sur la Parole* (pp. 366–373). (Cited on pp. 117, 124).

Eddelbuettel, D. & François, R. (2011). Rcpp: Seamless R and C++ Integration. *Journal of Statistical Software*, *40*(8), 1–18. (Cited on p. 162).

Engelhardt, P. E., Corley, M., Nigg, J. T., & Ferreira, F. (2010). The role of inhibition in the production of disfluencies. *Memory & Cognition*, *38*(5), 617–628. doi:10.3758/MC.38.5.617. (Cited on p. 69)

Engonopoulos, N., Sayeed, A., & Demberg, V. (2013). Language and cognitive load in a dual task environment. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 2249–2254). (Cited on p. 200).

Ericsson, K. A. & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*, 211–245. (Cited on pp. 21, 22).

Eshkol, I., Tellier, I., Taalab, S., & Billot, S. (2010). Apprendre à étiqueter le corpus oral à l'aide des connaissances linguistiques. In *Actes des 10es Journées internationales d'Analyse statistique des Données Textuelles*. (Cited on p. 102).

Eysenck, M. W. (2006). *Fundamentals of Cognition* (3rd ed.). Hove: Psychology. Retrieved from http://www.loc.gov/catdir/enhancements/fy0654/2005036752-d.html. (Cited on p. 17)

Eysenck, M. W. & Keane, M. T. (2015). *Cognitive Psychology: A student's handbook* (Seventh edition). Psychology Press. (Cited on p. 21).

Fehringer, C. & Fry, C. (2007). Hesitation phenomena in the language production of bilingual speakers: The role of working memory. *Folia Linguistica*, *41*(1-2), 37–72. (Cited on pp. 37, 43).

Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, *30*(2), 210–233. doi:10.1016/0749-596X(91)90004-4. (Cited on p. 51)

Ferreira, F. (1993). Creation of prosody during sentence production. *Psychological Review*, *100*(2), 233–253. doi:10.1037/0033-295X.100.2.233. (Cited on p. 51)

Ferreira, F. (2007). Prosody and performance in language production. *Language and Cognitive Processes*, *22*(8), 1151–1177. doi:10.1080/01690960701461293. (Cited on p. 53)

Ferreira, F. & Swets, B. (2002). How Incremental Is Language Production? Evidence from the Production of Utterances Requiring the Computation of Arithmetic Sums. *Journal of Memory and Language*, *46*(1), 57–84. doi:10.1006/jmla.2001.2797. (Cited on pp. 51, 52)

Ferreira, V. S. (1996). Is It Better to Give Than to Donate? Syntactic Flexibility in Language Production. *Journal of Memory and Language*, *35*(5), 724–755. doi:10.1006/jmla.1996.0038. (Cited on p. 51)

Fitousi, D. & Wenger, M. J. (2011). Processing capacity under perceptual and cognitive load: A closer look at load theory. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 781–798. doi:10.1037/a0020675. (Cited on p. 12)

Fodor, J. D. (2002). Psycholinguistics cannot escape prosody. In B. Bel & I. Marlien (Eds.), *Proceedings of Speech Prosody 2002*. (Cited on p. 50).

Fodor, J. A. (1983). *Modularity of mind: An essay on faculty psychology*. Cambridge: MIT Press. (Cited on p. 57).

Fon, J. (2006). Shape Display: Task Design and Corpus Collection. In *Proceedings of the 3rd Speech Prosody*. (Cited on p. 74).

Fortkamp, M. B. M. (2003). Working Memory Capacity and Fluency, Accuracy, Complexity and Lexical Density in L2 Speech Production. *Fragmentos*, *24*, 69–104. (Cited on p. 37).

Fougeron, C. & Jun, S.-A. (1998). Rate effects on French intonation: Prosodic organization and phonetic realization. *Journal of Phonetics*, *26*(1), 45–69. doi:10.1006/jpho.1997.0062. (Cited on p. 150)

Francis, A. L. & Nusbaum, H. C. (2009). Effects of intelligibility on working memory demand for speech perception. *Attention, Perception & Psychophysics*, *71*(6), 1360–1374. doi:10.3758/APP.71.6.1360. (Cited on p. 59)

Frazier, L., Carlson, K., & Clifton, C. (2006). Prosodic phrasing is central to language comprehension. *Trends in Cognitive Sciences*, *10*(6), 244–249. doi:10.1016/j.tics.2006.04.002. (Cited on pp. 29, 50, 137)

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *Journal of the Acoustical Society of America*, *115*(5), 2246. doi:10.1121/1.1689343. (Cited on p. 82)

Fromkin, V. A. (1971). The Non-Anomalous Nature of Anomalous Utterances. *Language*, *47*(1), 27–52. (Cited on p. 32).

Garrett, M. F. (1975). Syntactic process in sentence production. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 9, pp. 133–177). (Cited on p. 32).

Garrett, M. F. (1980). Levels of processing in sentence production. In B. Butterworth (Ed.), *Language production* (pp. 177–220). London: Academic Press. (Cited on p. 32).

Garrod, S. & Pickering, M. J. (Eds.). (1999). *Language processing*. Hove, East Sussex, UK: Psychology Press. (Cited on pp. 57–59).

Garrod, S. & Pickering, M. J. (2007). Automaticity of language production in monologue and dialogue. In A. Meyer, L. Wheeldon, & A. Krott (Eds.), *Automaticity and control in language processing* (pp. 1–20). Advances in behavioural brain science. Hove: Psychology. (Cited on pp. 36, 41–43).

Gee, J. P. & Grosjean, F. (1983). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, *15*(4), 411–458. doi:10.1016/0010-0285(83)90014-2. (Cited on p. 51)

Gennari, S. P. & Macdonald, M. C. (2009). Linking production and comprehension processes: the case of relative clauses. *Cognition*, *111*(1), 1–23. doi:10.1016/j.cognition.2008.12.006. (Cited on p. 40)

Georgila, K. (2009). Using integer linear programming for detecting speech disfluencies. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 109–112). doi:10.3115/1620853.1620885. (Cited on p. 116)

Georgila, K., Wang, N., & Gratch, J. (2010). Cross-Domain Speech Disfluency Detection. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 237–240). (Cited on p. 116).

Gerdes, K. & Kahane, S. (2009). Speaking in Piles: Paradigmatic Annotation of a French Spoken Corpus. In M. Mahlberg, V. González-Díaz, & C. Smith (Eds.), *Proceedings of the 5th Corpus Linguistics Conference* (Article 309). (Cited on p. 103).

Germesin, S., Becker, T., & Poller, P. (2008). Hybrid Multi-step Disfluency Detection. In A. Popescu-Belis & R. Stiefelhagen (Eds.), *Proceedings of the 5th International Workshop on Machine Learning for Multimodal Interaction* (Vol. 5237, pp. 185–195). doi:10.1007/978-3-540-85853-9{_}17. (Cited on pp. 117, 122)

Gernsbacher, M. A. (1991). Cognitive Processes and Mechanisms in Language Comprehension: The Structure Building Framework. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 27, pp. 217–263). Psychology of learning and motivation. San Diego: Academic Press. (Cited on p. 59).

Gile, D. (2009). *Basic concepts and models for interpreter and translator training* (Revised Edition). Benjamins translation library (0929-7316). Amsterdam: John Benjamins. (Cited on p. 184).

Gimino, A. (2002). Students' investment of mental effort: Paper presented at the annual meeting of the American Educational Research Association. New Orleans, LA. (Cited on p. 64).

Goldman, J.-P. (2011). Easyalign: an automatic phonetic alignment tool under Praat. In *Proceedings of Interspeech 2011* (Vol. 3233-3236). (Cited on p. 93).

Goldman, J.-P., Auchlin, A., Roekhaut, S., Simon, A. C., & Avanzi, M. (2010). Prominence perception and accent detection in French: A corpus-based account. In *Proceedings of Speech Prosody 2010*. (Cited on pp. 134, 148).

Goldman, J.-P., Auchlin, A., & Simon, A. C. (2009). Description prosodique semi-automatique et discrimination des styles de parole. In H.-Y. Yoo & E. Delais-Roussarie (Eds.), *Actes d'IDP 2009* (pp. 207–221). (Cited on p. 187).

Goldman, J.-P., Avanzi, M., Auchlin, A., & Simon, A. C. (2012). A Continuous Prominence Score Based on Acoustic Features. In *Proceedings of Interspeech 2012* (pp. 2454–2457). (Cited on pp. 97, 130, 132, 186).

Goldman, J.-P., Avanzi, M., Simon, A. C., Lacheret, A., & Auchlin, A. (2007). A methodology for the automatic detection of perceived prominent syllables in spoken French. In *Proceedings of Interspeech 2007* (pp. 98–101). (Cited on p. 129).

Goldman, J.-P., Prsir, T., Christodoulides, G., & Auchlin, A. (2014). Speaking style prosodic variation: an 8-hour 9-style corpus study. In N. Campbell, Gibbons, & D. Hirst (Eds.), *Proceedings of Speech Prosody 2014* (pp. 105–109). (Cited on pp. 29, 75, 92, 96).

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in Spontaneous Speech*. London: Academic Press. (Cited on pp. 146, 147).

Goldman-Eisler, F. (1972). Pauses, Clauses, Sentences. *Language and Speech*, *15*(2), 103–113. (Cited on p. 146).

Gopher, D. & Braune, R. (1984). On the Psychophysics of Workload: Why Bother with Subjective Measures? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *26*(5), 519–532. (Cited on p. 65).

Gorovoy, K., Tung, J., & Poupart, P. (2010). Automatic speech feature extraction for cognitive load classification. In *Proceedings of the 2010 Conference of the Canadian Medical and Biological Engineering Society*. (Cited on p. 80).

Gow, D. W. & Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(2), 344–359. doi:10.1037/0096-1523.21.2.344. (Cited on p. 82)

Greene, J. O. & Lindsey, A. E. (1989). Encoding Processes in the Production of Multiple-Goal Messages. *Human Communication Research*, *16*(1), 120–140. (Cited on p. 146).

Griffin, Z. M. & Ferreira, V. S. (2006). Properties of Spoken Language Production. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (pp. 21–59). London: Academic. (Cited on p. 31).

Grimes, D., Tan, D. S., Hudson, S. E., Shenoy, P., & Rao, R. P. (2008). Feasibility and Pragmatics of Classifying Working Memory Load with an Electroencephalograph. In *Proceedings of the 26th Annual CHI Conference on Human Factors in Computing Systems*. (Cited on p. 67).

Grosjean, F. & Collins, M. (1979). Breathing, Pausing and Reading. *Phonetica*, *36*(2). (Cited on p. 146).

Grosjean, F. & Deschamps, A. (1972). Analyse des variables temporelles du français spontané. *Phonetica*, *26*(3). (Cited on p. 149).

Grosjean, F. & Deschamps, A. (1975). Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de parole et variables composantes, phénomènes d'hésitation. *Phonetica*, *31*(144–184). (Cited on pp. 147, 149, 150).

Grosjean, F., Grosjean, L., & Lane, H. (1979). The patterns of silence: Performance structures in sentence production. *Cognitive Psychology*, *11*(1), 58–81. doi:10.1016/0010-0285(79)90004-5. (Cited on pp. 51, 149, 150)

Grosjean, F. & Lane, H. (1976). How the listener integrates the components of speaking rate. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 538–543. (Cited on pp. 146, 149).

Haapalainen, E., Kim, S. J., Forlizzi, J. F., & Dey, A. K. (2010). Psycho-Physiological Measures for Assessing Cognitive Load. In *Proceedings of the 12th ACM international conference on Ubiquitous computing* (pp. 301–310). New York, NY, USA: Association for Computing Machinery (ACM). doi:10.1145/1864349.1864395. (Cited on p. 67)

Hartsuiker, R. J. (2014). Monitoring and Control of the Production System. In M. A. Goldrick, V. S. Ferreira, & M. Miozzo (Eds.), *The Oxford handbook of language production* (pp. 417–436). Oxford: Oxford University Press. (Cited on pp. 43, 45–47, 116).

Hartsuiker, R. J., Catchpole, C. M., de Jong, N. H., & Pickering, M. J. (2008). Concurrent processing of words and their replacements during speech. *Cognition*, *108*(3), 601–607. doi:10.1016/j.cognition.2008.04.005. (Cited on p. 49)

Hartsuiker, R. J. & Kolk, H. H. (2001). Error monitoring in speech production: a computational test of the perceptual loop theory. *Cognitive Psychology*, *42*(2), 113–157. doi:10.1006/cogp.2000.0744. (Cited on p. 45)

Hartsuiker, R. J., Pickering, M. J., & de Jong, N. H. (2005). Semantic and phonological context effects in speech error repair. *Journal of experimental psychology. Learning, memory, and cognition*, *31*(5), 921–932. doi:10.1037/0278-7393.31.5.921. (Cited on pp. 48, 49)

Heeman, P. A., McMillin, A., & Yaruss, J. S. (2006). An Annotation Scheme for Complex Disfluencies. In *Proceedings of the 9th International Conference on Spoken Language Processing* (pp. 1081–1084). (Cited on pp. 117, 120).

Heldner, M. & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*(4), 555–568. doi:10.1016/j.wocn.2010.08.002. (Cited on p. 147)

Henry, S. & Pallaud, B. (2004). Amorces de mots et répétitions dans les énoncés oraux. *Recherches sur le Français Parlé*, *18*, 201–229. (Cited on p. 116).

Holub, E. & Rennert, S. (2011). Fluency and intonation as quality indicators: Paper presented at the Second International Conference on Interpreting Quality, Almuñécar, Spain. (Cited on pp. 184, 185).

Horton, W. S. [W. S.] & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, *96*(2), 127–142. doi:10.1016/j.cognition.2004.07.001. (Cited on p. 39)

Horton, W. S. [William S.] & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*, 91–117. (Cited on p. 39).

Hupin, B. & Simon, A. C. (2007). Analyse phonostylistique du discours radiophonique. *Recherches en communication*, *28*. (Cited on p. 197).

Huttunen, K., Keränen, H. I., Väyrynen, E., Pääkkönen, R. J., & Leino, T. (2011). Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights. *Applied Ergonomics*, *42*(2), 348–357. doi:10.1016/j.apergo.2010.08.005. (Cited on p. 81)

Jameson, A., Kiefer, J., Müller, C., Großmann-Hutter, B., Wittig, F., & Rummer, R. (2009). Assessment of a User's Time Pressure and Cognitive Load on the Basis of Features of Speech. In M. W. Crocker & J. Siekmann (Eds.), *Resource-adaptive cognitive processes*. Berlin: Springer. (Cited on pp. 78, 79).

Jørgensen, F. (2007). The Effects of Disfluency Detection in Parsing Spoken Language. In J. Nivre, H.-J. Kaalep, K. Muischnek, & M. Koit (Eds.), *Proceedings of NODALIDA 2007* (pp. 240–244). (Cited on p. 116).

Just, M. A. & Carpenter, P. A. (1992). A Capacity Theory of Comprehension: Individual Differences in Working Memory. *Psychological Review*, *99*(1), 122–149. (Cited on pp. 22, 60).

Kade, O. (1963). Der Dolmetschvorgang und die Notation. *Fremdsprachen*, *7*(1), 12–20. (Cited on p. 29).

Kassner, M., Patera, W., & Bulling, A. (2014). Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction: Technical Report. Retrieved from http://arxiv.org/abs/1405.0006. (Cited on pp. 169, 177, 201)

Keating, P. & Shattuck-Hufnagel, S. (2002). A Prosodic View of Word Form Encoding for Speech Production. *UCLA Working Papers in Phonetics*, *101*, 112–156. (Cited on pp. 51, 52).

Kemper, S., Herman, R. E., & Lian, C. H. T. (2003). The costs of doing two things at once for young and older adults: Talking while walking, finger tapping, and ignoring speech of noise. *Psychology and Aging*, *18*(2), 181–192. doi:10.1037/0882-7974.18.2.181. (Cited on pp. 38, 39)

Kemper, S., Schmalzried, R., Herman, R. E., Leedahl, S., & Mohankumar, D. (2009). The effects of aging and dual task demands on language production. *Neuropsychology, development, and cognition. Section B, Aging, neuropsychology and cognition*, *16*(3), 241–259. (Cited on p. 39).

Khawaja, A. M. (2010). *Cognitive Load Measurement using Speech and Linguistic Features* (Doctoral dissertation, University of New South Wales, Sydney, Australia). (Cited on p. 64).

Kipp, M. (2001). Anvil - A Generic Annotation Tool for Multimodal Dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)* (pp. 1367–1370). (Cited on p. 157).

Kirsner, K., Dunn, J., & Hird, K. (2005). Language Production: a complex dynamic system with a chronometric footprint. In *Proceedings of the 7th International Conference on Cognitive Systems*. (Cited on p. 147).

Kjelgaard, M. M. & Speer, S. R. (1999). Prosodic Facilitation and Interference in the Resolution of Temporary Syntactic Closure Ambiguity. *Journal of Memory and Language*, *40*(2), 153–194. doi:10.1006/jmla.1998.2620. (Cited on p. 61)

Klingner, J. (2010). *Measuring Cognitive Load During Visual Tasks by Combining Pupillometry and Eye Tracking* (PhD Thesis, Stanford University). (Cited on p. 68).

Koch, P. & Oesterreicher, W. (2001). Langage oral et langage écrit. In G. Holtus, M. Metzeltin, & C. Schmitt (Eds.), *Lexicon der romanistischen Linguistik* (Vol. I/2, pp. 584–627). Berlin, Boston: De Gruyter. (Cited on p. 29).

Kramer, A. F. (1991). Physiological metrics of mental workload: A review of recent progress. In D. Damos (Ed.), *Multiple task performance* (pp. 279–328). Taylor & Francis. (Cited on p. 66).

Krivokapić, J. (2007). Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics*, *35*(2), 162–179. (Cited on p. 146).

Krivokapić, J. (2012). Prosodic planning in speech production. In S. Fuchs, M. Weihrich, D. Pape, & P. Perrier (Eds.), *Speech Planning and Dynamics* (pp. 157–190). Peter Lang. (Cited on pp. 51–53).

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press. (Cited on p. 29).

Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J.-P., Obin, N., Pietrandrea, P., & Tchobanov, A. (2014). Rhapsodie: a Prosodic-Syntactic Treebank for Spoken French. In *Proceedings of the 9th International Language Resources and Evaluation Conference*. (Cited on pp. 92, 99).

Lacheret, A., Simon, A. C., Goldman, J.-P., & Avanzi, M. (2013). Prominence perception and accent detection in French: From phonetic processing to grammatical analysis. *Language Sciences*, *39*, 95–106. (Cited on pp. 128, 134).

Lackner, J. R. & Tuller, B. H. (1979). Role of efference monitoring in the detection of self-produced speech error. In M. Garrett, W. E. Cooper, & E. C. T. Walker (Eds.), *Sentence processing* (pp. 281–294). Erlbaum: Hillsdale. (Cited on p. 45).

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning* (pp. 282–289). (Cited on pp. 102, 132).

Landis, J. R. & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. (Cited on p. 131).

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(3), 451–468. doi:10.1037/0096-1523.21.3.451. (Cited on p. 11)

Lavie, N. (2000). Selective attention and cognitive control: Dissociating attentional functions through different types of load. In S. Monsell & J. Driver (Eds.), *Control of cognitive processes* (pp. 175–194). Cambridge, Mass.: MIT Press. (Cited on p. 11).

Lavie, N., Hirst, A., de Fockert, J. W., & Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of experimental psychology. General*, *133*(3), 339–354. doi:10.1037/0096-3445.133.3.339. (Cited on p. 11)

Lenglet, C. (2015). *Prosodie et qualité en interprétation simultanée : analyse et perseption* (PhD Thesis, Université de Mons). (Cited on p. 183).

Levelt, W. (1983). Monitoring and self-repair in speech. *Cognition*, *14*, 41–104. (Cited on pp. 44–48).

Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press. (Cited on pp. 34, 42–45, 47, 49, 51, 52, 116).

Levelt, W. (1999). Producing spoken language: a blueprint of the speaker. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 83–122). Oxford: Oxford University Press. (Cited on p. 52).

Lindblom, B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 403–439). Dordrecht, Netherlands: Springer. doi:10.1007/978-94-009-2037-8{_}16. (Cited on p. 78)

Little, D. R., Oehmen, R., Dunn, J., Hird, K., & Kirsner, K. (2013). Fluency Profiling System: an automated system for analyzing the temporal properties of speech. *Behavior research methods*, *45*(1), 191–202. doi:10.3758/s13428-012-0222-0. (Cited on p. 148)

Liu, Y., Shriberg, E., & Stolcke, A. (2003). Automatic Disfluency Identification in Conversational Speech Using MultipleKnowledge Sources. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*. (Cited on p. 116).

Liu, Y., Shriberg, E., Stolcke, A., & Harper, M. (2005). Comparing HMM, maximum-entropy and conditional random fields for disfluency detection. In *Proceedings of Interspeech 2005* (pp. 3313–3316). (Cited on p. 116).

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., & Harper, M. (2006). Enriching speech recognition with automatic detection ofsentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech and Language Processing*, *14*(5), 1526–1540. (Cited on p. 116).

Lively, S. E., Pisoni, D. B., van Summers, W., & Bernacki, R. H. (1993). Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences. *Journal of the Acoustical Society of America*, *93*(5), 2962–2973. (Cited on pp. 77, 78).

Logan, G. D. & Cowan, W. B. (1984). On the ability to inhibit thought and action: A theory of an act of control. *Psychological Review*, *91*(3), 295–327. doi:10.1037/0033-295X.91.3.295. (Cited on pp. 45, 47)

Ludusan, B., Origlia, A., & Cutugno, F. (2011). On the use of the rhythmogram for automatic syllabic prominence detection. In *Proceedings of Interspeech 2011* (pp. 2413–2416). (Cited on p. 128).

Macken, W. J., Tremblay, S., Houghton, R. J., Nicholls, A. P., & Jones, D. M. (2003). Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(1), 43–51. doi:10.1037/0096-1523.29.1.43. (Cited on p. 55)

Maffia, M., Pellegrino, E., & Pettorino, M. (2014). Labeling expressive speech in L2 Italian: the role of prosody in auto-and external annotation. In N. Campbell, Gibbons, & D. Hirst (Eds.), *Proceedings of Speech Prosody 2014* (pp. 81–85). (Cited on p. 74).

Mahr, A., Feld, M., Mehdi Moniri, M., & Math, R. (2012). The ConTRe (Continuous Tracking and Reaction) Task: A Flexible Approach for Assessing Driver Cognitive Workload with High Sensitivity. In A. L. Kun, L. N. Boyle, B. Reimer, & A. Riener (Eds.), *Adjunct Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 88–91). (Cited on p. 198).

Manning, C. D. (2011). Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In *Proceedings of the 12th International Conference CICLing 2011* (Vol. 6608, pp. 171–189). Computational Linguistics and Intelligent Text Processing. doi:10.1007/978-3-642-19400-9{_}14. (Cited on p. 103)

Marslen-Wilson, W. & Tyler, L. K. (1987). Against modularity. In J. L. Garfield (Ed.), *Modularity in Knowledge Representation and Natural-Language Understanding* (pp. 57–104). Cambridge, MA: MIT Press. (Cited on p. 57).

Marslen-Wilson, W. & Welsh, A. (1978). Processing Interactions and Lexical Access during Word Recognition in Continuous Speech. *Cognitive Psychology*, *10*, 29–63. (Cited on p. 57).

Martin, L. J., Degand, L., & Simon, A. C. (2014). Forme et fonction de la périphérie gauche dans un corpus oral multigenres annoté. *Corpus*, *13*(13), 243–265. (Cited on pp. 92, 93).

Martin, P. (2008). *Phonétique acoustique: Introduction à l'analyse acoustique de la parole*. Cursus Linguistique. Paris: Armand Colin. (Cited on p. 27).

Martin, R. C. & Slevc, L. R. [L. Robert]. (2014). Language production and working memory. In M. A. Goldrick, V. S. Ferreira, & M. Miozzo (Eds.), *The Oxford handbook of language production* (pp. 437–450). Oxford: Oxford University Press. (Cited on pp. 37–41).

Mateer, M. & Taylor, A. (1995). Disfluency annotation stylebook for the Switchboard corpus: Manuscript. (Cited on p. 118).

Mathot, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: an open-source, graphical experiment builder for the social sciences. *Behavior research methods*, *44*(2), 314–324. doi:10.3758/s13428-011-0168-7. (Cited on pp. 141, 165)

Mattys, S. L., Brooks, J., & Cooke, M. (2009). Recognizing speech under a processing load: dissociating energetic from informational factors. *Cognitive Psychology*, *59*(3), 203–243. doi:10.1016/j.cogpsych.2009.04.001. (Cited on pp. 81, 82)

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7-8), 953–978. doi:10.1080/01690965.2012.705006. (Cited on pp. 82, 83)

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: a hierarchical framework. *Journal of experimental psychology. General*, *134*(4), 477–500. doi:10.1037/0096-3445.134.4.477. (Cited on p. 82)

McAngus Todd, N. P. & Brown, G. J. (1996). Visualization of Rhythm, Time and Metre. *Artificial Intelligence Review*, *10*, 253–273. (Cited on p. 128).

Mertens, P. (1991). Local Prominence of Acoustic and Psychoacoustic Functions and Perceived Stress in French. In *Proceedins of the 12th International Congress of Phonetic Sciences* (Vol. 3, pp. 218–221). (Cited on p. 128).

Mertens, P. (1993). Intonational grouping, boundaries, and syntactic structure in French. In D. House & P. Touati (Eds.), *Proceedings of the ESCA Workshop on Prosody* (Vol. 41, pp. 156–159). Lund (S) Working Papers. Lund University. (Cited on pp. 30, 137).

Mertens, P. (2004). The Prosogram: Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model. In B. Bel & I. Marlien (Eds.), *Proceedings of Speech Prosody 2004* (pp. 549–552). (Cited on pp. 97, 130, 131, 138, 160).

Mertens, P. (2008). Syntaxe, prosodie et structure informationnelle: Une approche prédictive pour l'analyse de l'intonation dans le discours. *Travaux de Linguistique*, *56*(1), 87–124. (Cited on p. 51).

Mertens, P. (2014). Polytonia: a system for the automatic transcription of tonal aspects in speech corpora. *Journal of Speech Sciences*, *4*(2), 17–57. (Cited on pp. 28, 29, 137).

Mertens, P. & Simon, A. C. (2013). Towards automatic detection of prosodic boundaries in spoken French. In P. Mertens & A. C. Simon (Eds.), *Proceedings of the Prosody-Discourse Interface Conference 2013* (pp. 81–87). (Cited on pp. 51, 95, 186).

Meyer, A., Wheeldon, L., & Krott, A. (Eds.). (2007). *Automaticity and control in language processing*. Advances in behavioural brain science. Hove: Psychology. (Cited on p. 36).

Mieskes, M. & Strube, M. (2008). A Three-stage Disfluency Classifier for Multi Party Dialogues. In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (pp. 2681–2686). (Cited on p. 117).

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. *Psychological Review*, *63*, 81–97. (Cited on p. 17).

Mixdorff, H. (2004). Qualitative analysis of prosody in task-oriented dialogs. In B. Bel & I. Marlien (Eds.), *Proceedings of Speech Prosody 2004* (pp. 283–286). (Cited on p. 74).

Miyake, A. & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge: Cambridge University Press. (Cited on pp. 22, 23).

Mo, Y., Cole, J., & Lee, E.-K. (2008). Naïve listeners' prominence and boundary perception. In *Proceedings of Speech Prosody 2008* (pp. 735–738). (Cited on p. 129).

Moniz, H., Batista, F., Trancoso, I., & Mata, A. I. (2012). Prosodic Context-based Analysis of Disfluencies. In *Proceedings of Interspeech 2012*. (Cited on p. 121).

Moniz, H., Ferreira, J., Batista, F., & Trancoso, I. (2015). Disfluency detection across domains. In *Proceedings of DiSS 2015*. (Cited on p. 116).

Moniz, H., Trancoso, I., & Mata, A. I. (2009). Classification of Disfluent Phenomena as Fluent Communicative Devices in Specific Prosodic Contexts. In *Proceedings of Interspeech 2009* (pp. 1719–1722). (Cited on p. 116).

Moray, N. P. (1969). *Listening and attention*. Penguin science of behaviour. Harmondsworth: Penguin. (Cited on p. 56).

Moreno, I. & Pineda, L. (2006). Speech Repairs in the DIME corpus. *Research in Computing Science*, *20*, 63–74. (Cited on p. 116).

Moss, H. E. & Gaskell, M. G. (1999). Lexical semantic processing during speech comprehension. In S. Garrod & M. J. Pickering (Eds.), *Language processing* (pp. 59–99). Hove, East Sussex, UK: Psychology Press. (Cited on p. 58).

Nenkova, A., Gravano, A., & Hirschberg, J. (2008). High frequency word entrainment in spoken dialogue. In *Proceedings of ACL-08: HLT* (pp. 169–172). doi:10.3115/1557690.1557737. (Cited on p. 72)

Nespor, M. & Vogel, I. (1986). *Prosodic Phonology*. Dordrecht: Foris. (Cited on pp. 50, 51).

Niebuhr, O. & Michaud, A. (2015). Speech data acquisition: the underestimated challenge. *Kalipho*, *3*, 1–42. (Cited on pp. 71–75).

Nooteboom, S. G. (1980). Speaking and unspeaking : detection and correction of phonological and lexical errors in spontaneous speech. In V. A. Fromkin (Ed.), *Errors in linguistic performance : slips of the tongue, ear, pen and hand* (pp. 87–95). London: Academic Press. (Cited on pp. 46, 47).

Nusbaum, H. C. & Schwab, E. C. (1986). The role of attention and active processing in speech perception. In E. C. Schwab & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines* (Vol. 1, pp. 113–157). Academic Press series in cognition and perception. Orlando: Academic. (Cited on p. 83).

Nygren, T. E. (1991). Psychometric Properties of Subjective Workload Measurement Techniques: Implications for Their Use in the Assessment of Perceived Mental Workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *33*(1), 17–33. doi:10.1177/001872089103300102. (Cited on p. 64)

Obin, N., Rodet, X., & Lacheret, A. (2008). Un modèle de durée des syllabes fondé sur les propriétés syllabiques intrinsèques et les variations locales de débit. In *Actes des 27èmes Journées d'étude sur la parole*. Association Francophone de la Communication Parlée. (Cited on pp. 150, 151).

Obin, N., Rodet, X., & Lacheret, A. (2009). A Syllable-Based Prominence Detection Model Based on Discriminant Analysis and Context-Dependency. In *Proceedings of the Speech and Computer Conferecne 2009*. (Cited on p. 130).

O'Connell, D. C. & Kowal, S. (2008). *Communicating with one another: Toward a psychology of spontaneous spoken discourse*. New York: Springer. (Cited on p. 31).

Oehmen, R., Kirsner, K., & Fay, N. (2010). Reliability of the manual segmentation of pauses in naturalspeech. In H. Loftsson, E. Rögnvaldsson, & S. Helgadóttir (Eds.), *Advances in Natural Language Processing* (Vol. 6233). Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. (Cited on pp. 146, 147).

Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America*, *99*(3), 1718–1725. (Cited on p. 29).

Oomen, C. C. E. & Postma, A. (2001). Effects of Divided Attention on the Production of Filled Pauses and Repetitions. *Journal of Speech Language and Hearing Research*, *44*(5), 997. doi:10.1044/1092-4388(2001/078). (Cited on pp. 45, 47)

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive Load Theory: Instructional Implications of the Interaction between Information Structures and Cognitive Architecture. *Instructional Science*, *32*, 1–8. (Cited on p. 12).

Paas, F., Tuovinen, Juhani, E., Tabbers, H., & van Gerven, P. W. M. (2003). Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist*, *38*(1), 63–71. (Cited on pp. 64, 65, 67).

Paas, F., van Merrienboer, J. J., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and motor skills*, *79*(1 Pt 2), 419–430. doi:10.2466/pms.1994.79.1.419. (Cited on pp. 64, 67)

Pallaud, B., Rauzy, S., & Blache, P. (2013). Auto-interruptions et disfluences en français parlé dans quatre corpus du CID. *Travaux interdisciplinaires sur la parole et le langage*, (29). Retrieved from http://tipa.revues.org/995. (Cited on pp. 116, 118)

Parbery-Clark, A., Skoe, E., Lam, C., & Kraus, N. (2009). Musician enhancement for speech-in-noise. *Ear and hearing*, *30*(6), 653–661. doi:10.1097/AUD.0b013e3181b412e9. (Cited on p. 72)

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, *119*(4), 2382. doi:10.1121/1.2178720. (Cited on p. 72)

Patten, C. J. (2007). *Cognitive workload and the driver: Understanding the effects of cognitive workload on driving from a human information processing perspective* (Doctoral dissertation, Stockholm University, Stockholm). (Cited on pp. 13, 14).

Peshkov, K., Prévot, L., Rauzy, S., & Pallaud, B. (2013). Categorizing syntactic chunks for marking disfluent speech in French language. In *Proceedings of DiSS 2013* (pp. 59–62). (Cited on pp. 116, 117).

Petrov, S., Das, D., & McDonald, R. (2012). A Universal Part-of-Speech Tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (pp. 2089–2096). (Cited on p. 109).

Pfitzinger, H. R. (1996). Two approaches to speech rate estimation. In ASSTA (Ed.), *Proceedings of SST 1996* (Vol. 96, pp. 421–426). Australasian Speech Science and Technology Association. (Cited on p. 149).

Pfitzinger, H. R. (1998). Local speech rate as a combination of syllable and phone rate. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)* (pp. 1087–1090). Sydney. (Cited on p. 150).

Pomplun, M. & Sunkara, S. (2003). Pupil Dilation as an Indicator of Cognitive Workload in Human-Computer Interaction. In *Proceedings of Inter-*

national Conference on Human-Computer Interaction (Vol. 3, pp. 542–546). Mahwah: Lawrence Erlbaum Associates. (Cited on p. 68).

Postma, A. (2000). Detection of errors during speech production: a review of speech monitoring models. *Cognition*, *77*, 97–131. (Cited on pp. 43, 46, 48).

Postma, A. & Kolk, H. (1993). The Covert Repair Hypothesis. *Journal of Speech Language and Hearing Research*, *36*(3), 472. doi:10.1044/jshr.3603.472. (Cited on pp. 44, 45)

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, *1*(1), 81–106. doi:10.1023/A:1022643204877. (Cited on p. 132)

R Core Team. (2016). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/. (Cited on p. 155)

Raake, A. & Katz, B. F. (2006). SUS-based Method for Speech Reception Threshold Measurement in French. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (pp. 2028–2033). (Cited on p. 196).

Rasmussen, J. (1980). What can be Learned from Human Error Reports? In K. D. Duncan, M. M. Gruneberg, & D. Wallis (Eds.), *Proceedings of an International Conference on Changes in the Nature and Quality of Working Life* (pp. 97–113). John Wiley and Sons. (Cited on p. 12).

Rasmussen, J. (1986). *Information processing and human-machine interaction: An approach to cognitive engineering*. North-Holland series in system science and engineering. New York: North-Holland. (Cited on p. 12).

Rasmussen, J. (1987). Cognitive Control and Human Error Mechanisms. In J. Rasmussen, K. D. Duncan, & J. Leplat (Eds.), *New Technology and Human Error* (pp. 53–61). Wiley. (Cited on pp. 12, 13).

Reason, J. T. (1990). *Human error*. Cambridge: Cambridge University Press. (Cited on p. 12).

Rietveld, A. & Gussenhoven, C. (1985). On the relation between pitch excursion size and prommence. *Journal of Phonetics*, *13*, 299–308. (Cited on p. 127).

Roelofs, A. (2003). Goal-referenced selection of verbal action: Modeling attentional control in the Stroop task. *Psychological Review*, *110*(1), 88–125. doi:10.1037/0033-295X.110.1.88. (Cited on p. 42)

Ronnberg, J., Rudner, M., Foo, C., & Lunner, T. (2008). Cognition counts: a working memory system for ease of language understanding (ELU). *International journal of audiology*, *47*(Supplement 2), 99–105. doi:10.1080/14992020802301167. (Cited on p. 83)

Rossi, M. (1999). *L'intonation: Le système du français : description et modélisation*. Collection L'essentiel français. Gap: Ophrys. (Cited on pp. 30, 137).

Rossnagel, C. S. (2004). Lost in thought: cognitive load and the processing of addressees' feedback in verbal communication. *Experimental psychology*, *51*(3), 191–200. doi:10.1027/1618-3169.51.3.191. (Cited on p. 39)

Rossnagel, C. (2000). Cognitive load and perspective-taking: Applying the automatic-controlled distinction to verbal communication. *European Journal of Social Psychology*, *30*(3), 429–445. doi:10.1002/(SICI)1099-0992(200005/06)30:3{<}429::AID-EJSP3{>}3.0.CO;2-V. (Cited on p. 39)

Roze, C. (2009). *LEXCONN : Base lexicale des connecteurs discursifs du français* (Mémoire de Master, Université Paris Diderot). (Cited on p. 111).

Sajous, F., Hathout, N., & Calderone, B. (2013). GLÀFF, un Gros Lexique À tout Faire du Français. In *Actes de la conférence Traitement Automatique des Langues Naturelles* (pp. 285–298). (Cited on p. 111).

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*. (Cited on p. 104).

Schmidt, T. & Wörner, K. (2009). EXMARaLDA – Creating, analysing and sharing spoken languagecorpora for pragmatic research. *Pragmatics*, *19*(4), 565–582. (Cited on p. 155).

Schneider, W. & Shriffin, R. (1977). Controlled and automatic human information processing I: Detection, Search and Attention. *Psychological Review*, *84*(1), 1–66. (Cited on p. 15).

Schon, D., Magne, C., & Besson, M. (2004). The music of speech: music training facilitates pitch processing in both music and language. *Psychophysiology*, *41*(3), 341–349. doi:10.1111/1469-8986.00172.x. (Cited on p. 72)

Scontras, G., Badecker, W., Shank, L., Lim, E., & Fedorenko, E. (2015). Syntactic complexity effects in sentence production. *Cognitive Science*, *39*(3), 559–583. doi:10.1111/cogs.12168. (Cited on p. 40)

Seeber, K. G. (2011). Cognitive load in simultaneous interpreting: Existing theories - New models. *Interpreting*, *13*(2), 176–204. (Cited on p. 64).

Seeber, K. G. & Kerzel, D. (2012). Cognitive load in simultaneousinterpreting: Model meets data. *International Journal of Bilingualism*, *16*(2), 228–242. doi:10.1177/1367006911402982. (Cited on p. 184)

Selkirk, E. O. (1984). *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: MIT Press. (Cited on pp. 50, 51).

Seyfeddinipur, M., Kita, S., & Indefrey, P. (2008). How speakers interrupt themselves in managing problems in speaking: evidence from self-repairs. *Cognition*, *108*(3), 837–842. doi:10.1016/j.cognition.2008.05.004. (Cited on p. 49)

Shattuck-Hufnagel, S. & Turk, A. E. (1996). A Prosody Tutorial for Investigators of Auditory Sentence Processing. *Journal of Psycholinguistic Research*, *25*(2), 193–247. (Cited on p. 50).

Shi, Y., Ruiz, N., Taib, R., Choi, E., & Chen, F. (2007). Galvanic Skin Response (GSR) As an Index of Cognitive Load. In *Extended Abstracts on Human Factors in Computing Systems* (pp. 2651–2656). New York, NY, USA: Association for Computing Machinery (ACM). doi:10.1145/1240866.1241057. (Cited on p. 67)

Shlesinger, M. (1994). Intonation in the production and perception of simultaneous interpretation. In S. Lambert & B. Moser-Mercer (Eds.), *Bridging the Gap: Empirical Research in Simultaneous Interpretation* (pp. 225–236). Amsterdam: John Benjamins. (Cited on pp. 184, 185).

Shriberg, E. (1994). *Preliminaries to a Theory of Speech Disfluencies* (PhD Thesis, University of California, Berkeley). (Cited on pp. 43, 115, 117, 120).

Shriberg, E. (1999). Phonetic Consequences of Speech Disfluency. In *Proceedings of the 14th International Conference on Phonetic Sciences (ICPhS)* (pp. 619–622). (Cited on p. 121).

Shriberg, E. (2001). To 'errrr' is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, *31*(01). doi:10.1017/S0025100301001128. (Cited on pp. 115, 120, 121)

Shriberg, E., Bates, R., & Stolcke, A. (1997). A prosody-only decision-tree model for disfluency detection. In G. Kokkinakis, N. Fakotakis, & E. Dermatas (Eds.), *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech)* (pp. 2383–2386). (Cited on p. 116).

Shriffin, R. & Schneider, W. (1977). Controlled and automatic human information processing II: Perceptual Learning, Automatic Attending and a General Theory. *Psychological Review*, *84*(2), 127–190. (Cited on p. 15).

Simon, A. C. (2012). *La variation prosodique régionale en français*. Champs linguistiques. Recueils. Bruxelles: De Boeck / Duculot. (Cited on pp. 29, 72).

Simon, A. C., Auchlin, A., Avanzi, M., & Goldman, J.-P. (2010). Les phonostyles: une description prosodique des styles de parole en français. In M. Abecassis & G. Ledegen (Eds.), *Les voix des Français* (Vol. 93, 94, pp. 71–88). Oxford: Peter Lang. (Cited on p. 29).

Simon, A. C. & Christodoulides, G. (2016). Perception of Prosodic Boundaries by Naïve Listeners in French. In *Proceedings of Speech Prosody 2016*. (Cited on p. 141).

Simon, A. C. & Christodoulides, G. (2016, accepted). Frontières prosodiques perçues : corrélats acoustiques et indices syntaxiques. *Langue Française*. (Cited on pp. 141, 144).

Simon, A. C., Francard, M., & Hambye, P. (2014). The Valibel Speech Database. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *The Oxford handbook of corpus phonology*. Oxford handbooks in linguistics. doi:10.1093/oxfordhb/9780199571932.013.017. (Cited on pp. 92, 102, 113)

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and Linguistics Compass*, *3*(2), 621–640. doi:10.1111/j.1749-818X.2009.00125.x. (Cited on p. 72)

Slevc, L. R. [L. R.]. (2011). Saying what's on your mind: working memory effects on sentence production. *Journal of experimental psychology. Learning, memory, and cognition*, *37*(6), 1503–1514. doi:10.1037/a0024350. (Cited on p. 40)

Sluijter, A. M. C. & van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, *100*(4), 2471. doi:10.1121/1.417955. (Cited on p. 131)

Smith, C. L. (2011). Perception of Prominence and Boundaries by Naïve French Listeners. In *Proceedings of ICPhS 2011* (pp. 1874–1877). (Cited on p. 129).

Speer, S. & Blodgett, A. (2006). Prosody. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (pp. 505–537). London: Academic. (Cited on pp. 60, 61).

Spilková, H., Brenner, D., Öttl, A., Vondricka, P., & van Dommelen, W. (2010). The Kachna L1/L2 Picture Replication Corpus. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*. (Cited on p. 74).

Strangert, E. (1997). Relating prosody to syntax: Boundary signalling in Swedish. In G. Kokkinakis, N. Fakotakis, & E. Dermatas (Eds.), *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech)* (pp. 239–242). (Cited on p. 51).

Stroop, J. R. (1935). Studies of Interference in Serial Verbal Reactions. *Journal of Experimental Psychology*, *18*, 643–662. (Cited on p. 166).

Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, *12*, 257–285. (Cited on p. 12).

Tamburini, F. (2003). Automatic Prosodic Prominence Detection in Speech Using Acoustic Features: An Unsupervised System. In *Proceedings of Interspeech 2003* (pp. 129–132). (Cited on p. 129).

Tanguy, N., van Damme, T., Degand, L., & Simon, A. C. (2012). Projet FRFC «Périphérie gauche des unités de discours» - Protocole de codage syntaxique. Louvain-la-Neuve, Belgium. (Cited on p. 96).

Tehrani, H. (2009). EggWorks: A program for automated analysis of EGG signals. (Cited on p. 171).

Tellier, I., Duchier, D., Eshkol, I., Courmet, A., & Martinet, M. (2012). Apprentissage automatique d'un chunker pour le français. In *Actes des Journées d'études sur la parole et conférence annuelle du Traitement Automatique des Langues Naturelles* (Vol. 2 - TALN, pp. 431–438). (Cited on pp. 102, 109).

Tellier, I., Eshkol, I., Dupont, Y., & Wang, I. (2014). Peut-on bien chunker avec de mauvaises ´etiquettes POS? In B. Bigi (Ed.), *Actes de la conférence Traitement Automatique des Langues Naturelles*. (Cited on p. 103).

Terken, J. (1991). Fundamental Frequency and Perceived Prominence of Accented Syllables. *Journal of the Acoustical Society of America*, *89*, 1768–1776. (Cited on pp. 29, 127, 129).

Terken, J. & Hermes, D. (2000). The Perception of Prosodic Prominence. In M. Horne (Ed.), *Prosody: Theory and Experiment* (pp. 89–127). Dordrecht: Kluwer Academic Publishers. (Cited on p. 129).

Turco, G., Gubian, M., & Schertz, J. (2011). A quantitative investigation of the prosody of Verum Focus in Italian. In *Proceedings of Interspeech 2011* (pp. 961–964). (Cited on p. 74).

Turk, A. E. & Sawusch, J. R. (1996). The processing of duration and intensity cues to prominence. *Journal of the Acoustical Society of America*, *99*(6), 3782–3790. doi:10.1121/1.414995. (Cited on p. 127)

Tydgat, I., Stevens, M., Hartsuiker, R. J., & Pickering, M. J. (2011). Deciding where to stop speaking. *Journal of Memory and Language*, *64*(4), 359–380. doi:10.1016/j.jml.2011.02.002. (Cited on p. 48)

Vainio, M. & Järvikivi, J. (2006). Tonal features, intensity, and word order in the perception of prominence. *Journal of Phonetics*, *34*(3), 319–342. doi:10.1016/j.wocn.2005.06.004. (Cited on p. 127)

Valli, A. & Véronis, J. [J.]. (1999). ÉTIQUETAGE GRAMMATICAL DE CORPUS ORAUX: PROBLÈMES ET PERSPECTIVES. *Revue Française de Linguistique Appliquée*, *4*(2), 113–133. (Cited on p. 102).

van Gompel, R. P. G. & Pickering, M. J. (2007). Syntactic parsing. In M. G. Gaskell & G. Altmann (Eds.), *The Oxford handbook of psycholinguistics* (pp. 289–307). Oxford: Oxford University Press. (Cited on p. 60).

van Wijk, C. & Kempen, G. (1987). A dual system for producing self-repairs in spontaneous speech: Evidence from experimentally elicited corrections. *Cognitive Psychology*, *19*(4), 403–440. doi:10.1016/0010-0285(87)90014-4. (Cited on p. 47)

Vasilescu, I., Candea, M., & Adda-Decker, M. (2004). Hésitations autonomes dans 8 langues : une étude acoustique et perceptive. In *Actes de MIDL* (pp. 25–30). (Cited on pp. 116, 121).

Wagner, M. & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, *25*(7-9), 905–945. doi:10.1080/01690961003589492. (Cited on p. 138)

Wagner, P. (2005). Great Expectations - Introspective vs. Perceptual Prominence Ratings and their Acoustic Correlates. In *Proceedings of Interspeech 2005* (pp. 2381–2384). (Cited on p. 128).

Wagner, P., Origlia, A., Avesani, C., Christodoulides, G., Cutugno, F., D'Imperio, M., Escudero Mancebo, D., Gili Fivela, B., Lacheret, A., Ludusan, B., Moniz, H., Ní Chasaide, A., Niebuhr, O., Rousier-Vercuyssen, L., Simon, A. C., Simko, J., Tesser, F., & Vainio, M. (2015). Different parts of the same elephant: A roadmap to disentangle and connect different perspectives on prosodic prominence. In *Proceedings of ICPhS 2015*. (Cited on pp. 127–129).

Wagner, P., Tamburini, F., & Windmann, A. (2012). Objective, Subjective and Linguistic Roads to Perceptual Prominence: How Are They Compared and Why? In *Proceedings of Interspeech 2012*. (Cited on pp. 129, 132).

Wagner, P., Trouvain, J., & Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics*, *48*, 1–12. doi:10.1016/j.wocn.2014.11.001. (Cited on pp. 71, 73, 74)

Warren, R. M. (1970). Perceptual Restoration of Missing Speech Sounds. *Science*, *167*(3917), 392–393. doi:10.1126/science.167.3917.392. (Cited on p. 56)

Watson, D. & Gibson, E. (2004). The relationship between intonational phrasing and syntactic structure in language production. *Language and Cognitive Processes*, *19*(6), 713–755. doi:10.1080/01690960444000070. (Cited on p. 52)

Watson, D. & Gibson, E. (2005). Intonational phrasing and constituency in language production and comprehension. *Studia Linguistica*, *59*(2-3), 279–300. (Cited on p. 137).

Weinreich, U., Labov, W., & Herzog, M. I. (1968). Empirical foundations for a theory of language change. In W. P. Lehmann & Y. Malkiel (Eds.), *Directions for historical linguistics* (pp. 95–195). Austin: University of Texas Press. (Cited on p. 29).

Werner, S. & Keller, E. (1994). Prosodic Aspects of Speech. In E. Keller (Ed.), *Fundamentals of speech synthesis and speech recognition* (pp. 23–40). Chichester: Wiley. (Cited on p. 25).

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & R. Davies (Eds.), *Varieties of Attention* (pp. 63–101). New York: Academic Press. (Cited on p. 13).

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2012). *Engineering psychology and human performance* (4th International Edition). Psychology Press. (Cited on pp. 13, 14).

Wightman, C. & Ostendorf, M. (1994). Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, *2*(4), 469–481. (Cited on p. 129).

Xue, G., Aron, A. R., & Poldrack, R. A. (2008). Common neural substrates for inhibition of spoken and manual responses. *Cerebral cortex (New York, N.Y. : 1991)*, *18*(8), 1923–1932. doi:10.1093/cercor/bhm220. (Cited on p. 47)

Yap, T. F. (2012). *Speech Production Under Cognitive Load: Effects and Classification* (Doctoral dissertation, The University of New South Wales, Sydney, Australia). (Cited on pp. xiii, 6, 165, 170, 211, 212).

Yap, T. F., Ambikairajah, E., Choi, E. H. C., & Chen, F. (2009). Phase based features for cognitive load measurement system. In *Proceedings of ICASSP* (pp. 4825–4828). (Cited on p. 79).

Yap, T. F., Epps, J., Ambikairajah, E., & Choi, E. H. C. (2011). Formant Frequencies under Cognitive Load: Effects and Classification. *EURASIP Journal on Advances in Signal Processing*, *2011*(1), 219253. doi:10.1155/2011/219253. (Cited on p. 80)

Yap, T. F., Epps, J., Choi, E. H. C., & Ambikairajah, E. (2010). Glottal features for speech-based cognitive load classification. In *Proceedings of ICASSP* (pp. 5234–5237). (Cited on p. 80).

Yin, B., Ruiz, N., Chen, F., & Khawaja, A. M. (2007). Automatic cognitive load detection from speech features. In *OzCHI 2007 Proceedings*. (Cited on pp. 79, 80, 175, 213).

Zekveld, A. A. & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: insights from pupillometry. *Psychophysiology*, *51*(3), 277–284. doi:10.1111/psyp.12151. (Cited on pp. 69, 84)

Zellner, B. (1994). Pauses and the temporal structure of speech. In E. Keller (Ed.), *Fundamentals of speech synthesis and speech recognition* (pp. 41–62). Chichester: Wiley. (Cited on pp. 145, 146).

Zellner, B. (1998). Fast and Slow Speech Rate: a Characterization for French. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)* (Vol. 7, pp. 3159–3163). Sydney. (Cited on p. 150).