DA CORRENTE ACÚSTICA À PALAVRA:
ESTÁDIOS DO PROCESSAMENTO DA
PERCEPÇÃO DA FALA

Tânia Patrícia Gregório Fernandes

DOUTORAMENTO EM PSICOLOGIA GERAL

Tese orientada pelo Prof. Doutor Paulo Ventura Fernandes da Rocha
e co-orientada pela Prof. Doutora Régine Kolinsky

2007
The present work was supported by a doctoral grant of Fundação para a Ciência e a Tecnologia – Ministério da Ciência, Tecnologia e Ensino Superior, Portugal, ref SFRH / BD / 12290 / 2003.

Part of this research was also conducted under the Project POCI / PSI / 56901 / 2004 - “Visual phonology and auditory orthography”, funded by Fundação para a Ciência e a Tecnologia – Ministério da Ciência, Tecnologia e Ensino Superior – and European Community FEDER funding.
Palavras-chave:
Segmentação da Fala; Aprendizagem de Línguas Artificiais; Automaticidade; Condições de Audição; Pistas sublexicais.

Key Words:
Speech Segmentation; Artificial Language Learning; Automaticity; Listening Conditions; Sublexical cues.

American Psychological Association (PsycINFO Classification Categories and Codes):

2300 Human Experimental Psychology
2326 Auditory and Speech Perception
2343 Learning and Memory
2346 Attention
ACKNOWLEDGEMENTS

This was probably the most difficult section of this thesis to write. It is difficult to put into words (and even more difficult in just a few words) all the generosity I have received, both professionally and personally, throughout this scientific process.

Since the beginning of my passion for Cognitive Psychology I had two mentors; my first words of gratitude will be for them.

To Professor Carlos Brito-Mendes, who showed to me for the first time what the “black box” is capable of, who is a constant inspiration in all my work in science, and who I will never forget.

To Paulo Ventura, my supervisor, who I esteem immensely as a scientist and as a person, for having shared with me these four years of work with friendship, and for all his help in this scientific process: theoretical reasoning, preparation of experimental materials, data analysis, discussion of results. In difficult times, when I most needed a friendly and encouraging comment, he was always there.

I would also like to thank Régine Kolinsky, my co-supervisor, who I deeply admire, for all her generosity in accepting me as a PhD student, for sharing with me her objective and sharp reasoning, for many stimulating discussions, for all her support in my development as a research scientist, and for being a friend.

Paulo and Régine have been a vital and constant source of reference and stimulation, while at the same time always allowing me to pursue an unconstrained research path in the directions that best suited my intellectual thirst and curiosity.

To José Morais, I am sincerely grateful for all the stimulating, skeptical, and constructive questions that have enormously enriched this work, and for the conversations about Portugal and Literature.

I would also like to express my grateful acknowledgement to: Sven Mattys, for sharing with me his theoretical insights and for having welcomed me at the University of Bristol; and Nicolas Dumay, the Language Group of the Psychology Department of the
University of Bristol, the members of the Unité de Recherche en Neurosciences Cognitives (UNESCOG) and of the Laboratoire de Psychologie Expérimentale of the Université Libre de Bruxelles, for having discussed my work.

The writing of this thesis was enormously benefited by the comments of Paulo, Régine, José, and Sven.

I would also like to thank to Cristina Carvalho, Sandra Fernandes and Luís Querido for all the help in running some of the experiments. To them and to Isabel Leite I am also thankful for their being always available and willing to help me.

This thesis was partly supported by a doctoral grant awarded by the Fundação para a Ciência e a Tecnologia, Ministério da Ciência, Tecnologia e Ensino Superior, Portugal. Its support is here gratefully acknowledge. I would also like to thank to Centro de Psicologia Clínica e Experimental – Desenvolvimento, Cognição e Personalidade, at Faculdade de Psicologia e de Ciências da Educação da Universidade de Lisboa, Portugal, and to the Department of Psychology of Universidade de Évora, as well as to all the students from the Universities of Lisboa and Évora who participated in the experiments reported in this thesis.

And last, but not least, I would like to thank to my closest ones. To my parents, who have supported (both economically and emotionally) all my development, who made me who I am, who accept me unconditionally. To Miguel, Lina, and Daniel for believing in my capabilities, even when I did not. To Pedro who honor me with genuine and enduring care, love, and words of support. To Beatriz and Filipa, for being true friends even through difficult times, and for all the philosophical conversations and interest in this work. To Carmen, Joaquina, and Carmencita who always share great moments with me in Brussels.

This work is dedicated to my mother, who has always stimulated my curiosity, insisted on the importance of learning, and who supports me in everything since the very beginning – from whom I have received so much for so little.
“Céptico como os cépticos, crente como os crentes.

A metade que avança é crente,

a metade que confirma é céptica.

Mas o cientista perfeito é também jardineiro:

Acredita que a beleza é conhecimento.

(A pessoa bela tem um segredo. Descobriu algo).”

Gonçalo M. Tavares,

In “Breves Notas sobre Ciência”
ABSTRACT

Until now, the weighting of general domain and speech-specific cues in speech segmentation was left largely unspecified. In the present work, the impact of different (qualitative and quantitative) listening conditions on the weighting of sublexical sources of information: segmental (transitional probabilities - TPs), suprasegmental (universal prosody and lexical stress) and subsegmental (coarticulation), was investigated in artificial language learning (ALL) settings, on the grounds of Mattys and colleagues theoretical framework. In Experiments 1-3B we evaluated the impact of physical noise on these cues. In Experiment 1, with intact speech, coarticulation overruled TPs. However, its role was highly modulated by signal-quality, while TPs were very resilient to noise. In Experiment 2, universal prosody like TPs was also highly resilient to physical noise, with the former prevailing over the latter, and driving the segmentation process. The impact of stress was rather different than the one of universal prosody. Stress pattern effects only emerged in degraded conditions (Experiment 2 and Experiment 3B).

The impact of cognitive noise was evaluated in Experiments 4-6 through attention-load. In Experiment 3, the impact of cognitive noise on the weighting of TPs and coarticulation was in sharp contrast to the pattern found with physical noise. While coarticulation processing was largely unaffected by a reduction of attentional resources, TPs computation was penalized. In Experiments 5 and 6, we evaluated in a more fine-grained manner the impact of cognitive noise on TP-based segmentation. TPs computation is attentional-resources’ dependent. Yet, it occurs even when the AL is not the focus of attention and scarce resources are available, suggesting some automaticity of statistically-driven segmentation processes.

In Experiments 7-9 we combined ALL settings with conventional experimental techniques, demonstrating that adult listeners are able to use on-line statistical information. Listeners actually treated statistical learning outputs as potential words - the output of ALL exhibits a lexical competition signature revealed by the inhibitory priming effect of novel neighbors (e.g., cathedruke) on lexical decisions to real words (e.g., cathedral) - but only when segmentation cues were congruent (Experiment 8). In Experiment 9, no effect of lexicalization was observed with incongruent cues.

Thus, speech segmentation is largely the product of the available cues and of the listening conditions. Furthermore, these results suggest that current models of spoken-word recognition need to incorporate the role that congruency between segmentation cues play in both speech segmentation and word learning.
RESUMO

O problema de Segmentação da Fala é um dos grandes desafios da Psicolinguística Cognitiva e é, também, o tema central do trabalho descrito nesta tese.

Dois mecanismos mentais parecem estar envolvidos no processo de segmentação. Primeiro, múltiplos candidatos lexicais são activados e competem directamente para o reconhecimento (e.g., McQueen, Norris, & Cutler, 1994). Em segundo lugar, os ouvintes exploram múltiplos índices (ou pistas) sublexicalais, disponíveis na corrente acústica.

Estas pistas sublexicalais podem ser, de forma geral, classificadas como: pistas suprasegmentais, relativas à prosódia (e.g., o acento lexical); pistas segmentais, a informação estatística relativa à sequência das unidades linguísticas (e.g., fonemas, sílabas) que ocorrem numa língua natural (i.e., probabilidades fonotácticas, Vitevitch & Luce, 1998, 1999) ou numa língua artificial (probabilidade transitacional entre sílabas, Saffran, Aslin, & Newport, 1996a; Saffran, Newport, & Aslin, 1996b); e pistas subsegmentais relativas a aspectos acústicos de baixo-nível (e.g., a coarticulação, Mattys, 2004).

Contudo, até ao momento actual, a maioria das investigações, devotadas ao tema do papel das pistas sublexicalais na segmentação da fala em adultos, tem utilizado técnicas experimentais convencionais. Através destas técnicas, como o gating, o priming de modalidade cruzada, o word-spotting, e o registo dos movimentos oculares, o papel dos índices sublexicalais na segmentação é indirectamente inferido de acordo com a activação lexical de candidatos compatíveis com o input linguístico. Portanto, o recurso a estes paradigmas não permite avaliar o papel dos índices sublexicalais per se, uma vez que aspectos de activação e competição lexical (i.e., informação alto nível) estão igualmente envolvidos no desempenho dos ouvintes. Esta limitação não está presente no paradigma de Aprendizagem de Línguas Artificiais (ALAs) proposto por Saffran e col. (Saffran et al., 1996a, 1996b), permitindo o estudo de informação sublexical em condições de minimização da disponibilidade de informação de alto-nível.

Saffran et al. (1996b) demonstraram, recorrendo ao Paradigma de ALA, que mesmo para ouvintes adultos, para quem a informação de alto-nível é prioritária no processo de segmentação, a informação de tipo estatístico é utilizada na segmentação. Neste paradigma, constituído por duas fases, os ouvintes são primeiro familiarizados com uma corrente acústica contínua composta pela concatenação de um repertório limitado de sílabas Consoante-Vogal. Nesta corrente contínua não existe nenhuma informação acústica que possa indicar ao ouvinte onde uma “palavra” (da LA) começa e onde acaba. Após a fase de familiarização, o ouvinte realiza um teste de escolha forçada onde, em cada ensaio, deverá decidir entre dois estímulos, qual deles corresponde a uma “palavra” da LA. A comparação do desempenho dos ouvintes com o nível do acaso permite avaliar se houve aprendizagem, uma vez que esta será expressa por um desempenho significativamente superior ao nível do acaso (i.e., o ouvinte escolheu com maior
frequência os estímulos apresentados que na realidade eram “palavras” da LA do que os estímulos constituídos pelo mesmo repertório de sílabas mas que não eram “palavras” da LA).

Uma vez que os estímulos, apresentados na segunda fase, ocorreram embebidos numa corrente acústica contínua durante a primeira fase, a observação de efeitos de aprendizagem sugere que os ouvintes são sensíveis à informação que está disponível na corrente apresentada, sendo capazes de utilizá-la ao serviço da segmentação da fala.

Normalmente, a informação disponível corresponde a informação segmental de natureza estatística, i.e., a probabilidade transicional (PT) entre sílabas adjacentes. A PT corresponde à probabilidade com que uma primeira sílaba (X) é capaz de prever a que se lhe segue (Y), i.e., $PT = \frac{frequência_XY}{frequência_X}$. De facto, não só numa LA, mas em qualquer língua natural, a probabilidade de uma sílaba prever a seguinte é maior, quando as duas sílabas pertencem à mesma palavra, do que quando ocorrem em posição adjacente na corrente mas com um ponto de segmentação entre si (Perruchet & Peereman, 2004; Swingley, 2005).

Como esta computação estatística não exige qualquer conhecimento lexical, esta poderia ser a primeira pista de segmentação disponível aos bebés (Thiessen & Saffran, 2003), ocupando, por isso, uma posição central, não só na aquisição lexical, mas também na aquisição de outras pistas de segmentação (Mattys, White, & Melhorn, 2005).

O paradigma de ALAs é também uma metodologia especialmente vantajosa para o estudo da segmentação da fala em adultos. Como a informação lexical (prioritária para o sistema perceptivo adulto) não se encontra disponível numa LA, este paradigma permite o estudo rigoroso do papel de diferentes fontes de informação sublexical na segmentação. O controlo sistemático da informação disponível e do tempo de exposição permitem igualmente a redução do conjunto de variáveis não controladas e a análise sistemática do papel de diferentes pistas de segmentação.

Recentemente, Mattys e colaboradores (Mattys et al., 2005) apresentaram a primeira proposta teórica relativa à organização das diferentes fontes de informação, quer lexical quer sublexical, utilizadas ao serviço da segmentação da fala. Esta proposta considera que a segmentação da fala é produto, quer das fontes de informação que se encontram disponíveis, quer das condições interpretativas. As condições de audição têm um papel gradativo (e não de tipo tudo-ou-nada) na utilização de qualquer pista de segmentação. Nesta proposta teórica, os índices de segmentação encontram-se organizados hierarquicamente em três níveis. No nível superior da hierarquia encontra-se representada a informação lexical e pós-lexical, prioritária para a segmentação da fala em ouvintes com um léxico mental totalmente desenvolvido. Nos níveis inferiores estão representadas as informações sublexical que terão um papel determinante na segmentação da fala, quando a informação lexical se encontra indisponível ou é reduzida. No segundo nível da hierarquia encontram-se informações de tipo segmental (estatístico) e subsegmental. Estas serão a informação determinante para a segmentação quando apenas informação sublexical se encontre disponível na corrente. O nível inferior da hierarquia corresponde à informação suprasegmental, de natureza prosódica, que funciona como heurística de último recurso em condições de fraca qualidade do sinal (e.g., sobreposição de ruído branco).

Na investigação da segmentação da fala é fundamental adoptar uma perspectiva integrativa, no sentido em que, para além do estudo isolado de diferentes pistas de segmentação,
estas fontes de informação deverão ser estudadas em condições combinatórias. Por outras
palavras, o estudo integrado permite a compreensão de como diferentes fontes de informação são
integradas (i.e., quando sugerem as mesmas hipóteses de segmentação, a disponibilidade de
diferentes pistas terá um benefício na segmentação da fala?) e como são ponderadas (i.e.,
quando diferentes pistas de segmentação sugerem limites de palavra distintos, qual a pista
considerada mais fidedigna?).

Uma outra frente de investigação abordada no presente trabalho diz respeito às
condições de audição. Tal como proposto por Mattys e col. (2005), estas condições têm um
impacto modulador na ponderação das pistas de segmentação disponíveis na corrente (i.e., a
fiabilidade das pistas depende largamente da qualidade do sinal), e estas condições não se
circunscrevem apenas à qualidade física do sinal. Há mais no ruído do que a simples degradação
física da corrente. Contudo, até ao presente, nenhum estudo investigou o papel de tipos de ruído
qualitativamente diferentes na organização hierárquica das diferentes fontes de informação.

Com o objectivo de avaliar o impacto das condições de audição na ponderação de
diferentes tipos de informação sublexical, no presente estudo foi realizado um conjunto de
experiências adoptando o paradigma de ALAs.

Nesta investigação, foi avaliado o impacto do ruído físico (i.e., sobreposição de ruído
branco na corrente de fala a diferentes rácios sinal-ruído) e de ruído cognitivo (i.e., redução dos
recursos atencioanais disponíveis por realização de tarefas concorrente, atencionalmente
exigentes), na ponderação de fontes de informação sublexical, representadas na proposta de
Mattys e col., quer no mesmo nível (a co-articulação, informação de tipo subsegmental, acústico,
de baixo-nível vs. a PT entre sílabas adjacentes, informação segmental estatística; e também
entre diferentes tipos de informação suprasegmental – a prosódia universal vs. o acento lexical),
quer em níveis diferentes (PTs vs. prosódica) da hierarquia.

Nas Experiências 1-3B foi avaliado o impacto do ruído físico (i.e., por sobreposição de
ruído branco ao sinal de fala) na ponderação dos três tipos de informação sublexical.

Na Experiência 1, com sinal intacto, a coarticulação demonstrou-se uma pista de
segmentação importante, superando o papel da informação estatística na segmentação. Contudo,
o estatuto da informação coarticulatória é modelado pela qualidade física do sinal, enquanto a
informação estatística se mantém resistente a uma degradação deste tipo. De facto, a informação
estatística foi capaz de conduzir o processo de segmentação a níveis similares em condições de
sinal intacto e de sinal fortemente degradado.

Na Experiência 2, quer a informação prosódica universal, quer as PTs, demonstraram-se
insensíveis a degradação física do sinal. Além disso, a informação prosódica foi a pista
considerada mais fiável no processo de segmentação em qualquer condição de audição.

O papel do acento lexical (i.e., informação suprasegmental, específica da língua em
questão) na segmentação da fala (Experiência 3) é muito diferente do papel da informação
prosódica universal. De facto, apenas em condições de ruído físico, se observaram efeitos do
padrão de acento lexical na segmentação da fala.

O impacto do ruído cognitivo (i.e., sobrecarga atencional) na ponderação das pistas
sublexicais é muito diferente do impacto do ruído físico. Na Experiência 4, verificou-se que a
ponderação da informação estatística e da coarticulação observada em condições de ruído cognitivo é aproximadamente uma imagem em espelho da observada em condições de ruído físico. Em condições de redução drástica dos recursos atencionais disponíveis, a segmentação conduzida pela coarticulação não foi afectada, enquanto a informação estatística se revelou muito sensível a esta degradação. Nas Experiências 5 e 6, avaliámos de forma mais fina o impacto do ruído cognitivo na segmentação conduzida pela informação estatística. Nestas duas experiências foi demonstrado que a computação de PTs é dependente dos recursos atencionais, mas não de um mecanismo selectivo de atenção. Contudo, mesmo quando, os recursos atencionais disponíveis são drasticamente reduzidos, a informação estatística continua a ser capaz de conduzir o processo de segmentação.

Nas três últimas experiências (Experiências 7-9) deste estudo combinámos técnicas convencionais com o paradigma de ALAs. Na Experiência 7, utilizando o word-spotting, apresentámos uma evidência directa da utilização on-line da informação estatística na segmentação da fala por ouvintes adultos.

Nas Experiências 8 e 9 avaliámos qual a natureza das representações fonológicas extraídas de uma corrente continua por aprendizagem estatística. Na Experiência 8 foi demonstrado que ouvintes adultos tratam o output da aprendizagem estatística como potenciais palavras (i.e., efeito de priming inibitório dos novos vizinhos - produtos da segmentação da LA -; e.g., /livek/1 na decisão lexical de palavras reais (e.g., fivela). Este estatuto lexical foi apenas observado quando as pistas de segmentação disponíveis na corrente de fala eram congruentes. Na experiência 9, em condição de pistas incongruentes, não foi observado qualquer efeito de lexicalização do output do procedimento de segmentação utilizado pelos ouvintes.

Desta forma, a segmentação da fala é largamente o produto da ponderação das fontes de informação disponíveis na corrente, e das condições interpretativas (cf. Mattys et al., 2005). Para além disso, no presente trabalho é também sugerido que a ponderação das pistas sublexicais de segmentação depende de dois factores: (i) a sua “generalidade de domínio”; e (ii) o seu papel numa língua em particular (i.e., pista universal vs. pista específica da língua em questão; e.g., informação prosódica universal vs. acento lexical).

Assim, o padrão de resultados observado neste estudo sugere que os modelos actuais de reconhecimento da palavra falada deverão incorporar a importância do papel da congruência das pistas de segmentação, bem como da natureza das mesmas, quer na segmentação da fala (i.e., processamento on-line), quer na aprendizagem de novas palavras (i.e., alterações estruturais de longo prazo).
Acronyms & symbols

Artificial Language .......................... AL
Artificial Language Learning .................. ALL
Adaptive Resonance Theory .................. ART
Brazilian Portuguese ......................... BP
Consonant .................................. C
Distributed Cohort Model .................... DCM
European Portuguese ......................... EP
Fundamental Frequency ....................... F0
Implicit Learning ............................. IL
Inter-stimulus Interval ......................... ISI
Intonational Phrase .......................... IP
International Phonetic Alphabet .............. IPA
Long Term Memory ........................... LTM
Metrical Segmentation Strategy ............... MSS
Neighborhood Activation Model .............. NAM
Possible Word Constraint ...................... PWC
Rapid Visual Serial Presentation ............. RVSP
Reaction Time ................................ RT
Stimulus Onset Asynchrony ................... SOA
Transitional Probability ....................... TP
Universal Grammar .......................... UG
Voice Onset Time ............................ VOT
Vowel ........................................ V
Working Memory ............................. WM
# TABLE OF CONTENTS

## PART I – THEORETICAL BACKGROUND

1. General Introduction .................................................................................................................. 3

1.1. From the continuous stream into discrete units of meaning ............................................... 5
  1.1.1. In the beginning there was the word ................................................................. 5
  1.1.2. Crossroads into the speech stream ................................................................. 10
  1.1.3. Issues and issues ......................................................................................... 18

1.2. Conventional Experimental Paradigms ............................................................................... 23
  1.2.1. Gating ........................................................................................................ 23
  1.2.2. Word-spotting ............................................................................................ 24
  1.2.3. Eye-tracking ............................................................................................... 25
  1.2.4. Cross-modal priming ............................................................................... 29

1.3. The Artificial Language Learning Paradigm .................................................................... 31
  1.3.1. ALL and Implicit Learning ........................................................................ 33
  1.3.2. ALL and Speech Segmentation ................................................................. 35

1.4. The hierarchical organization frame of Mattys and colleagues ...................................... 37
  1.4.1. Experimental evidences ............................................................................. 38
  1.4.2. Developmental and cross-linguistic implications ........................................ 39

1.5. An overview of the present thesis .................................................................................... 42
  1.5.1. The weighting of different sublexical sources of information in speech segmentation .................................................. 43
  1.5.2. The listening conditions ............................................................................. 46
  1.5.3. An on-line evidence of statistical speech segmentation .................................. 51
  1.5.4. The nature of speech segmentation outputs .............................................. 54

## PART II – EMPIRICAL CHAPTERS

2. Statistical and Coarticulatory Cues to word boundaries ......................................................... 61

2.1. Introduction ....................................................................................................................... 61
  2.1.1. Coarticulation – a subsegmental cue .............................................................. 63
  2.1.2. TPs – a segmental cue .................................................................................. 64
6. On-line measures and lexicalization during ALL .................................................. 209

6.1. Introduction ........................................................................................................ 210
   6.1.1. The nature of statistical segmentation outputs ........................................... 213
   6.1.2. An overview of the experiments ................................................................. 216

Experiment 7. On-line statistically-driven speech segmentation

6.2. Method .................................................................................................................. 220
   6.2.1. Participants ................................................................................................. 220
   6.2.2. Material ..................................................................................................... 220
   6.2.3. Procedure .................................................................................................. 222

6.3. Results and Discussion ....................................................................................... 223
   6.3.1. Spotting Responses .................................................................................... 223
   6.3.2. RT data .................................................................................................... 226

Experiment 8. The status of statistical segmentation outputs

6.4. Method .................................................................................................................. 232
   6.4.1. Participants ................................................................................................. 232
   6.4.2. Material ..................................................................................................... 232
   6.4.3. Procedure .................................................................................................. 240

6.5. Results and Discussion ....................................................................................... 240
   6.5.1. ALL performance ....................................................................................... 240
   6.5.2. Lexicalization test ...................................................................................... 242

Experiment 9. The status of the output of the segmentation procedure adopted by listeners

6.6. Method .................................................................................................................. 249
   6.6.1. Participants ................................................................................................. 249
   6.6.2. Material ..................................................................................................... 250
   6.6.3. Procedure .................................................................................................. 251

6.7. Results and Discussion ....................................................................................... 251
   6.7.1. ALL performance ....................................................................................... 251
   6.7.2. Lexicalization test ...................................................................................... 253

6.8. General Discussion .............................................................................................. 256

PART III – GENERAL DISCUSSION AND CONCLUSIONS

7. General Discussion ................................................................................................. 273
   7.1. Main findings of the present study .................................................................. 274
     7.1.1. Speech segmentation with physically degraded signal ............................ 274

XX
## INDEX OF TABLES

**Table 1:** Orthographic translation of a sample of the stream heard in the familiarization phase in the three cue conditions of Experiment 1. .......................................................... 70

**Table 2:** Performance pattern (proportion of TP-Word responses, in percentage) according to cue condition (single cue; congruent cues; incongruent cues) and Signal Quality (intact speech; 22dB SNR; 10dB SNR), in Experiment 1, considering the TP gradient of the TP-words of the AL (High-TP; low-TP) .......................................................... 76

**Table 3:** Orthographic translation of a sample of the stream heard in the familiarization phase in the four cues conditions and of the AL-stimuli (TP-words; Part-words 3#12; Part-words 23#1) of Experiment 2. ......................... 101

**Table 4:** Average proportions of TP-word responses (in percentage), separately for cues condition (single cue; incongruent cues - 1st syl; incongruent cues - 2nd Syl; and congruent cues) and signal quality condition (Intact vs. Mildly Degraded), considering Part-words type (3#12 vs. 23#1) in Experiment 2. ......................... 107

**Table 5:** Orthographic translation of a sample of the stream heard in the familiarization phase in the four conditions of stress location of Experiment 3B. .......................................................... 124

**Table 6:** Average proportions of TP-Word responses (in percentage) in Experiment 2B, separately for each input-intelligibility condition (intact; mildly degraded; strongly degraded) and lexical stress location (TP-unstressed; 1st syllable; 2nd syllable; 3rd syllable), considering the TP-level of the AL-words (high- vs. low-TP). .................. 127

**Table 7:** Performance pattern (mean percentage of Correct Detections – CD – and false alarms – FA) for the Three AL-conditions, according to the number and congruence of segmentation cues available (single cue; congruent cues – congruent ; incongruent cues – incongruent), in the RSVP task, according to visual block (block1; block 2; block 3) and $d'$ scores in Experiment 4. .......................................................... 155

**Table 8:** Average proportions of TP-Word responses (in percentage), separately for each cue condition (single cue; congruent cues; incongruent cues) and cognitive noise condition (low vs. high load), considering the TP- gradient of the AL words (high- vs. low-TP-words) in Experiment 4. ................................. 157

**Table 9:** Average proportions of TP-Word responses (in percentage), separately for each familiarization time (7-; 14-; and 21-min) and attentional load (low vs. high load) condition, considering the TP level of the AL words (high- vs. low-TP-words) in Experiment 5. .......................................................... 183
Table 10: Average proportions of TP-Word responses (in percentage), separately for each learning condition (Incidental and Intentional) and attentional resources condition (low-attention vs. high-attention load), considering the TP-level of the AL words (high- vs. low-TP-words) in Experiment 6. .................................................. 195

Table 11: Examples of materials used (in phonological form, according to IPA) in the AL familiarization phase (AL-Fam.), two-forced choice (ALL-test), and lexicalization test (Lexical Decision – LD on base-words) in Experiment 8, separately for the two artificial languages (ALs: AL_A and AL_B). ................................. 231

Table 12: Base-Words characteristics by block (Blocks A and B; Experiment 8 and 9: mean log frequency; neighborhood density, mean transitional probability (TP between first and second syllable), Uniqueness Point (UP), and mean duration (in ms), as well as t-values (all ps > .5). .............................................. 233

Table 13: Mean RTs (measured from target onset) for Base-Words A and B (BWd A; BWd B) and the Priming Effect (P.E., the difference in lexical decision latencies between primed and unprimed conditions) separately for each AL-condition (AL_A and AL_B) and moment of testing (Immediately and Post-1 week) in Experiment 8. .... 243

Table 14: Examples of materials used in AL-familiarization phase (AL-Fam.), ALL-test, and Lexicalization test (Lexical Decision – LD on Base-Words) on Experiment 9, regarding the two Artificial Languages (ALs). ................................. 248

Table 15: Mean RTs (measured from target onset) for Base-Words A and B (BWd A; BWd B) and the Priming Effect (P.E., the difference between primed and unprimed conditions) separately for each AL-condition (AL_A2 and AL_B2) and moment of testing (Immediately and Post-1 week) in Experiment 9. ................................. 253
INDEX OF FIGURES

Figure 1: Hierarchical organization of speech segmentation cues according to listening conditions, as proposed by Mattys and colleagues. The triangle represents the weighting of segmentation cues as well as their resilience to physical degradation of the signal (i.e., the base of the triangle corresponds to the strongest sensitivity to physical noise). ................................................................. 38

Figure 2: Performance pattern (proportion of TP-Word responses, in percentage) according to cue condition (single cue; congruent cues; incongruent cues) and of the signal quality (intact speech; 22dB SNR; 10dB SNR) – Experiment 1. ................................................................. 78

Figure 3: ALL performance pattern (proportion of AL-word responses, in percentage), in Experiment 2, broken-down by Part-word Type (3#12; 23#1), according to Input-Intelligibility (Intact; Degraded) and Cues available (single cue; incongruent cues first syllable; incongruent cues second syllable; congruent cues). ........................................ 108

Figure 4: Listeners’ proportion of responses on the three-alternative forced choice stress location task on AL-stimuli, broken-down by syllable detection responses (1st syl; 2nd syl; 3rd syl) in the two acoustic cue conditions (Unstressed vs. Stressed) in Experiment 3A. ................................................................. 117

Figure 5: Performance pattern (proportion of AL-word responses, in percentage) broken-down by TP-level (High-TP; Low-TP) of AL-words, according to Signal condition (Intact; mildly degraded - 22dB SNR; strongly degraded – 10dB SNR) and Lexical Stress location (TP – No stress; 1st Syllable; 2nd Syllable; 3rd Syllable) in Experiment 3B. Vertical bars denote standard error of the mean on each condition. ........................................ 126

Figure 6: Performance pattern (proportion of AL word responses, in percentage) broken-down by TP level (High-TP-words; Low-TP-words), according to attentional load condition (High-attention Load; Low-attention Load) and familiarization time (7-min; 14-min; 21-min), in Experiment 5. ................................................................. 184

Figure 7: Performance pattern (proportion of AL-word responses, in percentage) broken-down by TP level (High-TP-words; Low-TP-words) of AL words, according to attentional resources condition (Figure 7a: High-attention Load; Figure 7B: Low-attention Load) and learning condition (Incidental learning; Intentional learning) in Experiment 6. ................................................................. 196

Figure 8: Mean Adjusted RT (in ms; from AL-stimuli offset) for trisyllabic responses in spotting TP-words and part-words, separately for initial and final AL-stimuli-bearing sequences (in the Preceding vs. Following lists, respectively), in Experiment 7. ................................................................. 227
Figure 9: Mean percentage of TP-word choices in the ALL-test (i.e., two-alternative forced choice test) for each participant, according to the moment of testing: immediately after the AL-familiarization phase (i.e., immediate) and one week after with no intervening AL-familiarization period in Experiment 8. .......................... 241

Figure 10: Effects of AL-familiarization, according to the AL to which listeners were exposed (LA; LB), on lexical decision’s response to existing lexical items immediately after AL-familiarization: Mean RTs (in milliseconds) for the two blocks of Base-Words (Base-Words A; Base-Words B) in Experiment 8 .......................... 244

Figure 11: Inhibitory Effect (computed as the difference between lexical decision latencies to the primed block of Base-Words and the unprimed Block of Base-Words), according to the AL to which listeners were exposed (AL_A; AL_B) and the moment of testing (Immediately after and 1-week after) in Experiment 8. .......................... 246

Figure 12: Mean percentage of TP-word choices in the ALL-test for each participant, according to the moment of testing: immediately after the AL-familiarization phase (i.e., immediate) and one week after with no intervening AL-familiarization period, in Experiment 9. .......................................................... 252

Figure 13: The theoretical proposal outlined in the present thesis allied with Mattys and colleagues proposal for language-specific cues (see also Figure 1 in section 1.4. of chapter one) ................................. 293
PART I
THEORETICAL BACKGROUND
Chapter 1.

General Introduction

“(…) meus ombros se retesavam não pelo que eu via, mas no afã de captar ao menos uma palavra. Palavrinha? Sem a mínima noção do aspecto, da estrutura, do corpo mesmo das palavras, eu não tinha como saber onde cada palavra começava ou até onde ia. Era impossível destacar uma palavra da outra, seria como pretender cortar um rio à faca…”

Chico Buarque (2004), In “Budapeste”

Understanding speech in our native language is accompanied by a subjective impression of “hearing” a discrete series of word-units clearly separated from one another. However, when we hear a speaker of a foreign language, our subjective impression is rather different. In that situation, any first attempt to segment the speech flow would be like “trying to cut a river with a knife” (Buarque, 2004). This is in fact closer to the objective reality: The speech stream (either in our native language or in an unknown one) is a continuous stream with few reliable cues to word-boundaries (e.g., Klatt, 1980; Liberman & Studdert-Kennedy, 1978). Thus,
Understanding how we parse this quasi-continuum signal into discrete lexical units is one of the main challenges faced by psycholinguistic research. This is the primary focus of the research reported on this thesis, in the context of a full, mature speech perception system.

In 1987, Frauenfelder and Tyler stated, in their manuscript on Cognition’s special issue about spoken word recognition, that until that moment in psycholinguistic History “spoken word recognition has been a neglected area of study”. After twenty years, a vast bulk of research has been devoted to this cognitive domain, in particular to speech segmentation.

The speech segmentation problem is the main focus of the present thesis. In this chapter I will present the mechanisms involved in speech segmentation – high-level word recognition and signal derived sources of information – as well as the evidences demonstrating that even in adulthood both types of mechanisms act together in speech segmentation. The conventional paradigms demonstrating the role of sublexical segmentation cues in adulthood will be also outlined. The artificial language learning paradigm implemented by Saffran and colleagues (Saffran, Aslin, & Newport, 1996a; Saffran, Newport, & Aslin, 1996b) will be described as a potential fruitful tool for understanding the role of sublexical cues in the absence (or almost so) of available high-level information. The association between this paradigm and the Implicit Learning field, as well as the empirical speech segmentation evidences derived from studies using this paradigm will also be described. After contextualizing the segmentation mechanisms on the grounds of empirical evidences, the theoretical framework that underlie the study described in the present thesis will be presented (i.e., the hierarchical organization frame of
Mattys and colleagues), considering both the experimental evidences and the developmental and cross-linguistic implications of this theoretical proposal. Finally, the present thesis will be briefly overviewed considering the theoretical aspects studied in this work as well as their relation with each one of the nine experiments presented through this thesis.

1.1. From the continuous stream into discrete units of meaning

Two major mechanisms that may help solve the speech segmentation problem were at first independently proposed on the grounds of their locus of operation (Gow & Gordon, 1995). First, at lexical level, multiple word candidates are activated by the input and compete with each other for recognition (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; McQueen, Norris, & Cutler, 1994; Norris, McQueen, & Cutler, 1995). Second, at sublexical level several signal-derived sources of information, probabilistically associated with word-boundaries, were pointed out as potential segmentation cues (e.g., Brent & Cartwright, 1996; Cairns, Shillcock, Chater, & Levy, 1997; Christiansen, Allen, & Seidenberg, 1998; Christiansen & Curtin, 2005).

1.1.1. In the beginning there was the word

*Lexical accounts* incorporate mechanisms by which the identification of individual lexical items contributes to the detection of word boundaries.

*Sequential Recognition and Speech Segmentation*

In sequential models of word recognition such as the initial version of the
Cohort Model (Marslen-Wilson & Welsh, 1978) the recognition system processes incoming speech in a maximally efficient manner. Although speech segmentation is not directly addressed, the process of spoken word recognition indirectly enables that operation. Multiple lexical candidates are activated according to the initial portion of a word, and these are progressively eliminated by mismatching input until a single lexical item remains. As a consequence, a word with an early uniqueness point (i.e., for which the sequence of sounds diverges from other lexical items early on, before its final segment) rapidly enables not only a faster recognition (e.g., Grosjean, 1980, 1985; Mattys & Clarke, 2002; for strategic evidences on uniqueness point see Radeau, Morais, Mousty, & Bertelson, 2000) but also predicts the location of upcoming word boundaries.

However, investigation on psycholinguistic lexical databases (e.g., Luce, 1986) has demonstrated that many words do not diverge from all other candidates until their final segment, being thus embedded in longer lexical items (e.g., cap embedded in captain; bone embedded in trombone). Furthermore, behavioral data also suggest that nonaligned candidates directly compete with each other for recognition. Shillcock (1990) found that words such as trombone, presented in auditory sentence contexts, produced priming of responses to the visual word rib, an associate of the embedded word bone. McQueen et al. (1994) demonstrated, using word-spotting, that lexical candidates straddling different parts of the input string compete with each other for recognition: the detection of an embedded word-target (e.g., mess) was penalized when the embedding string corresponded to a word onset (e.g., domes, the beginning of domestic) in comparison to when the string did not correspond to the beginning of any real word in listeners’ native language (e.g.,
Vroomen and de Gelder (1995) also demonstrated that, although lexical effects take time to develop, the direct competition between word candidates is an important component of word recognition and speech segmentation. In ambisyllabic nonsense contexts, in which the coda of the first syllable also belongs to the next syllable as onset (e.g., *melkeum* and *melkaam*), the cross-modal priming effect observed for a visual word-target (e.g., melk, in Dutch milk) was stronger when the second syllable of the disyllabic nonsense prime was the initial syllable of few words in Dutch (e.g., *melkeum*) in comparison to when it was the beginning of many words (e.g., *melkaam*). This pattern of results is due to the fact that in the former case the first syllable of the prime, which corresponded to the word-target, received less inhibition from the (few) competitors initial overlapping with the second syllable of the prime than in the latter case. Tabossi, Burani, and Scott (1995) also demonstrated using cross-modal priming that participants presented with fragments of speech that could be segmented either as a long carrier word (e.g., *visite* – Italian for visits) or as a shorter embedded word (e.g., *visi tediati* – Italian for faces bored), were still considering the longer word as a potential candidate for recognition when they heard the first syllable of the second word (e.g., “tediati”). Gow and Gordon (1995) also demonstrated that both two-word (e.g., *two lips*) and one-word (e.g., *tulips*) utterances led to similar priming effects on the recognition of a semantic associate of the one-word utterance (e.g., *flower*). While hard to reconcile with the Cohort Model (Marslen-Wilson & Welsh, 1978), these evidences support models in which segmentation is a byproduct of lexical competition.


**Lexical Competition and Speech Segmentation**

Models in which lexical competition is conceived as occurring directly between word candidates compatible with the speech input, consider speech segmentation as a byproduct of lexical competition. As clearly defined by McClelland and Elman (1986) on TRACE model:

“*Word identification and segmentation emerge together as part and parcel of the process of word activation*” (p. 61).

“*These remarkably simple mechanisms of activation and competition do a very good job of word segmentation, without the aid of any syllabification, stress, phonetic word boundary cues*” (p. 64).

Other models that incorporate lexical competition include the Neighborhood Activation Model (*NAM*; Luce & Pisoni, 1998; Luce, Goldinger, Auer, & Vitevich, 2000), the revised version of the Cohort Model (Marslen-Wilson, 1987; Marslen-Wilson, Moss, & van Halen, 1996), and the Shortlist Model (Norris, 1994).

These models differ on the specific mechanism of lexical competition: either by direct inhibition as proposed by TRACE (McClelland & Elman, 1986) and Shortlist (Norris, 1994) models, and also by the Distributed Cohort Model (*DCM*, Gaskell & Marslen-Wilson, 1997; 1999; 2002), or by the indirect mediation of a decision stage as suggested by the Cohort Model (Marslen-Wilson, 1987; Marslen-Wilson et al., 1996) and the NAM (Luce & Pisoni, 1998; Luce et al., 2000). Nevertheless, all models agree that lexical activation will be reduced when a greater number of lexical candidates are compatible with the same speech stream, and lexical candidates with large phonological overlapping will act as strong competitors for recognition.
Models such as TRACE (McClelland & Elman, 1986) and Shortlist (Norris, 1994) incorporate lexical competition between nonaligned candidates. For example, “cap” and all other words (even nonaligned candidates) in which the sequence /kæp/ is embedded (e.g., captain, captive) will directly compete for recognition through lateral inhibition. Thus, on both models, short words (e.g., cap) would only win the recognition process after its offset, when longer competitors are ruled out.

TRACE vs. Shortlist

Note however that although TRACE (McClelland & Elman, 1986) and Shortlist (Norris, 1994) suggest the same mechanism of lexical competition, expressed as lateral inhibition, between activated word candidates, they differ on three fundamental aspects. First, in TRACE (McClelland & Elman, 1986), all lexical nodes are potential candidates for recognition, continuously increasing or decreasing in activation as a function of their match with the incoming signal whereas in Shortlist (Norris, 1994) competition only takes place within a short list of word candidates that specifically match the input to some preset criterion. Second, TRACE (McClelland & Elman, 1986) is an expression of a highly interactive view of spoken word recognition in which there is a continuous two-way flow of information (i.e., both bottom-up and top-down); on the contrary, Shortlist (Norris, 1994) is an autonomous model, entirely bottom-up in its operation (see for a discussion between interactive vs. autonomous models on modularity grounds: Bowers, & Collin, 2004; McQueen, Norris, & Cutler, 2006). Importantly, while in TRACE speech segmentation is an exclusive byproduct of lexical competition, and thus no space remains for any effect of sublexical segmentation cues, Shortlist not only considers,
as TRACE, that segmentation can be a byproduct of lexical competition, but also incorporates explicit sublexical segmentation strategies (e.g., the Metrical Segmentation Strategy – MSS; Cutler & Norris, 1988), unified in the operation of the Possible Word Constraint (PWC; Norris, McQueen, Cutler & Butterfield, 1997).

1.1.2. Crossroads into the speech stream

At a fundamental level, the speech segmentation problem faced by adults and young children is the same (Brent & Cartwright, 1996): adults sometimes encounter and learn novel words and children often recognize familiar words within continuous utterances. However, since adults are already familiar with a vastly larger proportion of words a lexical segmentation strategy can be, in adulthood, a very proficient mechanism (e.g., McClelland & Elman, 1986; Norris, 1994) but it has virtually no value in the very beginning of language onset. The conception of a solely lexical segmentation strategy, during language acquisition, would pose a *chicken-and-egg* problem (cf. Cairns et al., 1997): the identification of a particular stretch of speech as a meaningful unit would presuppose recognizing what that unit is, but that would only be possible once segmentation has been carried out.

Since infant-directed speech does not reliably provide isolated words (Aslin, Woodward, LaMendola, & Bever, 1996), processing multi-word utterances is critical for word learning during language acquisition. On this view, children must consider as candidate words sound sequences that they have never heard in isolation. Thus, *sublexical* (bottom-up) sources of information could enable the *bootstrapping* of lexical boundary detection (e.g., Brent & Cartwright, 1996; Cairns, et al., 1997; Gleitman & Wanner, 1982), and hence would have a fundamental role in language
acquisition.

Sublexical sources of information can be classified in three types of information: suprasegmental, subsegmental, and segmental information.

**Suprasegmental Information**

Suprasegmental information, i.e., prosody, was one of the first sublexical sources of information to be proposed as a speech segmentation cue. Gleitman and Wanner (1982) suggested that prosody (i.e., rhythm, intonation) might play an important role in the acquisition of syntax, acting as a “prosodic bootstrapping”, and it seems to be one of the more precocious segmentation cues.

Newborns (as well as non-human primates: Ramus, Hauser, Miller, Mones, & Mehler, 2000; Tincoff, et al., 2005) are able to distinguish languages based on their rhythmic properties (e.g., Ramus, 2002; Ramus et al., 2000; Nazzi & Ramus, 2003). Two-month-olds are sensitive to Intonational Phrases\(^1\) (\(\text{IPs}\); e.g., Dehaene-Lambertz & Houston, 1998) and at 6-month to preboundary length, pitch, and pause in clause segmentation (Seidl, 2007). After 7.5 month, infants are also sensitive to prosodic edges (Seidl & Johnson, 2006).

Eight-month-olds English listeners are already able to use their native language metrical prosody for assisting speech segmentation (i.e., strong syllables are defined as possible word-beginnings; Curtin, Mintz, & Christiansen, 2005; Jusczyk, Houston, & Newsome, 1999a) and 9-month-olds treat strong-weak disyllables (i.e., trochaic units) rather than weak-strong ones (i.e., iambic units) as

---

\(^1\) A wide variety of terms are found in the literature, yet two levels above prosodic words are considered in prosodic hierarchy: Intonational Phrases (\(\text{IPs}\)), the highest prosodic unit, mostly corresponding to whole clauses or sentences and often marked by a pause at the end; and Phonological Phrases, below IPs, also referred as intermediate IPs or accentual groups (Grice, 2004; Werner & Keller, 1994a). Both types of phrases are acoustically characterized by a final lengthening with a falling pitch contour (at their right edge; see Christophe, Peperkamp, Pallier, Block, Mehler, 2004; Shukla, Nespor, & Mehler, 2007).
cohesive units (Morgan & Saffran, 1995). Eight-month-old English infants are also able to use lexical stress in speech segmentation, since trisyllabic stimuli constituted by a sequence of strong-weak-strong syllables are only correctly extracted from the stream as long as the first syllable receives primary stress, but not when primary stress is located at the last syllable (Houston, Santelman, & Jusczyk, 2004).

Twelve-month-olds French listeners (but not 8-month-olds) also use their native language’s prosodic units in segmentation (i.e., the syllable; Nazzi, Iakimova, Bertoncini, Frédonie, & Alcantara, 2006). Furthermore, 13-month-olds are also sensitive to prosodic boundaries, such as the ones of phonological phrases (Cristophe, Gout, Peperkamp, & Morgan, 2003).

In fact, prosody is a general term associated with many kinds of suprasegmental information, such as rhythm, intonation and word primary stress (for a review see Cutler, Dahan, & van Donselaar, 1997), which do not necessarily have the same role in speech segmentation. For example, the role of IPs’ contour seems to be universal, probably based on physiological mechanisms such as breath groups (Grice, 2006; Shukla, Nespor, & Mehler, 2007; Werner & Keller, 1994). On the one hand, adult listeners are sensitive to those prosodic contours even in an unknown/foreign language (Shukla et al., 2007; Experiments 5 and 6). On the other hand, these prosodic correlates, such as the ones of prosodic right-edges (which characterize IPs in many different languages; Werner & Keller, 1994), are acoustic marks of the slowing down of the articulators within a breath group, which is reflected in the signal as final lengthening and low pitch (Grice, 2006, see also Cutler et al., 1997). In contrast to universal prosodic cues such as IPs, lexical (or word
primary) stress, although generally correlated with duration, F0, and amplitude (stressed syllables are lengthened, have higher pitch, and are louder), is not acoustically marked in the same manner in all languages. For example, while in Finish pitch seems to be the most important correlate of lexical stress (e.g., Iivonen, Niemi, & Paananen, 1998), in European Portuguese\(^2\) (EP; the native language of participants of the present work) lexical stress is marked by syllable lengthening (Delgado-Martins, 2002; d’Andrade & Lacks, 1996). The location of stress and whether it obeys to a fixed- or free-pattern is also language-dependent (e.g., in Finish, always on the first syllable of a word: Iivonen et al., 1998; in EP lexical stress is by default on the penultimate syllable, although it can occur in any one of the three last syllables of a polysyllabic word, d’Andrade & Laks, 1996; Mateus & d’Andrade, 2000). Thus prosody is a broad category that comprises both universal (e.g., IPs) and language-specific (e.g., word primary stress) cues, which do not necessarily have the same role in speech segmentation.

**Subsegmental Information**

Subsegmental information corresponds to acoustic-phonetic cues (Davis, Marslen-Wilson, & Gaskell, 2002; Mattys, 2004). As a matter of fact, investigations of minimal pairs differing only in word boundary location (such has *play* taught and *plate* ought) identified a variety of acoustic (subsegmental) cues that are associated with word-boundaries (Lehiste, 1960). Allophonic differences in the articulation of segments in different contextual positions are observed in any natural language. For example, in EP, the allophones of [l] in *mel* (in EP, honey) and in *lua* (in EP, moon)

---

\(^2\) Although European Portuguese (EP) and Brazilian Portuguese (BP) share many properties, there are differences particularly as regards prosodic aspects (e.g., Frota & Vigário, 2001). For example, while vowel reduction is a prominent phenomenon in EP, it does not occur in BP (e.g., Abaurre & Galves, 1998). Thus in the present thesis we are exclusively considering EP.
are differently produced according to their position in syllables and words (Mateus & d’Andrade, 2000). Word-initial segments are also longer in duration than equivalent segments that are not word initial (Lehiste, 1960). Durational differences are also observed between the same segments and syllables in short and long words (Klatt, 1980). Therefore, this acoustic-phonetic knowledge could provide listeners with clues about word boundaries location in fluent speech (Liberman & Studdert-Kennedy, 1978). As a matter of fact, two-month-old infants are already sensitive to allophonic differences, such as *nigh rates* and *nitrates* (Hohne & Jusczyk, 1994). However, only at 10.5 month, infants are able to use allophonic information in speech segmentation’s assistance (Jusczyk, Hohne, & Bauman, 1999b).

The degree of coarticulation is also an important subsegmental cue (Fougeron & Keating, 1997) since an early phase in linguistic development (i.e., 8-month-olds; Johnson & Jusczyk, 2001). Coarticulation is usually defined as a change in the acoustic-phonetic content of a speech segment due to anticipation or preservation of adjacent segments (e.g., Kühnert & Nolan, 1999), and is the primary reason for the difficulty in specifying invariant acoustic properties of phonetic segments. Although all fluent speech is coarticulated, the extent of coarticulation between adjacent segments is influenced by the presence of a prosodic boundary. There is, generally, more coarticulation within words than between them (e.g., Byrd & Saltzman, 1998). On the one hand, the degree of overlap is greater for consonantal clusters at onset position than for adjacent consonants that belong to different words (Byrd, 1996). On the other hand, segment strength may convey information about the local coherence vs. disjuncture in connected speech (Fougeron & Keating, 1997; Keating, *in press*).
Segments at the beginning of prosodic domains are generally strengthened, while segments within those domains are not, and this domain-initial strengthening is a general phenomenon found in both stressed and unstressed syllables (Cho & McQueen, 2005).

**Segmental Information**

Another important source of information that can help listeners to locate word boundaries in any natural language is the statistical segmental information conveyed by the speech stream. The transitional probability (TP) between two linguist units (e.g., syllables or phonemes) corresponds to the conditional probability by which one first unit (X) is able to predict the immediately following one (Y):

$$TP(Y|X) = \frac{\text{Frequency}(XY)}{\text{Frequency}(X)}$$

As a matter of fact, within any language, the TP from one unit of sound to the next will generally be highest when the two units follow one another within a word, whereas the TP spanning a word boundary will be relatively low (e.g., Perruchet & Peereman, 2004). This pattern of higher TPs within words than across them is also characteristic of natural infant-directed speech (Swingley, 2005).

Human listeners, in different linguistic stages of development (i.e., infants, children and adults) and even nonhuman primates (Hauser, Newport, & Aslin, 2001) can use TPs between adjacent syllables to locate “word” boundaries within a continuous artificial language (AL) stream composed by nonsense syllables with no acoustic cues to “word” boundaries (e.g., Saffran et al., 1996a; 1996b).
Saffran and colleagues (e.g., Saffran et al., 1996a; 1996b) have implemented the AL learning (ALL) paradigm and proposed that a statistical learning mechanism is responsible for listeners’ ability to discover “words” embedded in an AL.

In this paradigm, in the first, familiarization, phase, listeners are exposed to a continuous stream of nonsense syllables (i.e., the AL; e.g., bupadapatubi, Saffran et al., 1996b), with no other cues to word boundaries than the TPs between (adjacent or nonadjacent) syllables. Within the AL continuous stream, AL-syllables belonging to the same AL “word” have higher TPs (henceforth, TP-words; e.g., bupada, patubi) than AL-syllables that are part of different “words” (i.e., part-word stimuli; e.g., da#patu), straddling a TP-boundary (marked in the previous example by “#”). In the second phase of the ALL paradigm, participants are presented with isolated AL-stimuli: AL-words vs. AL-nonwords³.

With infant listeners, the Headturn Preference Procedure (Jusczyk & Aslin, 1995) enables the evaluation of infants’ sensitivity to the statistical information conveyed by the stream. Indeed, the discrimination between TP-words and part-words is not merely based on the raw frequency of occurrence of the stimuli, since TP sensitivity is observed for 8-month-olds even when part-words occurred as often as TP-words during the familiarization phase (Aslin, Saffran, & Newport, 1998).

Furthermore, reliance on statistical information such as TPs can contribute to real-word acquisition. Since TP computation does not require any (even minimal) lexical knowledge, this conditional probability estimate might allow learners to segment speech from the discovery of troughs in the TPs distribution between

---

³ AL-nonwords are stimuli constituted by the syllables of the AL but with lower TPs than the TP-words, either because they did not occur in the familiarization phase (their TP being 0), or because, although having occurred, and even with the same raw frequency as TP-words (Aslin, Saffran, & Newport, 1998), their TP is lower. These last AL-nonwords are called Part-words, since their syllables occur in the stream as adjacent parts of different TP-words, hence straddling TP-words boundaries.
adjacent syllables. Coherently, infants have been shown to be sensitive to this statistical information from an early phase of linguistic development (Thiessen & Saffran, 2003; for evidences of visual statistical learning in 2-month-olds see Kirkham, Slemmer, & Johnson, 2002). Infants as young as 7-month-olds were shown to be able to segment words according to statistical cues (Saffran et al., 1996a; Thiessen & Saffran, 2003), 8-month-old infants to process the output of statistical learning as candidate words on their native language (Saffran, 2001), and 12-month-olds to segment TP-words and then discover syntactic regularities relating those TP-words (Saffran & Wilson, 2003). These findings suggest that the mental representations produced by the statistical learning mechanism are not just sets of syllables linked by statistics, but new units that could be available as input to subsequent learning processes. Statistical learning could thus constitute one of the earlier ways the perceptual system uses (i.e., infants’ “first window”, cf. Thiessen & Saffran, 2003) to segment a continuous acoustic flow into discrete lexical units.

Furthermore, the statistical learning mechanism acts not only on linguistic material but also on sequences of tones (e.g., Saffran, Johnson, Aslin, & Newport, 1999) and in other modalities (e.g., Conway & Christiansen, 2005). Whether statistical learning reflects one general process (e.g., Saffran, Pollack, Seibel, & Shkolnik, 2007) or multiple specialized processes (e.g., Conway & Christiansen, 2005; 2006) is not yet clear. Nonetheless, it appears to be subject to one unifying principle, namely a tendency toward “reduction of uncertainty” (Gibson, 1991): learners are driven to seek invariant structure in the stimuli array (Gómez, 2002). Therefore, the statistical learning approach postulates that participants are able to compute statistical information, using it not only in language processing but also in
any sequence-learning task (Boyer, Destrebecqz, Cleeremans, 2005; Cleeremans, Destrebecqz, & Boyer, 1998; Perruchet & Pacton, 2006; Saffran et al., 1999).

Infant listeners are also sensitive to statistical segmental information of their native language, i.e., phonotactic probabilities of segments (i.e., the relative frequencies of segments and sequences of segments in syllables and words, cf. Mattys & Jusczyk, 2001; e.g. Jusczyk, Luce, & Charles-Luce, 1994). Indeed, English 9-months-old infants have already begun to determine the way that phonotactic sequences line up with word boundaries in their language (Mattys & Jusczyk, 2001; Mattys, Jusczyk, Luce, & Morgan, 1999) and at twelve-month-old they are already sensitive to the PWC (Johnson, Jusczyk, Norris, & Cutler, 2003).

1.1.3. Issues and issues

The studies briefly reviewed above indicate that, towards the end of the first year, infant listeners are already sensitive to a variety of different sublexical segmentation cues. In Jusczyk’s (1999) words:

“this is a fortunate development because none of the sources of information about word-boundaries is sufficient for correctly segment all words from fluent speech” (p. 326).

Indeed, sublexical sources of information have often been considered too unreliable to be useful (Brent & Cartwright, 1996), and potential problematic sources of noise (Christiansen & Allen, 1997). Thus, for language learners, integrating multiple probabilistic signal-derived cues is one solution that can provide reliable evidence about word boundaries (e.g., Christiansen & Curtin, 2005).
Combining lexical-driven mechanisms, in which word knowledge guides the parsing of the input (see previous section 1.1.1.), to any one of the signal-derived cues may also be helpful to word segmentation (e.g., McQueen et al., 1994; Norris et al., 1995). This, however, holds true only when the speech signal can map onto stored representations. Remarkably, that seems to be the case since an early phase in linguistic development. While six-month-olds are already able to use their knowledge of familiar words (i.e., their own name and the word \textit{mommy}) to assist the parsing of adjacent novel words from a continuous speech stream (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005); at 1.5-year-old, word learning is already affected by lexical competition (Swingley & Aslin, 2007). Thus, early in development, the lexical entries seem to be adult-like in a seminal sense (e.g., Coady & Aslin, 2004; Swingley, 2005), with a compromise between lexically-driven and signal-derived informations being established in speech segmentation from an early phase in linguistic development.

Importantly, if the conjunction of both lexically-driven and signal-derived cues has a major role in speech segmentation during language acquisition, this seems to also hold true in the context of a full, mature speech perception system. Recently, it was demonstrated that sublexical sources of information, may not only assist language acquisition (e.g., Jusczyk, 1999; Thiessen & Saffran, 2003), but also modulate lexical activation in adulthood (e.g., Davis et al., 2002; Gow & Gordon, 1995; Salverda, Dahan, & McQueen, 2003; Salverda et al., 2007).

\textit{Gow & Gordon (1995)}

The cross-modal semantic priming study of Gow and Gordon (1995) was one
of the first pieces of evidence demonstrating that sublexical cues play a role in lexical activation, and hence in speech segmentation, for adult listeners. Two utterances segmentally identical (e.g., /selfʃ/) led to different priming effects based on subsegmental, acoustic-phonetic differences: while *selfish* facilitated the lexical decision of an associated target-word, no priming effect was found for that same target-word when the two-word utterance *self fish* was presented as prime. Thus, listeners are able of using acoustic-phonetic cues on the distinction between appropriate and inappropriate segmentations of words containing embeddings.

**Davis, Marslen-Wilson, & Gaskell (2002)**

Using cross-modal repetition priming, Davis et al. (2002; Experiment 2) evaluated the temporal course of lexical activation and demonstrated that acoustic-phonetic information (i.e., syllable duration) is available early on in the processing of embedded word stimuli to assist the perceptual system in distinguishing between short words (e.g., *cap* in *cap tucked*) and longer competitors (e.g., *captain*). Remarkably, in a gating task (Davis et al., 2002; Experiments 1 and 4), listeners’ responses to short words (e.g., *cap*) remained consistent (and correct) well before their acoustic offset when the lexical context was incompatible with longer word competitors (e.g., *cap looking*), contrary to what was observed when the lexical context was compatible with those longer word competitors (e.g., *cap tucked*), suggesting a strong influence of coarticulation on lexical activation and indirectly on speech segmentation.
Salverda, Dahan, & McQueen (2003)

Salverda et al. (2003) using eye-tracking, add to Davis et al. (2002) findings on adult listeners sensitivity to subsegmental information. In fact, when the first syllable of the target stimuli (e.g., hamster) was substituted by a recording of the monosyllabic word (i.e., with a long duration: e.g., ham), participants presented more transitory fixations to the picture representing this monosyllabic candidate than when the first syllable of the target-word was substituted by a record of the first syllable of that disyllabic word.

Davis et al.’s (2002) and Salverda et al.’s (2003) results suggest that speech segmentation is not solely lexically-driven, as suggested by “lexical competition” models (i.e., TRACE: McClelland & Elman, 1986; Shortlist: Norris, 1994; see section 1.1.1. of the present chapter). If that was the case no sensitivity to acoustic-phonetic cues would have been found. Thus, in the context of a full mature perception system, sublexical sources of information also assist the segmentation processing in conjunction with high-level information.

Moreover, these subsegmental cues have also a role in novel learnt words processing, and modulate lexical activation in bilingual lexicon(s).

Shatzman and McQueen (2006), using eye-tracking, demonstrated that, after an object-label learning task during novel items recognition, listeners were sensitive to differences in duration of syllables of mono- and disyllabic stimuli, even when these acoustic-phonetic differences were not available during the novel items’ learning phase.

Fine-grained acoustic-phonetic information also seems to modulate lexical activation in bilingual lexicons. Ju and Luce (2004) demonstrated using eye-tracking...
that Spanish-English bilinguals only fixated an interligual distractor (i.e., picture whose English name was phonologically similar to the Spanish target word: e.g., the picture of a *pliers, alicates* in Spanish, with the target-word being *playa – beach –*) when the auditory stimuli were presented with English voice onset time (*VOT*), but not when target words were presented with Spanish appropriate VOTs.

Adult listeners are also sensitive to prosodic cues (for a review see Cutler, et al., 1997), such as IPs’ edges (e.g., Shukla et al., 2007); phonological phrases boundaries (e.g., Christophe, Gout, Peperkamp, & Morgan, 2003), the metrical unit (e.g., Cutler & Butterfield, 1992; Cutler & Norris, 1988), and lexical stress (e.g., Mattys, 2000, Vroomen, Tuomainen, & de Gelder, 1998; for electrophysiological evidences see Böcker, Bastiaansen, Vroomen, Brunia & de Gelder, 1999; Cunillera; Toro, Sebastián-Gallés, & Rodríguez-Fornells, 2006; Friedrich, Kotz, Friederici, & Alter, 2004). Speech segmentation driven by statistical segmental information was also demonstrated based on both general TPs within ALs (e.g., Saffran et al., 1996a; for electrophysiological evidences see Sanders, Newport, & Neville, 2002) and listeners’ native language statistical information, such as phonotactic legality (e.g., McQueen, 1998; Norris et al., 1997) and diphones transitional probability (e.g., Mattys, White, & Melhorn, 2005). Therefore, a vast bulk of research (described in more detail in the following chapters) has demonstrated that adult listeners from different linguistic backgrounds are sensitive to suprasegmental, subsegmental, and segmental sources of information and that these sublexical cues assist high-level information in speech segmentation.
1.2. Conventional Experimental Paradigms

Most of the studies that demonstrated adult listeners’ sensitivity to sublexical segmentation cues used conventional experimental techniques such as gating, word-spotting, cross-modal (semantic or repetition) priming, and eye-tracking paradigms. Although all experimental tasks have pros and cons (Goldinger, 1999; Kolinsky, 1998), these conventional paradigms present the same broad limitation, since speech segmentation, as well as the role of sublexical sources of information on it, is indirectly inferred according to lexical activation of different word candidates.

1.2.1. Gating

The dynamics of spoken word recognition have often been examined using gating tasks (for an overview see Grosjean, 1996). This paradigm has provided valuable information on the relationship between the acoustic signal (and the available acoustic-phonetic cues) and lexical access (e.g., Davis et al., 2002). The gating technique provides absolute control over the duration of the speech signal presented to participants. In this task, speech is presented in fragments (i.e., gates) of increasing duration and participants are asked to propose the word being presented and to give a confidence rating of their response. For a given word, the isolation point is defined as the duration of the gate at which participants correctly identified the stimulus-word and did not subsequently change their guess. Since listeners are encouraged to focus on accurate responses independently of the time taken to select them, gating provides a working measure of the steady state of the word recognition
system. It also provides measures reflecting the activation and identification of competing lexical hypotheses as acoustic information accumulates.

However, gating is usually considered an off-line task that may include strategic effects, since it appeals to the sequential nature of the stimuli presented and possible bias can underlie listeners’ performance. Indeed, gating responses can be generated after the presentation of each gate, and most likely reflect lexical interpretation given the available sensory input after internal processing has reached an asymptotic stable state (e.g., Dahan & Gaskell, 2007).

1.2.2. Word-spotting

According to McQueen (1996) the word-spotting task appears tailor-made for investigating lexical segmentation. In the word-spotting task, listeners are asked to detect any real word embedded in nonsense (pseudo-continuous) strings, which gives the task some “ecological validity” (McQueen, 1996). Indeed, since subjects are not told which word they are asked to spot, at first sight, this task seems to resemble the problem that listeners face in natural settings, with continuous speech.

Kolinsky (1998) already suggested that tasks in which participant’s attention is not drawn to factors at study, reduce the use of late (i.e., metaphonological) representations involved in a post-recognition stage. However, despite the apparent resemblance between word-spotting task and recognition in connected speech, the slow reaction times (RTs) and high error rates suggest that this is a difficult task for listeners. Furthermore, it is not clear whether metaphonologic representations could be involved in listeners’ performance. In some metaphonological tasks, participants are asked to produce their response by deleting one linguistic unit (e.g., a syllable or
a phoneme) at a particular position of the stimuli presented. In some way, this seems to be similar to what participants are asked to do in word-spotting, particularly in conditions in which the location of the embedded word is given beforehand (whether it can occur in the beginning or in the end of the nonsense container string). Eventually, in order to correctly perform the word-spotting task, listeners are asked to delete one (or more) segment of the nonsense sequence and also evaluate whether this metaphonological operation results in any real word. This explicit, intentional analysis of recognition outputs would be far from the perceptual literacy-independent stage (see Kolinsky, 1998). Yet, this suspicion remains to be tested.

It is important to note, however, that word-spotting has often been used to analyze the role of sublexical information in speech segmentation, such as metrical prosody (e.g., strong syllables, in English: Cutler & Norris, 1988; lexical stress, in Finish: Vroomen et al., 1998), and segmental information, like phonotactics (McQueen, 1998), the PWC (Norris et al., 1997), vowel harmony (Vroomen et al., 1998) and syllable onsets (Dumay, Frauenfelder, & Content, 2002). Crucially, the role of these sublexical segmentation cues demonstrated with word-spotting was also replicated with other paradigms (e.g., cross-modal priming, Mattys et al., 2005; the ALL: Saffran et al., 1996a; Vroomen et al., 1998).

1.2.3. Eye-tracking

The eye movement paradigm (Allopena et al., 1998; for an overview see Tanenhaus & Spivey-Knowlton, 1996) has proved to be a valuable methodology for studying on-line lexical access in spoken-word recognition using continuous speech input. This fairly natural setting enables the exploration of subtle competitor effects
that possibly could not be observed with other conventional techniques.

In the visual world task, participants are instructed to fixate a central cross then followed by a spoken instruction to move (using a computer mouse) one of four objects displayed on a computer screen (e.g., “click on the beaker with the computer mouse”; Alloopena et al., 1998). Eye gaze is monitored as the spoken stimulus is heard and until listeners perform the instruction. Eye movements’ monitoring (i.e., location and latencies of eye fixations for each item presented in the visual display over time) allows the examination of spoken word recognition as the continuous speech stream unfolds, since eye movements occur concurrently with the spoken input in real time. Thus, fixations initiated at time $t$ reflect the spoken word processing stage reached at that time, determining both the available input and internal processing dynamics. This measure provides a fine-grained estimate of target and competitor consideration over time, being extremely sensitive to the uptake of information during lexical access. Indeed, Alloopena et al. (1998) demonstrated that the presence of a cohort competitor (e.g., beetle) in the display increased the latency of eye movements to the target (e.g., beaker) and induced participants’ frequent looks to that competitor. Thus, Alloopena et al. (1998) directly demonstrated that two referents with phonologically similar names were in fact competing as the target word unfolds. Dahan, Magnuson, and Tanenhaus (2001) also demonstrated with eye-tracking that frequency affects the earliest moments of lexical access. Indeed, when a referent picture (e.g., picture of a bench) was presented with two cohort members with high- and low-frequency (e.g., bed and bell, respectively), participants fixated more often the high-frequency competitor than the low-frequency one. Thus, frequency effects occur early on as spoken word unfolds, suggesting that the locus of
frequency effect is not late-acting nor has a decision-bias locus. Furthermore, eye-tracking data has been compared with simulations of connectionist models such as TRACE (e.g., Alloopena et al., 1998; Dahan et al., 2001), providing an important tool for theoretical testing.

Magnuson, Tanenhaus, Aslin, and Dahan (2003) have also used eye-tracking, after an object-label learning task, for examining the course of lexical activation and competition on the artificial lexicon in adulthood. Surprisingly, Magnuson et al. (2003) demonstrated signature processing effects of cohort and rhyme competitor modulated by target and neighbor frequency during “word” recognition in the artificial lexicon. Thus, artificial lexicons seem to exhibit the processing effects that characterize spoken word recognition with real words (Alloopena et al., 1998; Dahan et al., 2001). These lexical signature effects for artificial lexical items were observed even immediately after the first session of object-label (i.e., new visual shape - new spoken word) learning.

Recent eye-tracking studies (e.g., Ju & Luce, 2004; Salverda et al., 2003; Salverda et al., in press; Shatzman & McQueen, 2006) also demonstrated participants’ on-line sensitivity to signal derived cues, and their role on lexical competitors’ activation.

Although eye-tracking provides fine-grained information of the temporal course of lexical activation, it suffers from two particular limitations: (i) it only measures activation for (the limited) displayed items; and (ii) only word-targets with clear referents can be evaluated using this paradigm.

In a very recent study of Magnuson, Dixon, Tanenhaus, and Aslin (2007) a new version of the visual world paradigm that overcomes the first limitation pointed
out above was used. In this version the recognition of single monosyllabic words (varying in frequency and competitor’s environment) were displayed as a referent picture among three other unrelated distractors. Thus, competitors were never present in the same display as the target as it is usually done in eye-tracking experiments. In this new version was still possible to infer how the set of activated lexical candidates change as a word is heard, and hence to estimate the time course of lexical activation by measuring eye movements as target-words were presented. In fact, Magnuson et al.’s (2007) data demonstrated that while the competitor set changes dynamically, earlier on as the spoken word unfolds, stronger competition is observed between targets and cohorts (i.e., candidates sharing phonological onsets; cf. Marslen-Wilson, 1987). While in an early moment effects of word frequency and cohort density dominate, global neighborhood effects emerge later in the recognition process, and recognition occurs against a background of activated competitors that changes over time based on both fine-grained goodness-of-fit and competition dynamics.

Notably, even studies in which participants are only asked to look at the display with no explicit instructions to search for particular targets (i.e., a passive condition: e.g., Huetting & McQueen, 2007) demonstrate the same effects observed in conventional conditions in which participants are asked to perform an over task on the visual display presented (e.g., Alloopena et al., 2007).

Recently, some studies (e.g., Huetting & McQueen, 2007; McQueen & Viebahn, 2007) started to overcome the second limitation referred above by using printed words instead of pictures in the visual display. In these studies, the same
phonological effects observed with visual referents (e.g., Allopena et al., 1998) were demonstrated. However, different patterns of results are found with words and pictures (see e.g., Huetting & McQueen, 2007), suggesting that specific aspects may differentiate the processes involved in the treatment of these two types of material. Thus further research is required on this matter before implementing this printed words version of the “visual-world” paradigm.

1.2.4. Cross-modal priming

Both cross modal semantic (for an overview see Tabossi, 1996) and repetition priming allow the assessment of words activation in connected speech, and thus have also been used to examine the interaction of lexical and prelexical informations on speech segmentation (e.g., Gow & Gordon, 1995). In cross modal priming experiments participants usually perform a lexical decision task on visual targets after hearing an auditory prime. The inter-stimulus interval (ISI) between prime and target can also be manipulated. By comparing the RTs of lexical decision in the condition in which prime and target-word are related (e.g., semantic/associative relationship: tulips and flower) with the RTs in the condition in which no relation exists between prime and target (i.e., priming effect), it is possible to evaluate to which degree different lexical items are activated.

In the semantic version, in critical trials, the visual target-words are semantically and/or associatively related either to the prime (e.g., flower and tulips; Gow & Gordon, 1995; Experiment 1) or to part of the prime (e.g., kiss and warm lips; Gow & Gordon, 1995; Experiment 1). In the repetition version, in critical trials, target-words can either correspond to the prime (in the repetition priming task: Davis
et al., 2002; Experiment 2) or share phonological information with part of the prime (e.g., *diplenota* and *notable*; Mattys, 2004). In the repetition version of cross-modal priming, the possible confounds produced by differences in semantic or associative relatedness are avoided, while retaining sensitivity to effects of lexical competition and mismatch. As a matter of fact, Norris, Cutler, McQueen and Butterfield (2006) have proposed that associative priming is not a direct consequence of automatic speech processing, since the conceptual interpretation of the speech input occurs only in a subsequent phase separated from the activation of phonological forms of the current lexical candidates.

Importantly, the fragment version of cross-modal priming has been used to evaluate the weighting of different segmentation cues, incongruently available, and hence competing cues, in the speech input (i.e., the prime stimuli; Mattys, 2004; Mattys et al., 2005).

However, one possible limitation of the cross-modal priming paradigm regards the fact that the non-observation of priming effects does not ensure that a lexical competitor that (partially or totally) matches with the input is not at least weakly activated. It is possible that a competitor might be weakly activated but not sufficiently so to be detected in this task, leading to no significant priming effects.

As already suggested by Kolinsky (1998) the attribution of tasks to specific processing stages is not straightforward, and can easily become an *oversimplistic* approach. However, in all the conventional experimental tasks aforementioned, the role of sublexical cues in speech segmentation is indirectly inferred according to lexical activation of bottom-up matching lexical candidates. Thus, with those tasks it
is not possible to disentangle the contribution of sublexical sources of information \textit{per se} from the role of high-level information in speech segmentation, since lexical activation and competition are also intrinsically involved on listeners’ performance. Notably, this shortcoming is not present on the ALL paradigm implemented by Saffran and colleagues (Saffran et al., 1996a, 1996b) and hence this technique could be a valuable tool for studying the role of sublexical information in speech segmentation, independently of the impact of high-level information in speech segmentation. This is particularly important in the context of a full mature speech perception system, since in this case high-level information \textit{per se} has a major role in speech segmentation (e.g., McClelland & Elman, 1986; Norris, 1994). Thus, in order to disentangle the role of sublexical cues from the one of high-level information, it is important to adopt an experimental technique that enables the evaluation of those sublexical sources of information in the absence (or almost so) of available high-level information.

1.3. The Artificial Language Learning Paradigm

In fact, while in natural languages a rich set of correlated cues is available, which makes the isolation of each cues’ role in speech segmentation extremely difficult, ALs allow the systematic manipulation of particular sublexical sources of information, enabling the evaluation and preponderance of isolated cues over others. Additionally, in AL settings, the role of sublexical segmentation cues is not inferred according to high-level activation, as happens in conventional designs using the experimental techniques briefly reviewed in section 1.2. of the present chapter.
In the ALL paradigm, participants are presented with an unknown language (i.e., the AL), and thus high-level information is absent (or at least largely reduced) and hence cannot be used in this speech segmentation processing. The rationale of this paradigm (Saffran et al., 1996a; 1996b) conceives that if listeners show learning effects (e.g., in a forced-choice task) after an AL familiarization period, it is because the information available in the AL corresponds to the one available in any natural language and used in speech segmentation. During the familiarization phase of an ALL study, as already referred in the previous section 1.1.2., listeners are exposed to a continuous AL stream constituted by a limited repertoire of concatenated syllables. Usually statistical information is the only available cue that can help listeners to locate word-boundaries, since the TP between adjacent syllables is higher within AL-words (i.e., TP-words) than between them. Thus, any preference for those TP-words over other AL-stimuli with lower TP (i.e., nonwords or part-words) in a subsequent ALL-test (e.g., two alternative forced-choice test) is a demonstration that listeners can parse TP-words from the continuous speech stream through the adoption of a statistical segmentation procedure. This test provides a measure of learning, and consequently allows evaluation of the reliability of the sublexical segmentation cues available in the speech stream, independently of high-level (lexical and post-lexical) information.

Additionally, the ALL paradigm also provides the investigation of speech segmentation with a highly controlled situation, reducing the set of uncontrolled variables. This paradigm also permits the systematic manipulation of different sublexical cues simultaneously available in the AL stream, in conditions of both congruent (i.e., different available cues suggesting the same AL parsing) and
incongruent (i.e., different available cues suggesting incompatible and hence competing segmentation hypotheses) cues.

1.3.1. ALL and Implicit Learning

The resemblance between ALL and the Artificial Grammar Learning of Reber (1967) is obvious. In the 1960s, Arthur Reber asked people to look at strings of letters generated by a finite state grammar (which was artificial, since it did not correspond to any natural language structure). After that first phase, participants were informed that the order of letters obeyed to a complex set of rules and in the second phase they were presented with new sets of letter sequences, half of which obeyed the rules and half of which did not, and were asked to classify those sequences according to their artificial grammar legality. Participants could make these decisions with better-than-chance accuracy with a correct performance on about 60-70% of the trials, even though they found very difficult to report knowledge of the rules of the grammar. For example, participants could not recall correctly which letters began and ended the strings. Reber (1967) described these results as a:

“peculiar combination of highly efficient behavior with complex stimuli and almost complete lack of verbalizable knowledge about them” (p. 859).

Reber (1967) realized that such Artificial Grammar Learning could be used as a method for investigating what he termed Implicit Learning, i.e. the acquisition of implicit (unconscious) knowledge.

Statistical TP-based segmentation actually shares many features with other sequence learning effects studied within the framework of Implicit Learning (IL,
Perruchet & Pacton, 2006), in which on-line sensitivity to statistical information like TPs was already observed (Hunt & Aslin, 2001; Remillard & Clark, 2001). Additionally, the observation of an age-independent effect in ALL performance of 6/7-years-old children and young adults (Saffran, Newport, Aslin, Tunick, & Barrueco, 1997), as well as the same age-independent effect in IL conditions (Meulemans, Van der Linden, & Perruchet, 1998; Vinter & Perruchet, 2000) could suggest that a general statistical learning mechanism operates efficiently in different domains since an early phase and independent of cognitive development.

However, there is a fundamental difference in IL and Statistical TP-based segmentation studies as regards the type of tasks in which learning effects were observed. In IL (e.g., with Serial Reaction Time task, for an overview see Cleeremans et al., 1998), participants’ performance is directly assessed by RTs to probable and improbable sequences during the exposition to the statistical sequential material, and by participants’ performance in a subsequent transfer task in which previous statistical sequential knowledge does not apply to the sequence, and thus interference of this statistical disruption is also evaluated. These on-line measures reveal participants sensitivity to statistical contingencies. In contrast, in statistical speech segmentation, during the exposition to the statistical material (i.e., the AL stream), participants do not perform any overt task, and their sensitivity to statistical information is evaluated in a subsequent indirect off-line two-alternative forced-choice task. Even without endorsing the assumption that all knowledge expressed through a forced-choice measure is conscious; the effects revealed in such different tasks do not necessarily rely on the same aspects of learning. Remarkably, recently Turk-Browne, Jungé, & Scholl (2005) have demonstrated in a visual statistical
learning study, formally similar to the auditory ALL study of Saffran et al. (1996a), that the performance on a forced-choice task revealed the same visual ALL pattern than the one observed in a RT task (i.e., visual detection of a target shape in a predictable or unpredictable position according to statistical information).

Nevertheless, the relationship between the statistical, TP-based, learning mechanism and IL is far from clear. Although many authors (e.g., Boyer et al., 2005; Perruchet & Pacton, 2006; Perruchet & Vinter, 1998; Saffran et al., 1999) suggest that the underlying mechanism of learning effects observed in both fields can be closely related, the ALL studies within speech segmentation field have been mainly used in order to evaluate both the impact of statistical learning in speech segmentation (e.g., Saffran et al., 1996b) and the role of other sublexical sources of information in this processing (e.g., Saffran et al., 1996b; Vroomen et al., 1998).

1.3.2. ALL and Speech Segmentation

Neuroimaging evidences, such as electrophysiological correlates (N100 amplitude enhancement: Sanders et al., 2002; see also Cunillera et al., 2006) and neural signatures (left-lateralized fMRI signal increases in temporal cortices: McNealy, Mazziotta, & Dapretto, 2006) of on-line word segmentation in ALL suggest that processes that may be central to speech segmentation are also called upon in the segmentation of an AL. In addition, behavioral data also demonstrates that adult listeners attempt to integrate the output of statistical learning with knowledge of their native language acquired prior to the experimental task, (e.g., prosodic properties: Shukla, et al., 2007; Vroomen et al., 1998; phonotactic probabilities: Onnis, Monaghan, Richmond, & Chater, 2005). Indeed, Finish listeners
had better ALL performances when both primary stress and vowel harmony (sublexical informations that characterize Finish) were available in the AL stream than when only statistical information was available to help listeners locating AL-word-boundaries (Vroomen et al., 1998). Dutch listeners had also better ALL performances when besides statistical information, lexical stress was congruently available in the stream (i.e., initial syllable of TP-words acoustically marked with a F0 peak; Vroomen et al., 1998). The same ALL benefit promoted by the availability of different congruently available sublexical cues was also found for French listeners, when prosodic information was also available in the stream (i.e., final syllable of the TP-words acoustically marked by lengthening and/or a F0 peak; Bagou, Fougeron, & Frauenfelder, 2002) and for American-English listeners (i.e., vowel lengthening on the final syllable of TP-words; Saffran et al., 1996b; Experiment 2).

Until recently, within the context of a full mature speech perception system, sublexical sources of information have often been investigated individually and thus the question of how different sublexical segmentation cues interact with each other has been left largely unspecified. As a matter of fact, the integration of multiple available segmentation cues may provide evidence about linguistic aspects that cannot be derived from any single source, promoting the optimization of speech segmentation (see Christiansen et al., 1998; Christiansen & Curtin, 2005). Therefore, the realistic understanding of natural speech segmentation and how cues are weighted on its assistance can only be achieved if the role of different cues, simultaneously available, is evaluated both in congruent (i.e., when those cues
suggest the same segmentation hypotheses) and in incongruent (i.e., suggesting incompatible word-boundaries) conditions. Shedding light on the integration and role of multiple available sublexical segmentation cues in the context of a full mature speech perception system is the general aim of the work presented in this thesis.

1.4. The hierarchical organization frame of Mattys and colleagues

Recently, Mattys and colleagues (Mattys, 2004; Mattys et al., 2005) proposed the first integrated theoretical approach of the hierarchical organization of both sublexical and lexical sources of information in speech segmentation. In this theoretical framework, the involvement of any segmentation cue, either lexically-driven or signal-derived, is conceived as a graded rather than an all-or-none phenomenon, and speech segmentation is largely the product of both the differential weighting of the types of information available in the signal and of the listening conditions (Mattys, 2004; Mattys et al., 2005).

The hierarchical segmentation proposal of Mattys and colleagues (2005) defines three tiers according to the type and importance of the information available in the speech stream (for an illustration of the theory see Figure 1). The first tier consists of lexical and post-lexical knowledge, which, is supposed to be the most reliable information in optimal listening conditions. Thus, when these cues are available with phonetically intact signal, speech segmentation will be lexically driven, even if conflicting sublexical information is available in the speech flow. The

---

4 On may argue that the term “hierarchical organization” is misleading, since there is no structural relation between the sources of information of the different tiers: they are just weighted differently, with a bottom-up increase of reliability of the segmentation cues in optimal listening conditions. However, for sake of clarity, we kept this terminology when referring to Mattys et al.’s (2005) model.
other types of information – subsegmental, segmental and suprasegmental – early acquired, are confined to the below tiers and are called upon when lexical information is unavailable or reduced. The second tier consists in the conjunction of segmental and subsegmental information, which are able to drive segmentation processing when only sublexical information is available in the stream and the speech signal is phonetically intact. The lowest tier corresponds to metrical prosody, which would act as a last-resource segmentation heuristic (see also Creel, Tanenhaus, & Aslin, 2006; Valian & Levitt, 1996), prevailing over the available information from above tiers when lexical knowledge (tier I) is not available or it is reduced, and the signal is degraded, which impoverishes segmental and subsegmental information (tier II).

Figure 1: Hierarchical organization of speech segmentation cues according to listening conditions, as proposed by Mattys and colleagues. The triangle represents the weighting of segmentation cues as well as their resilience to physical degradation of the signal (i.e., the base of the triangle corresponds to the strongest sensitivity to physical noise)

<table>
<thead>
<tr>
<th>Sources of Information</th>
<th>Listening Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical Knowledge</td>
<td>Optimal</td>
</tr>
<tr>
<td>Segmental &amp; Subsegmental</td>
<td>Poor lexical information</td>
</tr>
<tr>
<td>Metric Prosody</td>
<td>Poor segmental information (physically degraded signal)</td>
</tr>
</tbody>
</table>

Adapted from Mattys et al. (2005)

1.4.1. Experimental evidences

Using the lexical decision task with cross-modal fragment-priming, Mattys
(2004; Mattys et al., 2005, Experiment 1A) demonstrated that the relative impact of (subsegmental) coarticulation and (suprasegmental) lexical stress information is modulated by signal quality. When these two cue types were put against each other, while with phonetically intact speech, the possible word boundaries indicated by coarticulation outweighed the ones indicated by lexical stress, when the signal was physically degraded by white-noise superimposition, the hierarchical organization of these cues was reversed (see Figure 1 for the illustration): listeners adopted a stress-based segmentation procedure, with the possible word boundaries indicated by stress outweighing the ones indicated by coarticulation. This metrical segmentation bias (cf. Mattys, 2004), dependent on physical degradation of the signal, was also observed when the incongruent information available was another sublexical cue represented at Tier II (i.e., phonotactics) and even when the incongruent cue available corresponded to high-level information, represented in Mattys et al.’s (2005) model at Tier I (i.e., lexical information and semantic-context: Mattys et al., 2005; Experiments 5 and 6A, respectively). Additionally, suprasegmental information was able to drive speech segmentation processing in strongly degraded conditions (i.e., white-noise superimposed at a signal to noise ratio – SNR – of -5dB) when multiple cues represented at each one of the other two above tiers were also available in the stream suggesting conflicting segmentation points (Mattys et al., 2005; Experiment 6B).

1.4.2. Developmental and Cross-linguistic implications

Part of the differential weighting, and hence of the hierarchical organization according to the resilience of cues to noise, might be related to their structural grain
(Mattys et al., 2005). For example, the use of cues defined at a lower structural level, such as coarticulatory cues, depends on the ability to process fine-grained, low-level acoustic properties, which are more easily masked by noise than higher-level units. Indeed, the perceptual salience of coarticulation is much more affected by the presence of noise than other cue types (Mattys, 2004; Mattys et al., 2005). When only sublexical information is available in the stream, segmentation driven by coarticulation is only observed with phonetically intact speech (Mattys, 2004), and when information of all three tiers is available (high-level and sublexical information) coarticulation has only an effect in mild-noise conditions (i.e., 5dB SNR; Mattys et al., 2005, Experiment 6B).

The representation of segmental and subsegmental information at the same middle tier by Mattys et al. (2005) is based on the fact that in any natural language segmental and subsegmental information tend to be intrinsically correlated, and hence they propose the same segmentation hypotheses. However, Mattys and colleagues also suggest that even cues represented at the same tier can be differentially weighted in particular languages, as long as the word-boundary predictability of those cues is not highly correlated in that language.

Cross-linguistic differences would be particularly observed on the weighting of sources of information that are language-specific cues, i.e., specifically related with the particular language, such as lexical stress (e.g., in Finish always at the first syllable of a word, while in EP by default in the penultimate syllable). For example, while in Finish a lexical stress segmentation strategy could be a very reliable cue even with intact signal, in EP this is unlikely, since lexical location is neither fixed nor indicating a highly probable word-boundary.
The hierarchical organization framework of Mattys and colleagues (Mattys, 2004; Mattys et al., 2005) is also underlain by a developmental conception of cues weighting, according to their role in different stages of linguistic development. Sublexical cues represented at the lowest tiers have a central role in development as guides to other segmentation cues, and hence they have a subordinated role, being lower-weighted, in the context of a full mature speech perception system. However, being a fundamental mechanism for language acquisition, though lower weighted in adulthood, confers it relative immunity to signal degradation. During linguistic development, these fundamental lowest-weighted sources of information were supplanted by other more reliable (and thus with high word-boundary predictability) and latter acquired segmentation cues.

For example, the ability to track TPs appears to be precocious (e.g., Kirkham et al., 2002; Thiessen & Saffran, 2003), rapid (e.g., Saffran et al., 1996a), and with a possible phylogenetic basis (Hauser et al., 2001), enabling statistical information to act as a pivot mechanism (cf. Mattys et al., 2005) in the acquisition of not only words but also other word-boundary cues (Thiessen & Saffran, 2003). When TPs and lexical stress (i.e., language-specific cue) are both available in the stream suggesting different segmentation points, English 6.5-months-old infants consider TPs the more reliable cue (Thiessen & Saffran, 2003) whereas at 9-month-old, the weighting of these cues is reversed, with lexical stress being the more reliable one (Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003). Furthermore, by 10.5 months, English infants are already able to integrate multiple sources of information, and lexical stress starts to lose its previous importance (Jusczyk et al., 1999a), remaining a last-resource segmentation heuristic in adulthood (Mattys et al., 2005).
Thus, the more resilient (but lower-weighted in normal listening conditions) cues in adult speech segmentation could correspond to the earliest acquired ones, which would also be the most critical cues at the onset of language development.

Nevertheless, since Mattys’ (Mattys, 2004; Mattys et al., 2005) framework is quite recent, two fundamental aspects, studied in the work presented in this thesis, remain underspecified. First, it is not clear how different types of sublexical information, both represented at the same tier (e.g., segmental and subsegmental information) and at different tiers (e.g., segmental and suprasegmental information), interact with each other. Second, Mattys et al. (2005) already suggested that listening conditions go beyond physical quality of the signal. However, until the moment, all the studies addressing this aspect have only evaluated impoverished conditions by physical degradation of the signal.

1.5. An overview of the present thesis

In the work described in this thesis we mainly adopted the ALL paradigm to evaluate two central aspects of the role of sublexical sources of information in speech segmentation on the grounds of Mattys et al.’s (2005) theoretical proposal. In particular, we evaluated the weighting of different sublexical cues and the role of (qualitatively) different listening conditions in speech segmentation in the context of a full, mature speech perception system.

The ALL paradigm allows the simultaneous evaluation of the role of different sublexical sources of information in speech segmentation. Using this paradigm, it is
possible to manipulate systematically those cues in congruent and in incongruent conditions. In the congruent cues condition, the available cues would suggest the same segmentation outputs. This would enable the evaluation of whether the congruency of different available cues promotes an optimization of speech segmentation processing in comparison to a condition in which only one segmentation cue is available in the speech stream. In the incongruent cues condition, when different cues are available in the stream suggesting incongruent (and hence competing) segmentation hypotheses, it is possible to access which cue prevails over the other. In this *incongruent cues* condition, in the second phase of the ALL paradigm (i.e., forced-choice testing), participants are presented with the segmentation outputs of the cues previously available in the AL stream, and are asked to decide which units are more plausible “words” of the new language. Note that in this incongruent cues condition the available cues suggest different segmentation hypothesis. Thus, based on participants’ preference for the outputs of one segmentation procedure over another, we can evaluate the weighting of the available sources of information (i.e., which sublexical cues listeners consider more reliable). Since this incongruent cues condition can be presented in different listening conditions, we are also able to evaluate whether the weighting of the available cues (i.e., which cue prevails over the other) is modulated by the listening conditions as suggested by Mattys and colleagues (Mattys et al., 2005).

1.5.1. The weighting of different sublexical sources of information in speech segmentation

Mattys et al. (2005) showed that when coarticulation (subsegmental) and
phonotactic (segmental) information (both at tier II) were available, indicating the same segmentation points, their segmentation hypotheses were the ones used by adult listeners, in two scenarios: (i) when lexical information was not available in the signal; (ii) when signal quality was sufficiently degraded as to reduce reliance on the semantic context, but, at the same time, allowing a relative availability of the acoustic-phonetic aspects. But, since in Mattys et al.’s study phonotactic and coarticulatory information were never put in conflict, we do not know whether one (and which one) of these two information types prevailed.

In fact, Mattys and colleagues (Mattys, 2004; Mattys et al., 2005) posit that segmental and subsegmental information are represented at the same tier, and consequently their model confers them the same importance. Nevertheless, it is plausible that even sublexical types of information represented in the same tier have independent impacts on segmentation and differ by their degree of (in)dependence on signal quality. Indeed, as already posited by Mattys and colleagues, and mentioned in section 1.4.2. of the present chapter, at least in some languages, segmental and subsegmental information may be weighted differently according to their word-boundaries predictability. This proposal was investigated in Experiment 1 described in chapter 2 of the present thesis, as regards coarticulation (a subsegmental, acoustic-dependent, cue) and transitional probabilities, a segmental statistical cue related to a higher structural level (here, syllabic, i.e., the TP between adjacent syllables). In an ALL setting, we investigated, through three signal-quality conditions, the relative impact of these two types of sublexical information in speech segmentation: segmental statistical information – TPs between adjacent syllables, and subsegmental information – coarticulation.
In addition, exclusively in physically degraded conditions for adult listeners, suprasegmental information seems to have a preponderant role over other sublexical sources of information such as statistical segmental cues (Mattys et al., 2005; Experiment 2). However these two types of information were only evaluated regarding word primary stress and phonotactics, and only with English listeners (Mattys et al., 2005). Indeed, Mattys et al. (2005) already suggested that the weighting of language-specific cues, such as lexical stress, probably depends on their word-boundary predictability in a specific language (cf. Mattys et al., 2005; see section 1.4.2. of the present chapter). Therefore, cross-linguistic differences would be particularly observed on the weighting of these language-specific cues. As a matter of fact, the majority of content words in stress-time languages like English have a metrically stressed syllable at word onset. Yet, that is far from true in EP, in which by default lexical stress occurs in the penultimate syllable of polysyllabic words. Thus, in EP, lexical stress does not generally indicate any word-boundary. Furthermore, as already mentioned above, prosody is a general term associated with many kinds of suprasegmental information, such as rhythm, intonation and word primary stress and no study until the present has evaluated whether these different types of suprasegmental information could act all as last-resource segmentation heuristics, only prevailing over (statistical) segmental information in physically impoverished conditions. In Experiments 2 and 3 presented in chapter 2 of the present thesis, we used two ALL settings to investigate the impact of physical noise on the weighting of prosodic information and statistical segmental information. In Experiment 2, we investigated the weighting of TPs computation and IPs right-edge
(i.e., acoustically marked by syllable lengthening and falling pitch). In order to evaluate whether the role of prosody in speech segmentation would depend on the particular type of suprasegmental information available in the speech stream, in Experiment 3B the suprasegmental information at study corresponded to a language-specific prosodic cue: word primary stress.

1.5.2. The listening conditions

One of the innovations and fundamental aspects of Mattys et al.’s (2005; Mattys, 2004) proposal regards the importance of listening conditions in the weighting of various speech segmentation cues. Until now this impact was only evaluated through physical degradation of the signal (i.e., physical noise), even though Mattys et al. (2005) already suggested that listening conditions go beyond this physical quality. Under this view, impoverishing listening conditions while maintaining intact the physical quality of the signal would also affect the kind of cues predominantly used in segmenting speech. A situation of cognitive noise could be achieved by reducing the available attentional resources necessary for extracting AL-words from the continuous stream.

In Experiment 4 presented in Chapter 4 of the present thesis, we used an incidental ALL setting to investigate the impact of cognitive noise (i.e., attentional load) on the weighting of two types of sublexical cues to speech segmentation: statistical information – TPs between adjacent syllables – and subsegmental information – coarticulation. Using the same sublexical sources of information studied in Experiment 1 (see chapter 2), we were also able to evaluate whether the type of noise (physical in Experiment 1, cognitive in Experiment 3) would have a
differential impact on the hierarchical weighting of those cues, represented at the same middle tier in Mattys et al. (2005) proposal. Furthermore, we also evaluated whether attention-load (i.e., \textit{cognitive noise}) has a graded rather than an all-or-none impact on speech segmentation, as already suggested for physical noise by Mattys (2004; Mattys et al., 2005).

Until now, no study has ever evaluated the impact of \textit{cognitive noise} on speech segmentation driven by coarticulatory information. However, for statistical information, two studies (i.e., Saffran et al., 1997; Toro, Sinett, & Soto-Faraco, 2005) evaluated the relationship between statistical learning (TP-based speech segmentation) and attentional resources.

\textbf{Statistical Learning and Attention}

The idea of a statistical segmentation procedure that is sensitive to temporal contingencies of the speech stream, operating largely independently of awareness and control (and hence in an automatic fashion) is attractive. This mechanism would allow the effortless acquisition of such a powerful cognitive ability as language, without the need to call on conscious (and controlled) strategies or processes. However, this does not imply that word segmentation based on TPs occurs automatically, regardless of the available attentional resources.

Several studies suggest that statistically-driven segmentation can occur in the absence of attentional allocation to the AL stream. Human infants (Saffran et al., 1996a) and non-human primates (Hauser et al., 2001), who cannot be formally instructed to attend to the speech stream, are sensitive to the available TPs. In
addition, human adults are able to use the statistical information conveyed by the speech stream even in incidental passive listening (Toro et al., 2005) and when they have to perform a concurrent drawing task (Saffran et al., 1997).

In an incidental ALL condition, in which participants were not informed about the subsequent performance of a task about their knowledge of the AL words, Saffran et al. (1997) reported that statistical speech segmentation was not affected by participants’ performance of an independent, primary visual color-drawing task, with the AL being presented as “background” nonsense auditory stream (Saffran et al., 1997). Since participants were told neither that the stream consisted of an AL, nor that they would be tested later in any way, the observation of an ALL effect in these conditions has been interpreted as evidence that statistical learning proceeds incidentally (Saffran et al., 1997) “as a by-product of mere exposure” (Saffran et al., 1999). Yet, incidental does not mean automatic; even if statistical learning occurs when it is not part of task requirement, the specific learning conditions on Saffran et al. (1997) do not permit the conclusion that this mechanism is automatic. Indeed, simply instructing participants to focus attention on another task is not sufficient to prevent attention-switching to the simultaneously presented speech stream. In situations of low perceptual load (cf. Lavie, 1995), such as the one used by Saffran et al. (1997), any attentional capacity not taken up by the processing of task-relevant stimuli (the color-drawing task) might involuntarily “spill over” to the perception of task-irrelevant information (e.g., Lavie, 1995, 2005). To ensure that attention is actually diverted from the AL stream, it is necessary to use another concurrent task of high processing load that engages all the available attentional resources (cf. Lavie, 2005).
Using such a situation, Toro et al. (2005) found a dramatic impairment in ALL effects. Indeed, participants were able to use TPs in mere incidental learning (a Passive-Listening condition) and in conditions of moderate attentional load, when their attention was diverted to a rapid visual serial presentation (RVSP) task in a relatively easy condition, namely with a stimulus onset asynchrony (SOA) of 750 ms (Toro et al., 2005; Experiment 2). Yet, when the primary diverting task had to be performed on an auditory stream (a separate one or the same as the AL stream; Toro et al., Experiments 1 and 3), or when the RVSP used a shorter SOA (500 ms; Toro et al., 2005; Experiment 2), ALL performance dropped to chance level (see also Turk-Browne et al., 2005, for a visual statistical learning situation). Thus, statistical speech segmentation seems to be affected by a severe reduction of the available attention. Toro et al.’s findings suggest that statistical learning, a fundamental mechanism available since an early phase of linguistic development (Thiessen & Saffran, 2003) is an attentional-dependent mechanism.

Since the beginning of the cognitive revolution the distinction between automatic and controlled processes casts largely on their reliance on attention. The traditional view (e.g., Posner & Snyder, 1975; Shiffrin & Schneider, 1977) conceived automaticity as an all-or-none concept. Thus, automatic processes (contrary to controlled ones) were by definition not dependent on attention. Yet, the distinction between automatic and controlled process is far more complex than simply relying or not on attentional mechanisms. Indeed, Stroop interference, generally thought to be unintentional and uncontrollable, is not independent of attentional limitations (e.g., Kahneman & Chajzyck, 1983). Thus, automaticity is not an all-or-none construct but
rather a graded one (Cohen, Aston-Jones, & Gilzenrat, 2004; Logan, 1988; Logan, Taylor, & Etherton, 1999; Moors & De Houwer, 2006).

In chapter 5 of the present thesis, in Experiment 5, we evaluated what is the specific impact of cognitive noise (i.e., high attentional load conditions in which the available attentional resources are severely reduced) in speech segmentation driven by TPs computation. Toro et al. (2005) already suggested that a dramatic reduction of the available attentional resources could possibly only slow down statistical learning and not completely prevent it. In Experiment 5 we evaluated at which extent the amount of AL-familiarization would influence ALL in conditions of low- and high-attention load (i.e., mild and strong cognitive noise conditions, respectively, similar to the ones used in Experiment 4 - see Chapter 4 of the present thesis). This was done by presenting different groups of participants with different familiarization time conditions (7-, 14-, and 21-minutes), in situations in which participants were either merely exposed to or had to actively process rapidly presented pictures.

However, attention is a very broad term and reaching a consensus regarding operational definitions of its various components has been one of the thorniest problems faced by cognitive scientists (e.g., Cohen et al., 2004; Moors & De Houwer, 2006). Roughly, two mechanisms of attention can be considered: (i) a selection mechanism that avoids overloading of the information processing system by filtering task-relevant aspects of incoming information; and (ii) a resources mechanism, or “source of energy” (Moors & De Houwer, 2006) of limited capacity that can be flexibly allocated to any stage in the information-processing chain.

In order to disentangle the role of cognitive noise as a reduction of attention
as limited-capacity resources and as a selection mechanism in statistical speech segmentation, in Experiment 6, presented in Chapter 5 of the present thesis, we manipulated orthogonally the available attentional resources (low- and high-attention load conditions, as in Experiment 5) and the focus of participants’ attention, using incidental and intentional learning instructions.

We adopted ALL settings through all the work presented in this thesis, suggesting that the ALL paradigm can indeed be a very fruitful tool in the understanding of how in speech segmentation sublexical sources of information are weighted in the context of a full mature speech perception system. The hierarchical framework of Mattys and colleagues (Mattys, 2004; Mattys et al., 2005) also enabled us to evaluate on theoretical grounds the role of the different segmentation cues (i.e., suprasegmental, segmental and subsegmental sources of information) at study, as well as the impact of different qualitative and quantitative types of listening conditions. However, two fundamentals aspects at the core of the ALL paradigm have been left unspecified. In particular: (i) the direct demonstration of on-line statistical speech segmentation; and (ii) the nature of statistical segmentation outputs. These aspects were assessed in chapter 6 of the present thesis by combining the ALL paradigm with conventional experimental techniques used on natural speech processing research.

1.5.3. An on-line evidence of statistical speech segmentation

The ALL effect revealed by listeners’ (above chance) preference for TP-words over AL-nonwords provides only indirect evidence of speech segmentation
processes, since it is based on an off-line measure derived from the performance on a subsequent test, usually the two-alternative forced-choice test. Furthermore, forced-choice testing uses isolated, already segmented, stimuli and thus do not ensure that test stimuli indeed correspond to listeners’ segmentation byproducts, since the AL potential ‘lexical-units’ are given beforehand.

In Experiment 7, presented in Chapter 6 of the present thesis, we checked whether adult listeners were actually able to use statistical, TP-based, information in an on-line segmentation task, in which no a priori segmented “lexical” solution was provided. To this aim, first listeners were familiarized with an AL with only statistical segmentation cue (TPs) available, and in the second phase, they performed a variant of word-spotting (Cutler & Norris, 1988), in which their task was to detect any AL-word embedded in four-syllabic nonsense strings.

However, even if, in line with IL findings (e.g., Hunt & Aslin, 2001; Remillard & Clarck, 2001), in Experiment 7, listeners would be able to use statistical information on-line for extracting the AL-words from the nonsense embedding sequences, we still do not know what is the status of those statistical outputs as regards adult listeners’ linguistic knowledge.

In the context of language acquisition, two studies assessed the extension to which statistical learning generates novel word-like units. Saffran (2001) has demonstrated that, immediately after an AL-familiarization period, in which statistical information was solely available, eight-month-olds demonstrated a preference for TP-words over part-words only when those AL-stimuli were embedded in linguistic familiar frames, but no preference was found when AL-stimuli were embedded either in nonsense linguistic frames or in nonlinguistic ones
(e.g., sequences of tones). Additionally, Estes, Evans, Alibali, and Saffran (2007) demonstrated that 17-month-olds were only successful in a subsequent object-label learning task, when labels corresponded to TP-words, but not when these were either nonwords or part-words. Therefore, it seems that word-segmentation and word-learning have a direct link, with statistical segmentation outputs being rapidly available for integration in infants’ linguistic knowledge.

Yet, whether this could also hold true for adults is unknown. Whereas in infants the statistical output could act as “stem cells” for lexical acquisition, since TPs computation does not require any (even minimal) lexical knowledge, acting as the “first window into the regularities of [infants’ native language]” (Thiessen & Saffran, 2003, p. 716), and occupying “a pivotal position in the acquisition of not only words, but also other word boundary cues” (Mattys et al., 2005, p. 493), adults have already acquired multiple speech segmentation cues and a fully consolidated mental lexicon. Therefore, once segmented from a continuous speech stream, adult listeners might continue to treat statistical segmentation outputs as relative coherent sound sequences with high internal probabilities, but with no particular status regarding their native language. Alternatively, TP-units could also be treated as potential “words” available for lexical integration. If this was the case, we would demonstrate that adult listeners are able to learn novel words, parsed from a continuous stream on a TP-basis.

---

5 Here the term ‘word’ loosely refers to phonological representations that are available for subsequent language processing (see Brent & Cartwright, 1996).
1.5.4. The nature of speech segmentation outputs

In Experiments 8 and 9, presented in Chapter 6, we evaluated whether statistical segmentation outputs could be lexicalized, adopting the approach of Gaskell and Dumay (2003a; 2003b) within an ALL setting.

In order to shed light on the mechanisms by which a novel word is integrated in mental lexicon (i.e., the lexicalization process\(^6\)), Gaskell & Dumay (2003a) familiarized adults with lists of isolated novel phonological sequences that strongly overlapped with existing words (e.g., cathedruke derived from cathedral) through a phoneme monitoring task. Since the engagement on lexical competition bears the very nature of a lexical item itself, Gaskell and Dumay (2003a; 2003b) conceived an irrefutable proof of lexicalization: The observation of an inhibition effect on the recognition of an already existing lexical item (e.g., cathedral) that strongly overlapped from onset with those novel ones (e.g., cathedruke) in both lexical decision and pause detection tasks. This inhibitory effect was observed, not only in a gradually fashion over the course of five days (Gaskell & Dumay, 2003a; Experiment 2), but also one week after the familiarization phase, with no intervening exposure period (Gaskell & Dumay, 2003a; Experiment 3). This effect was also observed after a 12-hour interval as long as it included one night’s sleep (Dumay & Gaskell, 2007).

Thus, we already know that the adult lexicon is a dynamic and open entity, at least when the potential novel words are presented already-segmented. Indeed, in natural settings adult listeners sometimes are confronted with new words embedded

---

\(^6\) The process by which a novel memory trace is engaged in lexical competition, affecting the recognition of other existing lexical items (i.e., lexical engagement; cf. Leach & Samuel, in press).
in continuous speech streams. Whether this lexicalization effect would also hold true for the output of statistical segmentation in a continuous speech flow was investigated in Experiments 8 and 9 of the work presented in chapter 6 of this thesis.

In both experiments two types of segmental information were available in the stream during the familiarization phase, namely the AL-statistical information (i.e., the TPs between adjacent AL syllables, which were always higher within than between TP-words) and the wordlikeness of trisyllabic AL-stimuli (i.e., according to both statistical segmental information on listeners native language and by the fact that AL-stimuli strongly overlapped from onset with real words of listeners’ native language – i.e., Base-Words). In the second phase of both Experiments 8 and 9, listeners performed the conventional forced-choice test and a Lexicalization test (i.e., lexical decision task on Base-Words, similar to the one used by Gaskell & Dumay, 2003a; 2003b). For clarifying whether any lexical engagement of the outputs of the segmentation procedure adopted by listeners was transient (observed only immediately after) or if it was stable and long-lasting, listeners also performed the ALL-test and the Lexicalization test post-one week, with no intervening AL-familiarization period.

The main difference between Experiments 8 and 9 was the congruency of those two available segmental cues during the AL familiarization phase.

In Experiment 8, adult listeners were familiarized with one of two ALs constituted by trisyllabic TP-words that differed from the Base-Words in the last syllable, after the Base-Words’ uniqueness point (UP). Thus, the two types of segmental information (i.e., AL-TPs and the wordlikeness of TP-words) suggested the same segmentation outputs in a congruent cues condition.
In Experiment 9, adult listeners were familiarized with two new ALs with the same phonological repertoire used in Experiment 8, and hence with the same two segmental sources of information available in the stream, but in an incongruent cues condition. In this experiment, the AL Part-words (even though these stimuli occurred adjacently in the stream during the familiarization phase, they had lower TPs than TP-words) were the AL-stimuli that differed from Base-Words after the UP. Thus, while AL-statistical information suggested that TP-words were the “units” of the new language, the wordlikeness of AL-stimuli suggested that part-words were the ones.

Thus, we were able to evaluate whether listeners’ speech segmentation outputs would be always lexicalized, even when another incongruent cue was available in the stream, as in Experiment 9, or whether the lexicalization process is highly dependent of the congruency of the segmentation cues available in the speech stream, as in Experiment 8.

In sum, in the work described in the present thesis we evaluated the weighting of different sublexical sources of information in speech segmentation, as regards the simultaneous availability of different types of sublexical cues (represented at the same and at different tiers in Mattys and colleagues theoretical proposal), as well as the role of listening condition in adulthood. In order to ensure that the differential weighting, or in Mattys et al.’s terms the hierarchical organization, of the available segmentation cues was not dependent on the activation of lexical (high-level) information, we mainly adopted the ALL paradigm through all work.

The ALL paradigm has revealed itself ideally suited for studying the weighting of various sublexical cues in the absence (or almost so) of high-level
information, in the context of a full mature speech perception system. Importantly, the present work also directly evaluated the potential usefulness of ALL paradigm combined with other conventional experimental techniques, as well as the status of the sublexical segmentation outputs on a full mature consolidated mental lexicon.

At the final end, in this series of experiments we had always as backdrop the natural settings’ perspective: a rich set of multiple cues available in the stream, assisting speech segmentation. Thus if we aim to obtain a realistic perspective of how these cues are weighted, it is important to study them in conjunction and also in different listening conditions. On the one hand, noise should no longer be confined to the physical quality of the signal. On the other hand, this enables the evaluation of speech segmentation’s optimization and the preponderance of cues over other incongruent ones.
PART II
EMPIRICAL CHAPTERS
Chapter 2.

Statistical information and Coarticulatory cues to word boundaries:

A matter of signal quality.

We investigated how statistical information – transitional probabilities, TPs – interacts with another sublexical cue – coarticulation – to word boundaries, and examined the impact of signal quality on the weighting of these cues, in this Experiment 1. In an artificial language learning setting, with phonetically intact speech, coarticulation overruled TPs, suggesting the prevalence of subsegmental, low-level information. However, while the role of coarticulation in segmentation was highly modulated by signal-quality, TPs were very resilient to noise. When coarticulation was made unreliable by strongly degrading the input (10dB SNR), only statistical information drove segmentation. In a milder degraded (22dB SNR) condition, when some acoustic properties were still available, coarticulation was exploited, although with less reliability than in optimal conditions. These results can be interpreted according to a hierarchical approach (Mattys, White, & Melhorn, 2005) in which both the available segmentation cues and the listening conditions have an important role in speech segmentation.

2.1. Introduction

As already presented in chapter one of this thesis, several signal-derived sources of information were pointed out as potential cues to word boundaries, and combining these cues to lexical-driven mechanisms may be helpful to word

---

7 The experiment presented in this chapter was published in Fernandes, Ventura, and Kolinky (2007).
segmentation (e.g., McQueen et al., 1994; Norris et al., 1995; see section 1.1.2. of chapter 1). However, since each of these signal-derived cues is only probabilistically associated with word boundaries, if we aim to obtain a realistic perspective of how listeners deal with multiple segmentation cues, it is essential to study them not only in isolation, but also in conjunction.

Furthermore, according to Mattys and colleagues (Mattys, 2004; Mattys et al., 2005), the involvement of any segmentation cue, either lexically- or signal-derived, is a graded rather than an all-or-none phenomenon. Mattys and colleagues defined a so-called *hierarchical organization* of both lexically-driven and signal-derived sources of information represented at three tiers, as already described in section 1.4. of the previous chapter. However, since Mattys’ (Mattys, 2004; Mattys et al., 2005) framework is quite recent, some of the possible interactions between the several information types available to the listeners remain underspecified. In particular, it is not clear how the various types of sublexical information interact with each other.

Mattys et al. (2005) showed that when coarticulation (subsegmental) and phonotactic (segmental) information (both at tier II) were available, indicating the same segmentation points, their segmentation hypotheses were the ones used by adult listeners, in two scenarios: (i) when lexical information was not available in the signal; (ii) when signal quality was sufficiently degraded as to reduce reliance on the semantic context, but, at the same time, allowing a relative availability of the acoustic-phonetic aspects. But, since in Mattys et al.’s study phonotactic and coarticulatory information were never put in conflict, we do not know whether one (and which one) of these two information types prevailed.
In addition, Mattys and colleagues (Mattys, 2004; Mattys et al., 2005) posit that segmental and subsegmental information are represented at the same tier, and consequently their model confers them the same importance. Nevertheless, it is plausible that even different sublexical types of information represented in the same tier would have independent impacts on segmentation and differ by their degree of (in)dependence on signal quality. Indeed, as already posited by Mattys and colleagues, at least in some languages, segmental and subsegmental information may be weighted differently according to their word-boundaries predictability. This proposal was investigated in this chapter (Experiment 1) as regards coarticulation (a subsegmental, acoustic-dependent, cue) and TPs, a statistical cue related to a higher structural level (here, syllabic).

2.1.1. Coarticulation – a subsegmental cue

Coarticulation is usually defined as a change in the acoustic-phonetic content of a speech segment due to anticipation or preservation of adjacent segments (e.g., Kühnert & Nolan, 1999; see section 1.1.2. of chapter one). Although all fluent speech is coarticulated, the extent of coarticulation between adjacent segments is influenced by the presence of a prosodic boundary. There is, generally, more coarticulation within words than between them (e.g., Byrd, 1996; Byrd & Saltzman, 1998), and segment strength may convey information about the local coherence vs. disjuncture in connected speech (Fougeron & Keating, 1997; Keating, in press). Indeed, domain-initial strengthening is a general phenomenon, found in both stressed and unstressed syllables (Cho & McQueen, 2005).
The importance of coarticulation in speech segmentation has been demonstrated, in optimal listening conditions, from an early age on (Johnson & Jusczyk, 2001) until adulthood (Mattys, 2004; Mattys et al., 2005). However, it seems to be strongly affected by the superimposition of noise to the speech input (Mattys, 2004). The main reason for this sensitivity to noise is possibly related to the nature of coarticulatory information (Mattys et al., 2005). Indeed, the use of a coarticulatory segmentation cue use depends on the ability to process fine-grained, low-level, acoustic properties. Noise probably masks such information easily, reducing the effectiveness of coarticulation in speech segmentation.

2.1.2. TPs – a segmental cue

Another important source of information that can help locate word boundaries is the statistical information conveyed by the speech stream. The extraction of statistical information is a general mechanism (e.g., Saffran et al., 1999; Conway & Christiansen, 2005) that seems to be universal in nature. As a matter of fact, within any language, the TP from one unit of sound (e.g., a syllable) to the next will generally be the highest when the two units follow one another within a word, whereas the transitional probabilities spanning a word boundary will be relatively low (see section 1.1.2. of chapter one). Human listeners in different linguistic stages of development can use this statistical information to locate “word” boundaries (e.g., Saffran et al., 1996a, 1996b).

---

8 While Saffran and colleagues considered statistical learning as based on statistical computations (Saffran et al., 1996a; Saffran et al., 1996b), another mechanism has been proposed: the formation of chunks (Perruchet & Vinter, 1998). It is difficult to decide between these interpretations on the grounds of their explanatory power (Perruchet & Pacton, 2006), but learning effects based on nonadjacent TPs (e.g., Kuhn & Dienes, 2005; Onnis et al., 2005) seem to challenge the chunking explanation.
Experiment 1. TPs and Coarticulation in physical noise

Until now, the question of how in speech segmentation such statistical cues interact with lower-level cues like those linked to coarticulation has been left unsolved in the context of a completely mature speech system. Indeed, to the best of our knowledge, no study on adult listeners has ever evaluated the relative power of these cues within the same experiment. We already know, however, that with a high quality signal 8-month-old infants consider coarticulation as a more reliable cue than TPs, when these indicate different segmentation hypotheses (Johnson & Jusczyk, 2001). Whether this would hold true in the mature system and whether the observed weighting would be modulated by signal quality, as may be predicted by Mattys and colleagues’ hierarchical framework (Mattys, 2004; Mattys et al., 2005), was examined in the Experiment 1 of the work described in the present chapter of this thesis.

To this aim, we adopted an AL paradigm, since this has been shown to be useful to understand how adult listeners segment natural speech (see section 1.3. of the previous chapter).

Nine groups of participants were presented with AL learning situations (cf. Saffran et al. 1996b). The groups differed in the input-intelligibility they were exposed to, namely intact, mild degraded or strongly degraded speech. Within each input-intelligibility condition, the same AL was presented in three conditions differing by the number and congruence of the segmentation cues available in the speech stream. To achieve realistic coarticulation, we used naturally produced utterances rather than synthesized speech, and created concatenated vs. coarticulated
versions of the AL stimuli. In the single cue condition, only TPs could help the listeners to locate word boundaries, since only the concatenated version of the AL was used. In the other two conditions, coarticulatory information was added. For these coarticulated versions, the TP-words (in the congruent cues condition), namely, the stimuli that correspond to the word boundaries defined by TPs, and the part-words (in the incongruent cues condition), namely, the stimuli that cross the TP-words boundaries, were recorded in their coarticulated versions. Thus, in the congruent cues condition, coarticulation and TPs pointed to the same word boundaries, while in the incongruent cues condition, the two cues pointed to different word boundaries, since coarticulation corresponded to the part-words of the AL. This allowed us not only to explore the relative power of each cue type when in conflict (in the incongruent case), but also to estimate the performance gain afforded by redundant (and hence potentially cooperating) cues (in the congruent case).

If, like infants (Johnson & Jusczyk, 2001), in good listening conditions, adult listeners consider coarticulatory cues as more reliable than statistical cues, in the forced-choice test aimed at estimating how well the listeners learned the “words” of the AL, in comparison to the TP condition one would observe not only a performance gain (i.e., more TP-word choices) in the congruent cues condition, but also a strong decrease of TP-word choices in the incongruent cues condition, since listeners would consider the part-words as being the lexical units of the AL. On the contrary, if statistical information was considered as more reliable, the TP-words would always tended to be considered the “lexical units” of the AL.

Even if the weighting of these cues, with intact speech, reveals itself similar to the one found in infants (Johnson & Jusczyk, 2001), it is possible that it is
modulated by signal-quality. Indeed, in adults, Mattys (2004; Mattys et al., 2005) demonstrated that listeners undervalue coarticulatory cues in conditions of noise superimposition. In the present work, we examined whether statistical information is more resilient to noise than coarticulatory cues by using two levels of input degradation.

The resilience of statistical information to noise may actually be linked to its special role in linguistic development. Indeed, since the computation of TPs does not require any (even minimal) lexical knowledge, it may provide infants with their first window into the acoustic regularities of their native language, thus playing a central role in speech acquisition (Thiessen & Saffran, 2003). This general segmentation mechanism may “occupy a pivotal position in the acquisition of not only words, but also other word boundary cues, such as stress, phonotactics, coarticulation, and allophony” (Mattys et al., 2005, p. 493). One of the qualities that would be expected of such a fundamental mechanism was to be relatively immune to signal degradation.

However, at a high level of noise like the 10dB SNR⁹ condition used here, the fine-grained acoustical information linked to coarticulation can be almost unavailable. Observing that TPs have more impact than coarticulation on the word segmentation process would thus not be surprising. But in milder degraded conditions like the 22dB SNR¹⁰ condition that we also used in the present experiment, the subsegmental information is still audible, and coarticulation could still facilitate segmentation (Mattys et al., 2005; Experiment 6b). Therefore, we were interested in checking whether in the last condition the mere presence of noise

---

⁹ The SNR ratio is measured as the noise intensity against the average signal intensity of the speech, here, the AL signal. Thus, a 10dB SNR means that if the signal intensity is 76dB, as it was the case here, then the noise intensity was set at 66 dB.

¹⁰ The 22dB SNR was chosen because this ratio reduced intelligibility by approximately 50%, a value similar to the one used by Mattys (2004) (for details, see pretest in section 2.2.2. of the present chapter).
already alters the reliability that listeners attribute to low-level subsegmental cues, like those linked to coarticulation. If this were the case, noise superimposition may allow statistical cues to override coarticulatory cues even at such a mild level of noise.

2.2. Method

2.2.1. Participants

Eighty-seven undergraduate psychology students at the University of Lisbon participated in the experiment for course credits. All were monolingual EP speakers, with no reported history of speech or hearing disorders. They were randomly assigned to nine groups according to the 3 (input intelligibility) x 3 (available segmentation cues) experimental design. Among them, 32 were randomly assigned to the intact speech condition (9 in the single cue, 11 in the congruent cues, and 12 in the incongruent cues condition), 28 to the Milder degraded (22dB SNR) condition (6 in the single cue, 10 in the congruent cues, and 12 in the incongruent cues conditions) and 27 to the Strongly degraded (10dB SNR) condition (9 in each cue condition).

2.2.2. Material

All natural speech stimuli were recorded in a soundproof room by an EP female native speaker, sampled at a rate of 22.05 kHz and 16-bit conversion. Editing of the digitized versions of the stimuli (i.e., syllables, words, and part-words) was
made with *Adobe Audition 1.5* and *Praat 4.2.24* (available at [http://fon.hum.uva.nl/praat](http://fon.hum.uva.nl/praat/)). These natural stimuli were closely matched to a synthesized stream (*see Pretest below*) in all relevant acoustic parameters, such as speaking rate and absence of lexical stress cues to word boundaries. This also allowed us to match the concatenated and coarticulated versions of the stimuli (*see below*) on mean duration (by applying a compression rate of maximum 30%) and on pitch contour (through flattening to monotone 220 Hz).

**AL Material**

The repertoire of the AL was constituted by four consonants (C) (/b/, /k/, /l/, /f/) and three vowels (V) (/a/, /i/, /u/), which when combined rendered 12 possible CV syllables.

In EP, there is a close relation between vowel quality and word stress (e.g., Mateus & d’Andrade, 2000). Thus, selection of the vowels was especially critical in order to avoid that the vocalic repertoire of the AL indicated a possible word boundary. In order to eliminate the possible role of lexical stress in the process of learning the AL, four aspects were taken into account (Mateus & d’Andrade, 2000): (1) nasal vowels were not selected because their position in a word can mark the location of lexical stress; (2) word-final position only allows a limited set of unstressed vowels (i.e., /a/, /i/, /u/, /ø/), which can appear at any position in the word: final and non-final, pre-stressed and post-stressed; (3) /æ, /i/, /u/ can also be stressed vowels; and (4) the unstressed vowel /ø/ is usually deleted in colloquial speech.

The syllables of the AL were combined to create six trisyllabic TP-words (/bæbuku/, /bukælæ/, /lufæbu/, /kæfubæi/, /fufibæi/, /kilæbæu/). The TPs of the trisyllables
(words and trisyllabic part-words of the AL) were computed by averaging the two TPs associated to each stimulus. The TPs between adjacent syllables were always higher within than between TP-words, with the mean TP of adjacent syllables between TP-words being 0.38. Yet, since some syllables occurred in more TP-words than others did, three TP-words presented higher TPs (ranging from 0.75 to 1.00) than the other three (from 0.50 to 0.58). This distributional gradient is probably similar to what happens in natural languages (Saffran et al., 1996b) and might allow a fine-grained evaluation of the effect of TPs in the learning process.

AL-stimuli (i.e., TP-words and part-words) are presented in Appendix I.

For the familiarization phase, (syllable-based) concatenated exemplars and (TP-word or part-word based) coarticulated exemplars of the AL stimuli were constructed from natural speech. Three versions of the AL were created (see Table 1). Each version included the same sequence of syllables, divided into three listening blocks of approximately seven minutes each. Each block was created by concatenating 105 tokens of each of the six words (1890 syllables, 630 tokens of words), with the criterion that two tokens of the same word never occurred adjacent in the stream.

Table 1: Orthographic translation of a sample of the stream heard in the familiarization phase in the three cue conditions of Experiment 1. The "#" defines word boundaries according to TPs; the "~" represents concatenation; coarticulated syllables are presented in capitalized letters and underlined.

<table>
<thead>
<tr>
<th>CUE CONDITIONS (available cues)</th>
<th>SPEECH STREAM</th>
<th>PART-WORDS PART-WORDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>single cue (TPs)</td>
<td>lu-fa-ba-#ki-la-bu-#ka-fu-bi-#ba-bu-ku-#bu-ka-la-#fu-fi-bu …</td>
<td>ba-#ki-la ka-la-#fu</td>
</tr>
<tr>
<td>congruent cues (coarticulated TP-words)</td>
<td>LUFABA-#ki-la-bu-#ka-fu-bi-#ba-bu-ku-#bu-ka-la-#FUFIBU …</td>
<td>BA-#ki-la ka-la#FU</td>
</tr>
<tr>
<td>incongruent cues (coarticulated part-words)</td>
<td>lu-fa-BA#KILA-bu-#ka-fu-bi-#ba-bu-ku-#bu-KALA#FU-fi-bu …</td>
<td>BA#KILA KALA#FU</td>
</tr>
</tbody>
</table>
Chapter 2. Statistical and Coarticulatory Cues to word boundaries

The three natural versions of the AL differed in what regards the type of information available in the speech stream, as explained in the Introduction and illustrated in Table 1, in which an orthographic translation of a sample of the speech stream in each of the three versions of the AL is presented.

**Pretest 1: Natural vs. Synthesized Speech**

A pretest checked whether the use of naturally produced utterances, chosen here to achieve realistic coarticulation, had turned the task easier (Thiessen & Saffran, 2003) than the use of synthesized speech, which is more common in AL learning experiments. Synthetic stimuli (created using text-to-speech MBROLA software, cf. Dutoit, Pagel, Pierret, Bataille, & van der Vrecken, 1996) were presented to 21 independent volunteer undergraduates in the same AL experiment as the main one, but using only the single cue condition. Syllables were concatenated (with no other cues to word boundaries than TPs) using an EP female diphone database (available at http://tcts.fpms.ac.be/synthesis/mbrola.html) at 22.05 kHz and with a speech rate of approximately 270 syllables per minute. We evaluated two input-intelligibility conditions (intact speech and 22dB SNR), since a facilitation effect might appear for natural in comparison to synthetic speech in the more difficult (noisy) situation. This was not the case. The TP-word preference was similar in the synthesized and natural single cue conditions in the two signal intelligibility situations (with intact speech: \(t(19) = .73; \) with 22dB SNR: \(t(13) = .56; p > .10\) in both cases). Thus, both in intact and noisy conditions, the natural speech stimuli used in the main experiment induced statistical learning based on TPs in the same way as the synthetic material more commonly used in AL learning studies does.
Pretest 2: Identification in Noise - degraded signal conditions

In order to create the two degraded signal conditions, white noise was superimposed to each block of all three natural versions of the AL at 22dB SNR and at 10dB SNR. These SNR were selected on the basis of a pretest ran on an independent group of 23 volunteer undergraduate students. In this pretest, five between-subjects conditions were used: intact speech; 22dB SNR; 10dB SNR; 5dB SNR; and 0dB SNR. All the words and part-words of the AL were presented in randomized order, one at a time. They were played through headphones at 76dB SPL (which is approximately the level of conversational speech). Participants were informed that on each trial they would hear a pronounceable trisyllabic nonsense sequence and were required to write it down. This allowed us to choose the SNR that would reduce the stimuli intelligibility by approximately 50%, a value similar to the one used by Mattys (2004), operationalizing “intelligibility” as the total number of stimuli correctly identified. As expected for unfamiliarized listeners, correct responses did not significantly differ between the TP-words and part-words of the AL ($t(22) = -1.534, p > .10$). The best performance was observed for intact speech (91.7% correct identification, on the average). The 22dB SNR reduced performance by nearly 50%, leading to 47.9% correct identification, on the average, while the other SNR conditions induced much poorer performances (16.6% for 10dB SNR; 5% for 5dB SNR; 0% for 0dB SNR). In addition, the mean number of phonemes correctly identified (in their correct order) with 22dB SNR was 5.37 in sequences of 6 phonemes each. This indicates that the incorrect responses were not due to a large inability to identify any of the phonemes of the stimuli, as was the case, for example,
in the 0dB SNR condition. Thus, although the 22dB SNR impoverishes the signal, it
does not affect phonetic information in a way that turns it inaudible. This makes the
22dB SNR a perfect option, since the signal is degraded, but noise is not so strong to
disable the availability of many phonetic aspects of the speech input. Nevertheless,
there are important differences between this pretest task, in which naïve participants
were required to identify the TP-words and part-words of the AL presented one at a
time, and the AL learning situation, in which participants are repeatedly exposed to
these stimuli embedded in longer streams of speech. Thus, possibly, the intelligibility
level obtained with a specific SNR in the identification in a noise task does not
correspond to the degree of intelligibility obtained with the same SNR in an AL
learning task. One should note that, usually, for words it is a 0dB SNR that reduces
word intelligibility by about 50%. Although our AL material is rather different from
real words, it is plausible that repeated exposition to the same material would lead to
a higher level of intelligibility in the AL situation than the one used by Mattys
(2004). Thus, another condition with superimposition of a higher level of white noise
was also chosen, namely 10dB SNR. At least in the pretest, the 10dB SNR had a
strong impact on the identification of the AL stimuli, reducing dramatically the
correct identification not only of the nonwords but also of the phonemes that
constitute them (3.9 out of 6).

The **forced-choice test** included the six TP-words and six part-words. These
part-words consisted of syllables of two different words that had appeared adjacently
in the speech stream. Three part-words (*Part-words 3#12* in Table 1) were formed by
the last syllable of one word (e.g., /br/ which is the last syllable of /lufbr/ and the
first two syllables of the next word (e.g., /kila/ which are the first two syllables of /kilsbu/). The others (Part-words 23#1 in Table 1) were constituted by the last two syllables of a word and the first syllable of the next word. For all conditions and groups, the stimuli used in the test phase were produced by concatenating the three syllables that constituted each word or part-word without white noise superimposed, thus avoiding responses based on acoustic matching between the stimuli of the familiarization and test phases.

2.2.3. Procedure

Presentation in the familiarization phase was done with Windows Media Player with all the auditory stimuli presented at a comfortable level through Sennheiser HD 280 headphones. For the test phase, stimuli were also presented through headphones, with presentation, timing and data collection controlled by E-Prime 1.1 (Schneider, Eschman, & Zuccolotto, 2002a; 2002b).

All participants were instructed to listen to a new language that contained “words”, but no meaning or grammar. Their task was to find out what words constituted the new language. No information about the structure, phonology or length of the words was given. Participants were informed that the experiment consisted of three short listening blocks, followed by a test of their knowledge of the words that constituted the language. Participants in the degraded signal conditions were warned of the poor signal quality. After each of the 7-minutes blocks, a 5-minutes break was provided. After the listening phase, participants were presented with a two-alternative forced-choice test. Each trial started with a warning tone,
followed by two trisyllabic strings, separated by 500 ms of silence, presented through headphones. One of these strings was a word from the AL, the other was a part-word. Each word was paired exhaustively with each part-word, rendering 36 trials. Immediately after participants gave their answer, another trial began. If no answer was registered after 10 seconds, the next trial began. Participants were told to always provide an answer, even if not totally sure about their decision. Nevertheless, accuracy was also emphasized. The test began with four practice trials, in which an animal and an environmental sound were presented and participants had to decide which of the two stimuli corresponded to the animal sound. Feedback was only provided for practice trials.

Order of presentation of test trials was randomized for each participant, and order of presentation of the stimuli within trials was counterbalanced within each group.

2.3. Results and Discussion

The percentage of TP-word choices was computed for each participant. In Table 2, the average AL learning performances are presented broken down by TP level (high vs. low level of TP of the TP-words).

In all input intelligibility conditions, participants presented with the single cue chose the TP-word significantly more often than the part-word: with intact signal, 67% (SD= 6.4) ($t(8) = 7.917, p < .0001$); with milder degraded signal (22dB SNR), 62% (SD=8.2) ($t(5) = 3.605, p < .025$); with strongly degraded signal (10dB SNR),
58% (SD=7.1) ($t(8) = 3.53, p < .01$). A significant learning effect was also found in all intelligibility conditions for the participants presented with congruent cues: with intact signal, 84% (SD=9.4) ($t(10) = 11.985, p < .001$); with 22dB SNR, 72% (SD=12.2) ($t(9) = 5.69, p < .001$); with 10dB SNR, 61% (SD=4.7) ($t(8) = 6.897, p < .001$).

Table 2: Performance pattern (proportion of TP-Word responses, in percentage) according to cue condition (single cue; congruent cues; incongruent cues) and Signal Quality (intact speech; 22dB SNR; 10dB SNR), in Experiment 1, considering the TP gradient of the TP-words of the AL (High-TP; low-TP).

<table>
<thead>
<tr>
<th>SIGNAL QUALITY</th>
<th>Intact</th>
<th>22dB SNR</th>
<th>10dB SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High TP</td>
<td>Low TP</td>
<td>High TP</td>
</tr>
<tr>
<td>AL Condition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single Cue</td>
<td>66.1</td>
<td>66.6</td>
<td>62.2</td>
</tr>
<tr>
<td>Congruent Cues</td>
<td>82.2</td>
<td>85.0</td>
<td>66.6</td>
</tr>
<tr>
<td>Incongruent Cues</td>
<td>39.4</td>
<td>41.6</td>
<td>50.0</td>
</tr>
</tbody>
</table>

Chance level corresponds to 50%.

In sharp contrast with this pattern, with intact signal, participants presented with incongruent cues discarded the TP-words, with an average of only 41% TP-word choices (SD=10.5). Thus, in this condition participants chose the part-words as the lexical units of the new language significantly more often than the TP-words ($t(11) = -3.040, p = .01$). With milder degraded signal, performance with incongruent cues did not differ from chance ($t(11) = -.389, p = .704$), with 50% TP-word choices (SD= 12.5). Thus, no learning effect was observed. It was only in the strongly degraded condition that participants presented with incongruent cues significantly preferred the TP-words over the part-words, reaching, on average, 57% TP-word choices [SD=7.6; $t(8) = 2.794, p < .025$].
In order to evaluate directly both the weighting of the two studied cues and the impact of signal quality on this weighting, we ran an ANOVA including cue condition (single cue; congruent cues; incongruent cues) and signal quality (intact speech, 22dB SNR and 10dB SNR) as between-subject factors. In addition, we included the high vs. low level of TPs of the TP-words as a within-subject factor in order to check whether there was a TP gradient.

The cue condition effect \(F(2, 78) = 49.1, \ p < .001, \ MSE = 5.66\) was significantly modulated by signal quality, as revealed by the interaction between these two factors \(F(4, 78) = 11.551, \ p < .001\). No other main effect or interaction was significant [all \(Fs < 2, \ p > .1\)].

We further investigated the nature of the significant interaction through pairwise comparisons, using the Bonferroni-corrected alpha rate of .017.

As clearly illustrated in the left-hand part of Figure 2, with intact signal a main effect of cue condition was found \(F(2, 29) = 64.77, \ p < .001, \ MSE = 10.87\), with congruent cues leading to better performance than both the single cue \(F(1, 29) = 16.55, \ p < .0005\) and the incongruent cues \(F(1, 29) = 128.74, \ p < .0001\) conditions. Not surprisingly, the incongruent cues condition, which led participants to prefer part-words over TP-words, also differed from the single cue condition \(F(1, 29) = 43.47, \ p < .001\).

As can be seen in Figure 2, with milder degraded (22dB SNR) signal, [main effect of cue condition: \(F(2, 25) = 11.24, \ p < .001, \ MSE = 17.33\], incongruent cues led to a lower performance than congruent cues \(F(1, 25) = 20.29, \ p < .001\).

As displayed in the right-hand part of Figure 2, contrary to what had been observed in the former two input-intelligibility conditions, with a strongly degraded...
(10dB SNR) signal, all three cue conditions led to a similar performance level \[ F(2, 24) = .73, p = .48, MSe = 2.815 \].

Figure 2: Performance pattern (proportion of TP-Word responses, in percentage) according to cue condition (single cue; congruent cues; incongruent cues) and of the signal quality (intact speech; 22dB SNR; 10dB SNR) – Experiment 1. Vertical bars denote 0.95 confidence intervals. Chance level corresponds to 50%.

The data show that there was actually no major impact of signal quality in the single cue condition \[ F(2, 21)= 3.33, p > .05, MSe = 6.58 \]. Thus, the statistical learning mechanism operating on TPs between adjacent syllables seems very resilient to noise, allowing speech segmentation to be almost as efficient when the signal is quite distorted (i.e., 10dB SNR) as when it is highly intelligible \[ F(1,78) = 3.84, p > .05 \].

In contrast, signal quality significantly affected performance in the congruent cues condition \[ F(2, 27) = 14.94, p < .001, MSe = 11.46 \]. A significant linear trend
[\(F(1, 78) = 30.05, p < .001\)] suggests that congruent cues to word boundaries are integrated in good listening conditions, allowing the optimization of the speech segmentation process. But integration was affected and hence the redundancy gain\(^{11}\) largely reduced as signal degradation increased. Indeed, the number of TP-words choices was significantly higher with intact speech than with noise superimposition \([22\text{dB SNR and 10dB SNR}, F(1, 78) = 8.47 \text{ and } 30.05, \text{ respectively, } p < .005\]). In addition, performance was also better with milder (22dB SNR) than with strongly (10dB SNR) degraded speech \([F(1, 78) = 6.72, p = .011]\).

This pattern of progressive reduction of the redundancy gain seems related to the strong sensitivity of coarticulatory cues to noise. Indeed, as reported above, the redundancy gain observed with congruent cues compared to the single cue condition was significant only with intact speech. It was still (numerically) present but not statistically significant anymore with mildly degraded speech (22dB SNR), and no longer observed at all with strongly degraded speech (10dB SNR).

The sensitivity to noise of coarticulatory cues to physical noise is even more clearly revealed by the performance pattern observed with incongruent cues. As a matter of fact, with incongruent cues the effect of noise was significant \([F(2, 30) = 6.16, p < .01, MSE = 14.53\)], and modulation of performance by signal quality is reflected by a significant linear trend \([F(1, 78) = 15.74, p < .001\]). As already reported, participants relied on coarticulation rather than on statistical information only when exposed to intact speech, a listening condition that differed significantly from strongly degraded (10dB SNR) speech \([F(1, 78) = 15.74, p = .0001\]). Indeed, with low noise intensity (22dB SNR), the inconsistency between the two types of

\(^{11}\) The term “redundancy gain” is used to emphasize the improvement in the segmentation process revealed by the superior AL learning performance in the congruent cues condition with intact speech, as the outcome of the availability of different consistent segmentation cues in the stream.
information disrupted performance, but the influence of coarticulation totally vanished only with the most degraded input (10dB SNR), a situation in which listeners considered only statistical information, performing at the same level as listeners exposed to either a single cue or congruent cues.

2.4. General Discussion

In Experiment 1, within an ALL setting, we investigated, through three signal-quality conditions, the relative impact of two types of sublexical information in speech segmentation: statistical information – TPs between adjacent syllables, and subsegmental information – coarticulation. The results presented in this chapter suggest, in accordance with Mattys and colleagues’ (2004; Mattys et al., 2005) proposal, that the segmentation process that listeners adopt varies as a function of both the types of cues available in the speech flow and signal quality.

Indeed, with phonetically intact speech, coarticulation was a powerful segmentation cue, able to drive the segmentation process. Most importantly, coarticulation overruled statistical information when both cues were put in conflict in the speech stream, which is in line with Johnson and Jusczyk’s (2001) infant data. Mattys (2004; Mattys et al., 2005) had already demonstrated coarticulation reliability in adults, when this cue was in conflict with lexical stress. Our results thus add to this evidence. Thus, in good listening conditions, coarticulation is given priority to either lexical stress (Mattys, 2004; Mattys et al., 2005) or a general segmental statistical cue (TPs, in the present experiment). It is not clear however, if it is either the local coherence given by coarticulation or the perception of edges between concatenated
and coarticulated parts of the speech stream (or even both factors) that are the basis for a coarticulation-driven segmentation procedure.

In any case, the perceptual salience of coarticulatory information is extremely affected by signal degradation. As a matter of fact, both Experiment 1 of the present study and Mattys’ data (Mattys, 2004; Mattys et al., 2005) show that coarticulation is much more affected by the presence of noise than other cue types. Indeed, when only sublexical information is available in the stream, segmentation driven by coarticulation is only observed with phonetic intact speech. When information of the three tiers is available, involving both lexical and sublexical information, coarticulation only has an effect in mild-noise conditions (Mattys et al., 2005). Remarkably, as happened in the mild degraded condition of the present study, in a condition of moderate-noise superimposition Mattys et al. (2005) also observed no preponderance of any of the incongruent sublexical cues. Nevertheless, it seems that in particular conditions of signal degradation, in which some acoustic properties are still available, the cue less reliable in noisy conditions (i.e., coarticulation) is still sufficiently available to disrupt the segmentation driven by an inconsistent cue (lexical stress, in Mattys et al., 2005, and TPs, in the present experiment). However, in this condition, this cue becomes unable to drive the segmentation process by itself and to override statistical information. This pattern is reflected, in the present 22dB SNR condition, by two pieces of evidence: (i) the (non significant) trend toward a redundancy gain observed when coarticulation and statistical cues suggested the same word-boundaries; (ii) the (significant) interference effect (leading to no AL learning) when these two cues pointed towards conflicting segmentation points.
In short, the incongruent cues condition showed that coarticulation overrides TPs only in intact speech, while TPs override coarticulation in strongly degraded speech. The pattern observed with congruent cues suggests an additive effect of converging cues, at least in good listening conditions, allowing the optimization of the speech segmentation process. However, whether this redundancy gain is synergistic (greater than the sum of the isolated cues effect) or conjunctive (Christiansen & Curtin, 2005) cannot be derived from the results of Experiment 1, since no “coarticulation-only” condition was assessed (the single cue condition was a TP-only condition). In any case, there is a performance gain when statistics and coarticulation are consistent in comparison with the condition in which only statistical information is available. But integration was affected and hence the redundancy gain is largely reduced as signal degradation increased.

The pattern of results reported here was not modulated by the TP level of the TP-words. In fact, we did not find any hint of a TP gradient effect: TP-words that presented higher TPs (ranging from 0.75 to 1.00) were not better learned than the others (with TPs from 0.50 to 0.58). Since such a gradient was reported by Saffran et al. (1996b), at least in a condition in which only statistical information was available, we suspect that this null result is partly due to the fact that the TP range within the TP-words used in our study is narrower (from 0.50 to 1.00) that the one used by Saffran et al (which varied between 0.31 and 1.00).

Much more important is the fact that the present pattern of results cannot be attributed neither to the use of natural speech, nor to the frequency difference of TP-words and part-words.
Indeed, it has been suggested that the use of natural speech in AL learning studies could make the task easier than with synthesized speech (Thiessen & Saffran, 2003). However, the similarity between the performance levels of the TP groups exposed to synthesized and natural speech (see pretest in section 2.2.2.) clearly demonstrates that this dimension cannot account for the present results.

As regards the raw frequency differences between TP-words and part-words, we know that, although this factor is potentially important to word discovery by infants (Brent & Cartwright, 1996) and can explain participants’ performance in some studies (e.g., Dahan & Brent, 1999), it cannot always explain the impact of TPs (Aslin, et al., 1998). In Experiment 1, we tried to minimize the potential confound between raw frequency of occurrence of the stimuli and TPs. In AL conditions in which coarticulatory information was available, the speech stream was constituted by concatenated syllables, and the number of coarticulated tokens of words in the congruent cues condition was equivalent to the one of part-words in the incongruent cues condition. Thus, the raw frequency of coarticulated exemplars was matched in these two conditions. The learning pattern (or the significant absence of learning, in the incongruent cues condition with intact signal) shows that listeners did not rely (or at least not only) on the absolute frequency of the items. Had listeners used only frequency of the trisyllabic items to analyze the speech stream, performance should have been similar in the three cues conditions – all groups should have learned the AL equally well, with performance depending exclusively on signal-quality. On the opposite, the AL learning patterns differed according to both the segmentation cues available in the speech stream and noise contingency.
The present experiment also shows that both the influence of noise and the involvement of the various sources of information available vary in a graded manner. In addition, the weighting modulation by noise seems to be mainly related to the listeners’ inability to exploit coarticulatory information when the signal is distorted. It is only in that case that the segmentation process is driven solely by statistical information. Indeed, in 10dB SNR condition, a similar AL learning performance was observed independently of the number and congruence of the segmentation cues available in the stream. This suggests that coarticulation was no longer available to help segmentation. Thus, the weighting change does not seem to reflect that the less one cue type is used, the more others will gain importance. Instead, it largely depends on both the unavailability of coarticulatory cues, and the high resilience of statistical information to noise superimposition.

In summary, noise superimposition does not affect at a similar degree all segmentation cues, neither when these pertain to the distinct tiers defined by Mattys et al. (Mattys, 2004; Mattys et al., 2005) nor when they are both sublexical, as was the case here. Does this imply that the segmental and subsegmental information should be assigned to qualitatively different “tiers”, contrary to what posited by Mattys et al.’s (2005) hierarchical model? Providing a definite answer to this question is difficult on the basis of the available evidence. In conditions in which segmental (phonotactic, cf. McQueen, 1998) and subsegmental (coarticulation, cf. Mattys, 2004) information were independently put against metrical prosody, both overruled the last cue. This may support the need for a broad distinction between sub-lexical (either segmental or sub-segmental) cues and prosodic information. However, as already suggested by Mattys et al. and as clearly demonstrated by the
present results, although in any natural language segmental and subsegmental information tend to be intrinsically correlated, they might be differentially weighted.

Part of this differential weighting, and hence of the resilience of cues to noise, might be related to the structural grain of the cues, as proposed by Mattys et al. (2005) and already commented on. Indeed, the use of cues defined at a lower structural level depends on the ability to process fine-grained, low-level acoustic properties, which are more easily masked by noise than higher-level units like syllables (involved here in the TPs). Note however that cue reliability cannot be reduced to perceptual salience *per se*. Indeed, if it was always easier to extract information from highly salient syllables than from perceptually less salient sub-syllabic (segmental or sub-segmental) structures, one would never observe coarticulation to overrule statistical information when both cues are put in conflict. Yet, this was the case with intact speech, both in the present experiment and in Johnson and Jusczyk’s (2001) infant study. Nevertheless, to fully understand the role of the structural grain of the cues, it would be interesting to contrast, under various noise conditions, the power of (subsegmental) coarticulation and of distributional regularities defined at the segmental level, such as phonotactic cues. Phonotactics have been shown to intervene in AL learning, at least when nonadjacent TPs are considered (Onnis et al., 2005).

It would also be interesting to compare, as it is presented in Experiment 4 (*see chapter 4 of the present thesis*), the resilience to cognitive noise of the two types of cues we used here through the use of (potentially) interfering tasks of various levels of difficulty (cf. Toro et al., 2005). If TPs were found to be the most resilient
segmentation procedure, not only to physical noise, as in the present experiment, but also to cognitive noise, any interpretation based only on physical masking of acoustic properties would be dismissed.

In fact, the differential weighting of cues may not depend only on the structural grain of the cues, but also on two basic factors, the first one being *domain generality*. Indeed, the ability to track TPs involves a domain-general learning mechanism, as demonstrated by the fact that TPs are also extracted in tone (Saffran et al., 1999) and visual (Fiser & Aslin, 2001) sequences. On the contrary, coarticulatory information is speech-specific. Within the speech domain, it might be useful to consider the further distinction between *universal cues*, like intonational phrases which partly correspond to physiological mechanisms like breath groups (Shukla et al., 2007), and *language-specific cues*, namely properties that depend on the particular language. Indeed, the latter properties obviously need to be learned, while both domain-general statistical mechanisms and speech-specific universal cues are available since a very early phase in language acquisition and are used by adults even with unknown or foreign languages (Shukla et al., 2007).

Both domain-general statistical mechanisms and speech-specific universal cues might therefore have a central role in development as guides to other segmentation cues. To act as reliable guides, such cues should be relatively immune to signal degradation. In other words, as suggested by Mattys et al.’s (2005), the more resilient (but lower-weighted in normal listening conditions) cues in adult speech segmentation could correspond to the earliest acquired, and hence, to the most critical cues at the onset of language development. Further work should be aimed at testing these propositions. For example, since various languages have
different patterns of coarticulation that often reflect the phonetic contrasts that are emphasized (see Manuel, 1999, for review), the importance of the language-specificity vs. universality of the cues may be assessed by contrasting the role of TPs to the role of language-specific as opposed to language-general patterns of coarticulation in speech segmentation.

Contrasting universal prosodic cues used by Shukla et al. (2007) to language-specific prosodic cues like word stress patterns used by Mattys (2004) and Mattys et al. (2005) may also shed light on the relevance of this distinction. This was done in Experiments 2 and 3B of the work presented in this thesis (see Chapter 3). In the next chapter we have evaluated the weighting of both domain-general statistical mechanisms and speech-specific cues, such as the universal prosodic properties examined by Shukla et al. (2007) under physically degraded conditions (see Experiment 2, presented in chapter 3). Indeed, with intact speech Shukla et al. observed that phrasal prosodic cues seem to act as a filter, suppressing possible word-like sequences (trisyllabic sequences with high TPs) that straddle two prosodic constituents (this proposal is presented in more detail in Chapter 3). Whether this would hold true in noisy situations was tested in the experiments presented in the next chapter of this thesis. Importantly, although Mattys (2004; Mattys et al., 2005) already demonstrated that lexical stress acts as a last resource segmentation heuristic in speech segmentation for English adult listeners, we do not know whether this could also hold truth for EP listeners. As a matter of fact, this was evaluated in Experiment 3B of the next chapter. Thus, in Experiment 3 we were able to evaluate directly whether the nature of sublexical cues (i.e., their domain generality and their
role in particular languages) is in fact important on their weighting in different listening conditions.

In summary, the AL learning patterns observed in Experiment 1 of the present work have shown three important facts: (i) the modulation of coarticulation reliability by signal quality; (ii) the high resilience of statistical information based on TPs to noise superimposition; (3) the strong signal contingency of the weighting of the cues used to segment speech into words. This pattern of results can be well accommodated by Mattys’ (2004; Mattys et al., 2005) hierarchical proposal, and highlights the importance of studying speech segmentation in the context of multiple cues (e.g., Christiansen & Curtin, 2005) and in different listening conditions (Mattys, 2004; Mattys et al., 2005). Importantly, an integrated approach must be able to apprehend the role of sublexical information in speech segmentation, how sublexical cues interact with lexical and supra-lexical information, and how the weighting of several cue types is affected by listening conditions.
Chapter 3.

Universal and Speech-specific Prosodic cues in Artificial Language Learning:

The role of physical noise.

Until now, the weighting of general domain and speech-specific cues in speech segmentation was left largely unspecified. In the present chapter, the impact of physical noise on the weighting of three sublexical cues: transitional probabilities (TPs), universal prosody and lexical stress, was investigated. In artificial language learning, while both universal prosody and TPs were highly resilient to signals’ physical degradation, with the former prevailing over the latter, being able to drive the segmentation process even when incongruent cues were available in the stream (Experiment 2), stress pattern effects only emerged in degraded conditions (Experiment 2 and Experiment 3B). Thus, speech segmentation does not seem to be the product of one preponderant cue acting as a filter of the outputs of another lower weighted cue. Instead, it mainly depends on the weighting of cues, according to their nature and the listening conditions.

3.1. Introduction

The present chapter will focus on two types of sublexical cues that are available since a precocious phase in linguistic development, in particular: prosodic (suprasegmental) and statistical (segmental) sources of information.

As already refer in section 1.1.2. of chapter one of the present thesis, prosody
was one of the first sublexical speech segmentation cues to be proposed (e.g., MMS, Cutler & Norris, 1988; Gleitman & Wanner; 1982), and listeners are sensitive to it from the very beginning of language onset.

Adult listeners are also sensitive to several types of prosodic cues (for a review see Cutler, Dahan & van Donselaar, 1997), such as IPs’ edges (e.g., Shukla et al., 2007); phonological phrases boundaries (e.g., Christophe, Gout, Peperkamp, & Morgan, 2003), metrical units (e.g., Cutler & Norris, 1988), and lexical stress (e.g., Mattys, 2000), and these cues play some role in lexical activation (e.g., Christophe, Peperkamp, Pallier, Block, & Mehler, 2004; McQueen et al., 1994; Norris et al., 1995; Salverda, Dahan, & McQueen, 2003).

Another important speech segmentation cue is the statistical information conveyed by the speech stream. In particular, many studies have focused on TPs between adjacent syllables. Indeed, within any natural language, the TP from one syllable to the next is generally higher within a word than between words (e.g., Perruchet & Peereman, 2004; Swingley, 2005). The ability to track TPs appears to be precocious (e.g., Kirkham et al., 2002; Thiessen & Saffran, 2003), rapid (e.g., Saffran et al., 1996a), involuntary (e.g., Saffran et al., 1997; but see Toro et al., 2005) and age-independent (e.g., Saffran et al., 1997). Since TPs computation does not require any, even minimal, lexical knowledge, it might allow learners to segment speech from the discovery of troughs in the TPs distribution between adjacent syllables, acting as a pivot mechanism (cf. Mattys et al., 2005) in the acquisition of not only words but importantly also other word-boundary cues (Thiessen & Saffran, 2003).

The integration of multiple available segmentation cues may provide
evidence about linguistic aspects that cannot be derived from any single source, promoting the optimization of speech segmentation (see Christiansen, Allen, & Seidenberg, 1998; Christiansen & Curtin, 2005). Therefore, the realistic understanding of natural speech segmentation and how cues are weighted on its assistance can only be achieved if the role of different cues, simultaneously available, is evaluated in both congruent and incongruent conditions, namely when these cues suggest either the same or incompatible segmentation hypotheses as regards the location of word boundaries, respectively.

3.1.1. The interaction between Prosody and TPs in speech segmentation

Recently, a considerable amount of research has been devoted to the integrated study of statistical and suprasegmental cues. In optimal listening conditions (i.e., with intact speech), prosodic information (e.g., strong syllables, syllable lengthening, lexical stress) seems to be underestimated by adult listeners when other speech segmentation cues are available in the stream. Using word-spotting, McQueen (1998) found that phonotactic legality is a preponderant cue able to drive speech segmentation processes independently of metrical prosody. Within an artificial language learning (ALL) setting, Saffran et al. (1996b) also showed that statistical learning is not disrupted by the presence of an incongruent prosodic cue (i.e., initial syllable’s lengthening). As a matter of fact, in such a condition American listeners were as able to extract from the stream the AL words based on the high TPs between their syllables (henceforth, TP-words) as listeners only exposed to the
statistical information. Statistical learning immunity to incongruent prosodic cues is also observed with listeners from other linguistic backgrounds (e.g., French listeners: Vroomen et al., 1998). Using the cross-modal priming effect in a lexical decision task, Mattys (2004; Mattys et al., 2005) also demonstrated that listeners disregarded lexical stress when this suggested segmentation hypotheses that were incongruent with coarticulation, phonotactics, or lexical context.

Even when prosodic cues are congruent with other sources of information, the expected benefit or redundancy gain (cf. Fernandes, Ventura, and Kolinsky 2007; see chapter two of the present thesis) is far from being consistently observed. On the one hand, Saffran et al. (1996b) reported that when TPs and prosodic information (final syllable lengthening) were congruent, listeners’ ALL performance was improved in comparison to a condition in which only statistical information was available in the stream. The same benefit was found for Finish and Dutch listeners (with initial stressed syllables marked by a F0 peak; Vroomen et al., 1998) and for French listeners (with lengthening and/or F0 peak of the last syllable; Bagou et al., 2002). But, on the other hand, there are numerous studies in which the congruency of prosody with other cues did not have any (positive or negative) impact. In ALL, Valian and Levitt (1996) only found a benefit promoted by “phrase prosody” (i.e., the rising pitch contour on the first two-word phrase of a sentence and a falling pitch on the other last two-word phrase) when listeners were unable to use other cues, since neither marker frequency nor a reference field were available. Mattys (2004; Mattys et al., 2005) also showed that when coarticulation, phonotactics, or lexical cues were available in the stream, the same pattern of results was found whatever the congruency of lexical stress with these cues. Toro-Soto, Rodríguez-Fornells and
Sebastián-Gallés (in press) evaluated whether Spanish listeners’ ALL performance would beneficiate from the availability of an increasing pitch in the “stressed” syllable of TP-words. Surprisingly, while listeners familiarized with the AL with TP-words stressed on the penultimate syllable (i.e., the default lexical stress pattern in Spanish) were not able to learn the TP-words, when stress occurred either in the first or in the last syllable of trisyllabic TP-words (which are also legal stress patterns in Spanish), listeners performed at a similar level than when only TPs were available in the stream. However, Toro-Soto et al.’s results need to be interpreted with cautious, since it is not clear that pitch is a strong correlate of lexical stress in Castilian Spanish. In fact, in Spanish, duration seems to be the strongest acoustic correlate of lexical stress, regardless of the presence of a pitch accent (Ortega-Llebaria, 2006).

The impact of prosodic cues on speech segmentation is quite different when the speech signal is degraded by noise superimposition. In this case, lexical stress is able to override any other incongruent segmentation cue, either lexically-driven (e.g., the semantic context) or signal derived, like coarticulation and/or phonotactics (Mattys, 2004; Mattys et al., 2005). Thus, speech segmentation is largely the product of the differential weighting of the types of information available in the signal and of the listening conditions (Mattys, 2004; Mattys et al., 2005; see section 1.4. of chapter one).

Shukla, Nespor, & Mehler (2007)

In apparent contradiction with this model, Shukla et al. (2007) recently showed that IPs (i.e., one of the highest prosodic structures; Grice, 2004) act as important segmentation cues even when TPs between adjacent syllables (an
information represented at a higher tier in Mattys et al.’s 2005 model) are available in a phonetically intact signal. Using a variant of the ALL paradigm, when IPs contour was available (with left edge of IPs marked by a raising pitch and shorten duration of the beginning syllables, and right edge marked by a falling pitch and lengthen duration of final syllables), only TP-words within IPs (i.e., at middle positions) were correctly extracted from the stream. In contrast, TP-words straddling IPs (with TP-words’ first syllables at the right-edge of one IP and the last syllable at the left-edge of the next IP) were not selected by listeners as “words” of the new language, probably because, although statistically cohesive, these TP-words straddled a prosodic boundary. Importantly, when some TP-words were aligned with prosodic edges (and thus were cohesive units on both statistical and prosodic bases), while others were in IP’s middle position (not prosodically marked and hence only cohesive units on statistical grounds), only TP-words supported by the two types of information (i.e., TP-words at IP edges) were correctly extracted from the stream. Based on these results, Shukla et al. (2007) proposed that prosody could act to filter the output of statistical computation, with only TP-words compatible with it being selected.

While Shukla et al.’s (2007) results may seem at odds with the unreliability of metrical prosody observed in previous studies with intact speech, the differential weighting of cues possibly depends also on two basic factors (cf. Fernandes et al., 2007; see chapter two), the first one being domain generality. The ability to track TPs involves a domain-general learning mechanism, since TPs are also extracted in non-linguistic materials (e.g., musical tones: Saffran et al., 1999; visual sequences: Fiser & Aslin, 2001; Kirkham et al., 2002; but see Conway & Christiansen, 2005,
2006) with a possible phylogenic origin (e.g., Hauser et al., 2001). In contrast, prosody is speech-specific. Within the speech domain, including prosodic information, we can also consider the further distinction between universal cues and language-specific cues. Universal cues are probably physiologically based (Grice, 2006; Shukla et al., 2007; Werner & Keller, 1994), used in any language from the onset of development as well as by adults confronted with unknown/foreign languages (Shukla et al., 2007). In contrast, language-specific cues are properties that depend on the particular language and thus latter acquired. For example the role of IPs’ contour seems to be universal, since some of its prosodic correlates such as the ones of prosodic right-edges (which characterize IPs in many different languages; Werner & Keller, 1994) are acoustic marks of the slowing down of the articulators within a breath group, which is reflected in the signal as final lengthening and low pitch (Grice, 2006, see also Cutler et al., 1997). On the opposite, lexical (or word primary) stress, although generally correlated with duration, F0 and amplitude (stressed syllables are lengthened, have higher pitch, and are louder), is not acoustically marked in the same manner in all languages. For example, while in Finish pitch seems to be the most important correlate of lexical stress (e.g., Iivonen et al., 1998), in EP lexical stress is marked by syllable lengthening (Delgado-Martins, 2002; d’Andrade & Lacks, 1996). The location of lexical stress and whether it obeys to a fixed or varied pattern is also language-dependent: while in Finish it is always on the first syllable of a word: Iivonen et al., 1998; in EP lexical stress is by default on the penultimate syllable, although it can occur in any one of the three last syllables of a polysyllabic word (d’Andrade & Laks, 1996; Mateus & d’Andrade, 2000).

Thus, while both general-domain and universal speech-specific cues are used
Chapter 3. Universal and Speech-specific cues in ALL

from the very beginning of language onset, language-specific cues will only occupy a role latter on speech segmentation, which is possibly modulated by their ability to predict word boundaries in specific languages (cf. Mattys et al., 2005). In fact, when TPs and lexical stress (i.e., a language-specific cue) suggest different segmentation points in the speech stream, while 6.5-months-old infants seem to consider TPs as the more reliable cues (Thiessen & Saffran, 2003), at 9-month-old, the weighting of these cues is reversed (Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003). At this age, infants are also sensitive to statistical segmentation cues of their native language, namely to phonotactics (the relative frequencies of segments and sequences of segments in syllables and words; cf. Mattys & Jusczyk, 2001), and by 10.5 months, infants are already able to integrate multiple sources of information, while language-specific prosodic cues start to loose their previous importance (Jusczyk et al., 1999a; 1999b). In line with Mattys et al.’s (2005) proposal, these language-specific sublexical segmentation cues, lower weighted in adulthood, such as lexical stress, seem to be early acquired, having a predominant critical role within a transitory phase in infancy. After this period, they gradually loose their reliance or are supplanted by other sublexical cues with higher word boundaries predictability.

3.1.2. An overview of the experiments presented in this chapter

Until now, the question of how, in speech segmentation, universal cues, like domain-general TPs and speech-specific IPs interact with language-specific cues such as lexical stress has been left unsolved. To the best of our knowledge, no study on adult listeners has ever evaluated the relative power of these three types of cues in
different physical listening conditions. This was the general aim of the experiments presented in this chapter.

On the basis of Fernandes et al.’s (2007; see Experiment 1 presented in the previous chapter of this thesis) work, we already know that universal, general-domain, segmentation cues such as TPs are very resilient to the physical degradation of the signal, being able to drive segmentation at similar levels in both good and noisy listening conditions. That is probably due to the fundamental role that TPs play in speech segmentation. According to this logic, IPs - another universal and fundamental cue - would have the same resilience to signal degradation. However, these two cues would be differentially weighted. IPs contour is a speech-specific cue, probably processed as encapsulated information of a modular speech perception system (Fodor, 1983, 2000; for neuroanatomic evidence, see e.g. Dehaene-Lambertz, Hertz-Pannier, & Dubois, 2006; Vouloumanos & Werker, 2007) and hence its processing would be mandatory and automatic. Whatever the listening conditions, this would confer such a universal speech-specific cue a higher weighting in speech segmentation, than TPs, because extraction of the latter corresponds to a general-domain mechanism.

However, since in Shukla et al. (2007) the potential AL words were always suggested either by TPs alone or by TPs and congruent IP-edges, we do not know whether IPs-edges may act independently and overrule TPs. This can only be evaluated in a condition in which the statistical outputs (i.e., the TP-words) would be incongruent with the prosodic outputs, namely when plausible “words” according to TPs straddle an IPs-edge, while plausible “words” according to IPs-edges straddle TP-boundaries. The last stimuli are called part-words, since they are AL stimuli of
the same length as TP-words, constituted by the same AL syllables repertoire, but with lower TPs than TP-words, since they are constituted by syllables that occurred adjacently in the stream but that straddle a TP-word boundary. If listeners preferred prosodically plausible words even when these straddle a TP-boundary, they would discard the TP-words and consider the part-words as the correct units of the AL.

This possibility was evaluated in Experiment 2. Within an ALL setting, two input-intelligibility conditions were evaluated (between-participants), namely intact (with no white noise superimposition) and mildly degraded signal (at 22dB SNR\textsuperscript{12}). Within each input-intelligibility condition, the same AL was presented in four (between-participants) conditions differing by the number and congruence of the segmentation cues available in the speech stream: a single cue condition, in which only TPs were available in the stream; a congruent cues condition, in which statistics suggested the same segmentation hypotheses as the IPs’ right-edges defined, in agreement to what is observed in natural speech, by syllable lengthening and falling pitch (e.g., Grice, 2006; Saffran et al., 1996b; Shukla et al., 2007); and two incongruent cues conditions, in which statistical and prosodic information suggested different segmentation hypotheses, with TP-words spanning the prosodic boundary defined by IPs-edges.

The use of two incongruent cues conditions allowed the independent evaluation of universal prosody (i.e., right-edge marking) and language-specific prosody (i.e., primary word stress). Indeed, one of the acoustic correlates of prosodic right-edges we used, namely syllable lengthening (but not pitch, cf. Grønnum & Viana, 1999), is also a correlate of lexical stress in Portuguese. Thus, syllable

\textsuperscript{12} A 22dB SNR means that if the signal intensity is 76dB, as was the case here, then the noise intensity was set at 54 dB. This SNR was chosen based on the Identification in Noise pretest of Experiment 1, in order to reduced intelligibility by approximately 50% (see also chapter two, section 2.2.2.).
lengthening may possibly be perceived as corresponding to the primary stress of TP-words. For this reason, we manipulated the position of the prosodically marked syllable. In the *incongruent cues-1st syl condition*, the first syllable of TP-words was prosodically marked as a right edge, which does not correspond to the default stress pattern in the listeners’ native language. In the *incongruent cues-2nd syl condition*, on the contrary, it was the second syllable of TP-words that was prosodically marked which does correspond to the default pattern of lexical stress in Portuguese. Thus, primary stress effects may emerge in this condition, since participants may consider the TP words as paroxytones (i.e., as stressed on the penultimate syllable).

In Experiment 3A, using a three-alternative forced-choice task, we ensured that duration is actually considered by EP listeners as a strong correlate of primary word stress. Thus, any impact observed when TP-words were acoustically marked as paroxytones could only be due to the role of primary stress in speech segmentation.

In Experiment 3B, the weighting of TPs and primary word stress was evaluated within an ALL setting through three (between-participants) input-intelligibility conditions: in addition to the intact and mildly degraded (22dB SNR) conditions already used in Experiment 2, a *strongly degraded* (10 dB SNR\textsuperscript{13}) was used. Within each intelligibility condition, the same AL was presented in four (between-participants) conditions according to the location of the stressed syllable: in one condition stimuli were unstressed, and in the others they were stressed on the first, second, or third syllable.

\textsuperscript{13} The 10dB SNR corresponds to the superimposition of white noise with an intensity of 66 dB (see section 2.2.2. of the previous chapter).
Experiment 2. Domain-general vs. Speech-specific Universal

Cues in Segmentation

In the present experiment speech-specific universal prosody was added to the AL stream by acoustically marking prosodic right-edges by syllable lengthening and falling pitch, in agreement to what is observed in natural speech (e.g., Grice, 2006; Saffran et al., 1996b; Shukla et al., 2007).

Since the acoustic correlate of prosodic right-edges used could possibly be perceived as corresponding to the primary stress of TP-words, two procedures adopted in the present experiment allowed the independent evaluation of universal prosody (i.e., right-edge marking) and language-specific prosody (i.e., primary word stress). First, in the present experiment allowed the independent evaluation of universal prosody (i.e., right-edge marking) and language-specific prosody (i.e., primary word stress). First, as already explained and as illustrated in Table 3, two incongruent cues conditions were used, one in which TP-words were incongruent with both IP-edges and the default stress pattern in the listeners’ native language, and one in which TP-words were incongruent with IP-edges but presented the Portuguese default stress pattern (they were paroxytones).

In addition, the use of two types of trisyllabic part-words allowed evaluating whether in degraded conditions listeners would treat the acoustic markers (duration and F0) as correlates of either universal prosody or primary stress. These part-words that also occurred in the continuous stream during the familiarization phase were confronted with TP-words in the test phase (a two alternative forced-choice test). As illustrated in Table 3, part-words 23#1 (where “#” defines a statistical boundary)
were constituted by the two last syllables of one TP-word (i.e., 23) and the first syllable of the next TP-word (i.e., 1), while part-words 3#12 were constituted by the last syllable of one TP-word (i.e., 3) and the first two syllables of the next TP-word (i.e., 12). Thus, in any AL familiarization condition, during the forced-choice test, listeners were presented with two types of “prosodically cohesive” AL stimuli: one whose acoustically marked syllable was located at the right edge (i.e., marking a prosodic boundary), and the other whose acoustically marked syllable corresponded to the penultimate syllable thus obeying to the default word primary stress pattern in Portuguese (for a clear illustration of these types of AL stimuli, see Table 3).

Table 3: Orthographic translation of a sample of the stream heard in the familiarization phase in the four cues conditions and of the AL-stimuli (TP-words; Part-words 3#12; Part-words 23#1) of Experiment 2. The “#” defines word boundaries according to TPs; the “-” represents concatenation; the prosodic marked syllable is underlined and prosodic right-edges are marked by “]”.

<table>
<thead>
<tr>
<th>CUE CONDITIONS (available cues)</th>
<th>SPEECH STREAM (familiarization phase)</th>
<th>PART-WORDS 3#12</th>
<th>PART-WORDS 23#1</th>
</tr>
</thead>
<tbody>
<tr>
<td>single cue (TPs)</td>
<td>...-#bu-ka-la-#fu-fi-bu-#lu-fa-ba-#ki-la-bu-#...</td>
<td>ba-#ki-la</td>
<td>ka-la-#fu</td>
</tr>
<tr>
<td>congruent cues</td>
<td>...-#bu-ka-la-#fu-fi-bu-#lu-fa-ba-#ki-la-bu-#...</td>
<td>ba-#ki-la</td>
<td>ka-la-#fu</td>
</tr>
<tr>
<td>incongruent cues 1st syllable</td>
<td>...-#bu-ka-la-#fu-fi-bu-#lu-fa-ba-#ki-la-bu-#...</td>
<td>ba-#ki-la</td>
<td>ka-la-#fu</td>
</tr>
<tr>
<td>incongruent cues 2nd syllable</td>
<td>...-#bu-ka-la-#fu-fi-bu-#lu-fa-ba-#ki-la-bu-#...</td>
<td>ba-#ki-la</td>
<td>ka-la-#fu</td>
</tr>
</tbody>
</table>

3.2. Method

3.2.1. Participants

Ninety undergraduate psychology students at the University of Lisbon participated in the experiment for a course credit. All were monolingual European-
Portuguese speakers, with no reported history of speech or hearing disorders. Among them, 45 were randomly assigned to the intact speech condition (12 in the single cue, 11 in the congruent cues, 12 in the incongruent cues-1\textsuperscript{st} syl, and 10 in the incongruent cues-2\textsuperscript{nd} syl condition), and 45 were assigned to the physically degraded (22dB SNR) condition (9 to the single cue, 12 to the congruent cues, 12 to incongruent cues-1\textsuperscript{st} syllable and 12 to the incongruent cues-2\textsuperscript{nd} syllable condition).

3.2.2. Material

All speech stimuli were synthesized using text-to-speech MBROLA software (Dutoit et al., 1996) with an EP female diphone database (available at http://tcts.fpms.ac.be/synthesis/mbrola.html) at 22.05 kHz and with a speech rate, close to conversational level, of about 270 syllables per minute.

\textit{AL Material}

The AL used in the present experiment, as well as in the other two experiments of this chapter, was already described in Chapter Two (see section 2.2.2). Note that selection of the vowels was especially critical, since there is a close relation between vowel quality and lexical stress in EP (Mateus & d’Andrade, 2000). The three vowels that constituted the AL phonological repertoire (i.e., /\textit{u}/, /\textit{i}/, /\textit{u}/) in EP can occur in any position within a word (i.e., final and non-final, pre-stressed and post-stressed) and can also be either stressed or unstressed (Mateus & d’Andrade, 2000).

\textit{TP-words}. Three TP-words presented higher TPs (“\textit{high-TP-words}”, ranging from 0.75 to 1.00, i.e., /\textit{lu\textsuperscript{u}ebul}/, /\textit{fu\textsuperscript{u}ibu}/, /\textit{ki\textsuperscript{u}ebul}/) than the other three (“\textit{low-TP-}
words”, from 0.50 to 0.58, i.e., /bukulu/, /bubuku/, /kufubi/). This distributional gradient is probably similar to what happens in natural languages (Saffran et al., 1996b) and might allow a fine-grained evaluation of any effect of TPs (i.e., TP-gradient) in ALL.

**Part-words.** These AL-stimuli were constituted by syllables of two different TP-words that occurred adjacently in the speech stream during the familiarization phase, and corresponded to the part-words used in Experiment 1 (see chapter two). Three part-words (i.e., Part-words 3#12; e.g., /bul/) consisted of the last syllable of one TP-word (e.g., /b/ which is the last syllable of /lufb/ and the first two syllables of the next (e.g., /kil/ which are the first two syllables of /kilbu/). The other three part-words (i.e., Part-words 23#1; e.g., /fibulu/) consisted of the last two syllables of a TP-word (e.g., /fibu/ which are the last syllables of /fufibu/) and the first syllable of the next (e.g., /lu/ which is the first syllable of /lufb/). The AL-stimuli are presented in Appendix I.

The TPs of the AL-stimuli (TP-words and part-words) were computed by averaging the two TPs associated to each stimulus, with TPs between adjacent syllables always higher within than between TP-words (0.68 and 0.38, respectively).

**Familiarization phase.** Four synthesized versions of the AL were created. Each version included the same sequence of syllables, divided into three listening 7-minutes blocks (rendering 21-minutes). Each block was created by concatenating 105 tokens of each TP-word (1890 syllables, 630 tokens of words) with the only criterion that two tokens of the same TP-word never occurred adjacently in the stream.

The difference between the single cue version of the AL and the other three versions consisted on the number of segmentation cues available in the speech
stream during the familiarization phase. In the single cue version, only TPs between adjacent syllables could help listeners to locate word-boundaries, with no acoustic cues available, since the stream presented a flat 220Hz pitch and the average duration of all syllables was equivalent (i.e., 222 ms). In the other three versions, both statistical and prosodic informations were available. The prosodic information was added by acoustically marking one particular syllable of each TP-word through its lengthening on 150ms and linear decreasing its pitch (20Hz variation: from 220 Hz to 200 Hz), while the other syllables remained unchanged (i.e., on average a duration of 222 ms; flat 220Hz pitch).

The three conditions in which prosody and TPs were available differed in the congruency of the cues (see Table 3 for an illustration). In the congruent cues condition both statistical and prosodic information suggested the same word-boundaries, since the last syllable of each TP-word was acoustically marked defining a prosodic right-edge. In the two incongruent cues conditions, statistical and prosodic information suggested different segmentation hypotheses. In the incongruent cues-1\(^{st}\) syl condition, the first syllable of each TP-word was acoustically marked defining the prosodic right-edge. Thus, while TPs suggested that TP-words were the “lexical units” of the AL, they straddle a prosodic boundary, and prosodic information suggested that the part-words (straddling TP-boundaries) that end with the prosodically marked syllable were plausible words of the new language (in this cues condition, the part-words 23#1). In the incongruent cues-2\(^{nd}\) syl condition, the prosodically marked syllable corresponded to the second syllable of each TP-word, and thus also here the two segmentation cues were put against each other (part-words 3#12 were the prosodic cohesive units).
The *mildly degraded signal condition* was made by superimposing white noise at a 22dB SNR to each 7-minutes block of all versions of the AL using the same method of Experiment 1 (*see* section 2.2.2. of Chapter Two).

**Forced-choice test phase.** The three syllables that constituted each AL-stimulus (TP-words and part-words) were synthesized with the same average duration and 220Hz flat pitch (with no prosodic acoustic correlates available), and concatenated without white noise superimposed, for avoiding that participants could respond on the basis of any acoustic matching between the stimuli of familiarization and test phases.

The forced-choice test phase included 36 trials rendered by the exhaustive combination of all six TP-words and six part-words. In half of the trials, TP-words were confronted with part-words 3#12, and in the others they were confronted with part-words 23#1.

3.2.3. Procedure

Identical to the one of Experiment 1 (*see* section 2.2.3. of the previous chapter). Participants were tested individually or in groups of two in a sound-attenuated room, as done in both familiarization and test phased of Experiment 1. Importantly, as in the previous experiment, no information about the structure, phonology, length of the words or about the available prosodic cues was given. Participants in the mildly degraded signal conditions were warned of the poor signal quality.
3.3. Results and Discussion

First, we evaluated whether any TP-gradient effect was found on listeners' ALL performance. In the mixed ANOVA ran on raw TP-word choices with TP-level of AL-words (high vs. low) as within-subject factor, signal quality (intact vs. degraded) and cues (single cue; incongruent cues – 1\textsuperscript{st} syl; incongruent cues – 2\textsuperscript{nd} syl; congruent cues) as between-subject factors, only the main effect of cue condition was significant \([F(3, 82) = 25.1, \; p < .0001, \; MSe = 12,54; \; \eta^2 = .38]\). No other significant main effects or interaction was found \([Fs < 2, \; p > .10]\).

Next, we evaluated whether listeners' TP-word choices were influenced by the type of part-words to which TP-words were confronted with in the ALL test phase. In the mixed ANOVA with part-words type (part-words 3\#12; part-words 23\#1) as a within-subject factor, and signal quality and cue condition as between-subject factors, the main effect of cue condition was significant \([F(3, 82) = 25.1, \; p < .0001, \; MSe = 12,54; \; \eta^2 = .48]\) and was modulated by part-word type \([F(3, 82) = 10.0, \; p < .0001, \; MSe = 6,41; \; \eta^2 = .27]\). The three way interaction between all factors at study was also significant \([F(3, 82) = 3.4, \; p < .05, \; MSe = 6,41; \; \eta^2 = .12]\). No other effect was found \([Fs < 1]\).

Average ALL performances are presented in Table 4, as well as local one-sample \(t\)-test comparisons with chance-level.

In order to specifically analyze the weighting of prosodic and statistical information, we evaluated the effect of cue condition and part-words type separately in each input-intelligibility condition (see Figure 3).
With intact signal, the main effect of cue condition was significant \([F(3, 41) = 15.0, p < .0001, MSe = 12.77, \eta^2_p = .52]\). No main effect of part-word type was found \([F < 1]\), but the interaction between, this factor and cue condition was significant \([F(3, 43) = 8.0, p < .0005, MSe = 8.57, \eta^2_p = .37]\).

Table 4: Average proportions of TP-word responses (in percentage), separately for cues condition (single cue; incongruent cues 1st syl; incongruent cues 2nd Syl; and congruent cues) and signal quality condition (Intact vs. Degraded), considering Part-words type (3#12 vs. 23#1) in Experiment 2. Standard error of the mean for each condition is presented in parentheses as well as examples of the AL-material.

<table>
<thead>
<tr>
<th>SIGNAL QUALITY</th>
<th>Intact</th>
<th>Mildly Degraded</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Part-words</td>
<td>Part-words</td>
</tr>
<tr>
<td>Cues Condition</td>
<td>3#12</td>
<td>23#1</td>
</tr>
<tr>
<td></td>
<td>(ba#kila)</td>
<td>(fibu#lu)</td>
</tr>
<tr>
<td>Single Cue</td>
<td>64.4 (5.7)*</td>
<td>61.6 (4.7)*</td>
</tr>
<tr>
<td>(fufibu#lufaba#kilabu)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Congruent cues</td>
<td>88.4 (2.6)**</td>
<td>85.4 (3.8)**</td>
</tr>
<tr>
<td>(fufiBU#lufaBA#kilaBU)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incongruent cues 1st syl</td>
<td>58.8 (4.7)</td>
<td>37.5 (6.3)*</td>
</tr>
<tr>
<td>(FUfibu#LUfaba#Kilabu)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incongruent cues 2nd syl</td>
<td>47.8 (9.6)</td>
<td>74.4 (3.3)**</td>
</tr>
<tr>
<td>(fuFibu#LuFAba#KiLABu)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Local one-way t-tests (in comparison to chance level, i.e., 50%):

* \(p < .05\)

** at least \(p < .01\)

In both the single cue and the congruent cues condition, listeners were able to correctly choose TP-words as the lexical units of the new language, with overall above-chance performance levels \((t(11) = 2.9, p = .01 \text{ and } t(10) = 13.4, p < .0001, \text{ respectively})\). In both cases, performance was not affected by part-words type [both
Interestingly, the absence of a part-words type effect observed in the congruent cues condition suggests that listeners were not affected by primary stress. Had this been the case, we would have found worst performance on trials in which TP-words were confronted with part-words 23#1. With congruent cues, listeners performance was driven by the available statistical and prosodic cues, which enabled them to present the best performance in comparison to both the single cue \[F(1, 43) = 16.6, p < .0005\] and the incongruent cues conditions \[vs. \text{ incongruent cues-1}^{st}\text{ syl condition: } F(1, 43) = 43.6, p < .0001; \text{ vs. incongruent-cues 2}^{nd}\text{ syl condition: } F(1, 43) = 17.6, p < .0001\].

Figure 3: ALL performance pattern (proportion of AL-word responses, in percentage), in Experiment 2, broken-down by Part-word Type (3#12; 23#1), according to Input-Intelligibility (Intact; Degraded) and Cues available (single cue; incongruent cues first syllable; incongruent cues second syllable; congruent cues). Vertical bars denote standard error of the mean on each condition. Chance level corresponds to 50%.

In clear opposition with the good performances observed in the single cue and

108
congruent cue conditions, with intact signal overall performance was at chance in the two incongruent cues conditions \([1^{\text{st}} \text{syl condition: } t(11) = -.5, \ p > .5; \ 2^{\text{nd}} \text{syl condition: } t(9) = 1.8, \ p = .1]\). As clearly illustrated in Figure 3, in both incongruent conditions TP-words choices were affected by part-words type \([\text{incongruent-cues } 1^{\text{st}} \text{syl condition: } F(1, 43) = 10.3, \ p < .005; \ \text{incongruent-cues } 2^{\text{nd}} \text{syl condition: } F(1, 43) = 13.4, \ p < .01]\).

As a matter of fact, when the first syllable of TP-words was acoustically marked, performance was at chance when TP-words were confronted with part-words \([-12].\) Nevertheless, this level of performance did not differ from the one displayed in both the single cue \([F(1, 43) = .5, \ p = .50]\) and incongruent cues-2\(^{nd}\) syl \([F(1, 43) = 1.7, \ p = .20]\) conditions. However, when TP-words were confronted with part-words \([23\#1],\) listeners considered these part-words as the lexical units of the AL, discarding the TP-words (see Table 4). This performance level was significant below the one found in any one of the other cue conditions at study \([\text{vs. single cue: } F(1, 43) = 13.2, \ p < .001; \ \text{vs. congruent cues: } F(1, 43) = 50.1, \ p < .0001; \ \text{vs. incongruent cues-2}\^{\text{nd}} \text{syl: } F(1, 43) = 28.4, \ p < .0001]\). Thus, when the statistical segmentation outputs (i.e., TP-words) were confronted with the lower-TP units that were consistent with the universal prosodic cue (part-words \([23\#1] \text{ like ka-la-}\#fu],\) where \(\#\) signals the prosodic edge, see Table 3), the latter were considered as the correct items of the new language, demonstrating the preponderance of the universal prosodic cue over general-domain statistics, with the former being able to drive segmentation even in the presence of an incongruent cue.

In the other incongruent cues condition (when the second syllable of TP-words was acoustically marked), TP-words were chosen as being the AL units
significantly above chance (see Table 4) only when they were confronted to part-words that were not supported by any other available cue (i.e., to part-words 23#1 like ka-la-[#fu]). This performance tended to be higher than the one in the single cue condition \(F(1, 43) = 3.4, p = .07\). Importantly, when TP-words were confronted to AL stimuli that have lower TPs but form cohesive units according to IP edges (i.e., part-words 3#12 like ba-#ki-la), listeners were no longer able to choose TP-words as the “lexical units” of the new language, performing at chance (see Table 4), because in this case the universal prosodic cue defined a right-edge at the last syllable of these part-words.

With degraded signal (i.e., 22dB SNR), as already observed with intact speech, the main effect of cues condition \(F(3, 41) = 11.1, p < .0001, MSe = 12.31; \eta_p^2 = .45\) and its interaction with part-words type \(F(3, 41) = 3.3, p < .05, MSe = 4.24; \eta_p^2 = .20\) were significant, as was the case with intact speech. No main effect of part-words type was found \(F(1, 41) = 2.4, p > .15\).

As can be seen in Figure 3, the performances of listeners exposed to the single cue and to the incongruent cues-1st syllable condition were similar with physically degraded signal as with intact signal. Indeed, in the single cue condition, overall performance was above chance \((t(11) = 2.5, p < .05)\), with listeners choosing more often the TP-words than the part-words, independently of part-word type \(F < 1\). In fact, statistical information was able to drive the segmentation process at a similar level with degraded and intact speech \(F < 1\), confirming its resilience to physical noise (cf. Fernandes et al., 2007). In the incongruent cues-1st syl condition, listeners had the worst performance, in comparison to both the single cue \(F(1, 41) = 5.9, p = .01\) and the congruent cues \(F(1, 41) = 30.7, p < .0001\) conditions, and even
to the other incongruent cues condition [$F(1, 41) = 15.0, p < .0005$]. They presented an overall performance almost below chance ($t(11) = -2.0, p = .065$), which was mainly due to listeners’ preference for part-words 23#1 over TP-words (see Table 4). Thus, confronted with TP-words and part-words that were prosodically cohesive according to IP edges (part-words 23#1), the latter units were considered as the plausible “words” of the AL. Thus, even when the signal is physically degraded, the universal prosodic cue is still able to drive segmentation processing, overruling the incongruent statistical cue at a similar level to the one already observed with phonetically intact speech [$F < 1$].

In opposition to what was observed with intact speech, when the signal was physically degraded performance with congruent cues was modulated by part-word type [$F(1, 41) = 8.3, p < .01$]. Indeed, listeners were more proficient on TP-words selection when TP-words were confronted to part-words 3#12 than when they were confronted to part-words 23#1. The proportion of TP-words chosen as the AL units over part-words 23#1 (which were acoustically marked on the penultimate syllable – here, second –, obeying to the default lexical stress pattern in EP) was lower than the one found for the same condition with intact speech [$F(1, 41) = 5.9, p = .01$], and did not differ from the one found in the single cue condition [$F(1, 41) = 1.7, p > .10$]. Thus, in this particular case, although the integration of statistical and universal prosodic cues was still able to drive the segmentation process, the confrontation between TP-words (outputs of this integration) and AL units with the default Portuguese stress pattern (part-words 23#1) induced a performance cost. This was not due to a general inability of congruent cues to promote a redundancy gain, since when confronted with part-words that were not supported by any cue (i.e., part-words
3#12, see Table 4), TP-words led to a well above-chance performance that was significantly better than in the single cue condition \([F(1, 41) = 8.9, p < .005]\), being in fact equivalent with degraded and intact signal \([F < 1]\).

In the incongruent cues-2\textsuperscript{nd} syl condition, in which TP-words were acoustically marked on the penultimate syllable, which is incompatible with a universal prosodic edge but compatible with the default stress pattern in Portuguese, the performance pattern differed sharply from the one found with intact signal. Indeed, overall performance was above chance \((t(11) = 2.9, p = .01)\), was not affected by part-words type \([F < 1]\), and was similar to the one found in the single cue condition \([F = 1.4, p > .10]\). Furthermore, preference for TP-words over the part-words 3#12 (which were marked by a prosodic edge) was stronger here than when the signal was intact \([F(1, 42) = 4.5, p < .05]\), revealing that lexical stress was possibly assisting TP segmentation when the signal was impoverished and hence that the conjunction of these two cues enabled paroxytone TP-words to be extracted from the stream even when confronted with stimuli supported by a universal prosodic cue (part-words 3#12).

In agreement with Shukla et al.’s (2007) conclusions, the present results thus corroborate two facts. First, universal prosodic cues are more preponderant in speech segmentation than general-domain TPs. Second, listeners are sensitive to different (and even incompatible) segmentation byproducts, with their ALL performance being largely dependent on the type of test-stimuli presented to them (see Shukla et al., 2007; Experiment 4). This is an important result since it demonstrates that when different segmentation cues suggest incompatible “lexical units”, listeners do not simply adopt an all-or-none procedure, filtering out from the beginning of
segmentation processing the lower weighted cues’ outputs. Instead, although listeners are able to select from two incompatible outputs (based on different cues) the one supported by the more reliable source of information, this seems to occur in a graded fashion.

The present results also add to previous findings. In the present experiment, it was demonstrated that universal prosody is as resilient to physical noise as the general-domain statistical cue constituted by TP, with the former maintaining its preponderance over the latter even in physically degraded conditions. The present results also converge with Mattys’ data (2004; Mattys et al., 2005) by showing that, in Portuguese, primary stress pattern effects emerge only in degraded listening conditions.

Since one of the acoustic correlates of prosody that was available in the AL stream used in the present experiment could be either associated with universal prosodic right-edges or with primary stress pattern, in Experiment 3A, we checked whether Portuguese listeners treat in fact syllable duration as an acoustic correlate of primary stress.
Experiment 3A: Duration is an acoustic correlate of word
primary stress in EP.

The present experiment was in fact a post-test of Experiment 2 and a pre-test of Experiment 3B on the “psychological reality” of syllable duration as an acoustical correlate of primary stress in Portuguese. Native EP listeners were presented with trisyllabic AL stimuli similar to the ones used in the test phase of Experiment 2 (six TP-words and six part-words) either in a stressed condition, in which duration marked the location of stress within AL stimuli, or in a control, unstressed condition, in which no acoustic correlate of primary stress was available. Participants were not informed about the presence or absence of an acoustic correlate to stress, and all had to perform an off-line three-alternative forced-choice task: they were required to locate the stressed syllable of the stimulus, by choosing among the first (antepenultimate), second (penultimate), or third (last) syllable.

3.4. Method

3.4.1. Participants

Thirty-eight undergraduate psychology students at the University of Lisbon participated in the experiment for a course credit. All were monolingual European-Portuguese speakers, with no reported history of speech or hearing disorders, and
none of them participated in the previous experiment. Half were assigned to the *unstressed condition*, and the others to the *stressed condition*.

### 3.4.2. Material

All trisyllabic AL-stimuli (the six TP-words and six part-words, *see* Appendix I) used in Experiment 1 and 2 were synthesized using the same method of Experiment 2 (*see* section 3.2.2 of this chapter).

Four exemplars of each AL-stimuli were created according to primary stress location (i.e., no stress; stress on the 1<sup>st</sup>; 2<sup>nd</sup>; and 3<sup>rd</sup> syllable). The unstressed version was constituted by the AL-material used on the test-phase of Experiment 2. Stress was acoustically marked by lengthening on 100ms the “stressed” syllable and reducing on 50ms each unstressed syllable, while pitch remained flat (220 Hz). Unstressed syllables were reduced because in EP vocalic reduction is a prominent phenomenon, while stressed syllables are lengthened, unstressed vowels are, in fluent speech, reduced and in most cases completely eliminated from the stream (Mateus & d’Andrade, 2000).

One list with all 12 unstressed AL-stimuli (6 TP-words and 6 part-words) was created, which was only presented to participants in the unstressed condition (i.e., control group). The 36 stressed AL stimuli were distributed across participants through three lists, each one including both TP-words and part-words stressed in one of the three possible syllables. The three stressed lists had the same proportion of TP-words and part-words with each lexical stress pattern, but each AL stimulus presented a different stress pattern in each one of the three lists.
3.4.3. Procedure

Participants were tested individually or in groups of two. Presentation, timing and data collection was controlled by *E-Prime 1.1* (Schneider et al. 2002a; 2002b).

All participants performed an off-line three-alternative forced-choice task. Participants in the *unstressed condition* were only presented with the unstressed items and participants in the *stressed condition* were randomly assigned to one of the three stressed lists (6 participants in list 1 and 2 and 7 in list 3). All participants were informed that on each trial they would hear through headphones a warning tone immediately followed by a trisyllabic pseudoword. Their task was to identify the stressed syllable of the stimulus, presented in each trial, by pressing the 1, 2, or 3 key on the keyboard corresponding to the first (antepenultimate), second (penultimate), or third (last) syllable, respectively. Participants were not informed about either the presence/absence or the type of acoustic correlate used, and response accuracy was emphasized. Immediately after participants gave their answer, another trial began. If no answer was registered after a maximum of 10 seconds, the next trial was presented. Three practice trials were provided, prior to the test, in order to clarify its structure and enable practice with key presses. On practice trials, the stimuli presented were real words with primary stress located on the first (i.e., /ˈpenɪk/ *panic*), second (i.e., /bəˈnaːnə/ *banana*) or third (i.e., /ˈwɪld boʊər/ *wild boar*) syllable and feedback was provided only on these trials.

Order of presentation of test trials was pseudo-randomized for each participant with the only criterion that two stimuli with the same stress pattern did not occur immediately after one another.
3.5. Results and Discussion

Listeners’ choices in the unstressed condition and correct responses in the stressed condition are presented in Figure 4.

Performance in the unstressed condition was compared with chance level (i.e., 33.33%). Overall performance was at chance: on average 33.3% (SE = 2.7), which was expected in a situation in which no acoustic correlates of stress was available. However, closer inspection of the data shows that while listeners’ choices of the stressed syllable was at chance for the first and the third syllable responses [on average, 28% (SE = 3.0): \( t(18) = -1.3, p > .10 \); and 27% (SE = 3.0): \( t(18) = -1.7, p = .10 \), respectively], listeners considered in 45% of the trials (SE = 3.0) that AL stimuli were stressed on the second syllable. This above-chance performance \( (t(18) = 4.2, p < .001) \) is probably due to the fact that this response corresponds to the default stress pattern in their native language.

Figure 4: Listeners’ proportion of responses on the three-alternative forced choice stress location task on AL-stimuli, broken-down by syllable detection responses (1st syl; 2nd syl; 3rd syl) in the two acoustic cue conditions (Unstressed vs. Stressed) in Experiment 3A. Note that only the performance of the stressed group corresponds to correct detections, since in the unstressed group no syllable was acoustically marked. Vertical bars denote standard error of the mean. Chance level corresponds to 33.3%.
In the stressed condition, listeners were able to correctly locate the stressed syllable of each AL stimulus, performing above chance for all stress locations (see Figure 2), with an average proportion of correct responses of 51% \([SE = 2.8] t(18) = 2.8, p = .01\) for stimuli with stress located on the first syllable, 63% \([SE = 2.7] t(18) = 4.9, p < .001\) for stress on the second syllable, and 43% \([SE = 2.5] t(18) = 2.5, p < .05\) for stress on the last syllable.

The mixed ANOVA ran on the proportion of responses in the unstressed condition and on the proportion of correct responses in the stressed condition, with group (unstressed vs. stressed) as a between-participants factor and located stress (1\textsuperscript{st}; 2\textsuperscript{nd}; 3\textsuperscript{rd} syllable) as a within-participants factor, revealed a significant main effect of group \([F(1, 36) = 26.2, p < .001; MSe = 0.04; \eta_p^2 = .42]\), since listeners in the stressed condition presented a high proportion of (correct) detections than the proportion of responses found in the unstressed condition. This result demonstrates that EP listeners are indeed sensitive to syllables duration as an acoustic correlate of primary stress.

The main effect of stress location was also significant \([F(2, 72) = , p < .001; MSe = 0.04, \eta_p^2 = .19]\), with higher proportion of responses to the second syllable than to any one of the other two \([vs. 1\textsuperscript{st} syllable: F(1, 36) = 10.0, p < .005; vs. 3\textsuperscript{rd} syllable: F(1, 36) = 16.2, p < .001]\), which did not differ from one another \([F < 1]\). No interaction between the two factors was found \([F < 1]\).

The present results allowed attesting that syllable duration is in EP a correlate of primary stress. Although listeners were presented with impoverished synthesized speech and with stimuli that were not words in their native language, they were still
quite able to correctly locate the stress of trisyllabic stimuli when the stressed syllable was lengthened in comparison to the unstressed syllables. Listeners who were presented with stimuli that included no acoustic marker of primary stress often considered that these stimuli were paroxytones, but not as often as those that were presented with lengthened stressed syllables, and their first- and second-syllable responses were at chance.

Besides, the EP listeners’ tendency to consider acoustically neutral stimuli as paroxytones is by itself fascinating. We cannot ensure whether this effect corresponds to a decisional bias or to a perceptual “stress illusion” (see Dupoux, Kakehi, Hirose, Pallier, & Mehler, 1999), but it clearly illustrates the fact that suprasegmental properties of the native language affects processing of unfamiliar items (e.g., Dupoux, Pallier, Sebastian-Gallés, & Mehler, 1997).

Porlex Psycholinguistic Database (Gomes & Castro, 2003)

Inspection of EP Porlex database (Gomes & Castro, 2003) reveals that from the 29238 entries, about 98% (i.e., 28 859) are words with two or more syllables with 59% of them being paroxytones (i.e., stressed on the penultimate syllable). Considering only words with the same phonological structure as the AL-stimuli used in the present study (i.e., CV.CV.CV; which corresponds to about 14% of the entries), the predominance of words with stress on the penultimate syllable is even more clearer (i.e., 82%), while only 16% and 2% corresponds to proparoxytones and oxytones, respectively. Taking these statistical facts into account, stressed syllables in Portuguese likely correspond to the penultimate syllable of a word, which in the case of trisyllables correspond to the second syllable. Against what is observed in
English (e.g., Cutler et al., 1997; Mattys, 2004), in Portuguese the stressed syllable does not define a possible word onset (nor a word ending), which would seem to reduce the role of lexical stress in speech segmentation. However, the results found in Experiment 2 suggest that the lexical stress pattern helps speech segmentation in degraded listening conditions. This possibility was further evaluated in Experiment 3B, which used three input-intelligibility conditions, with a *strongly degraded condition* (10dB SNR; the strongly degraded condition of Experiment 1, *see* chapter two) in addition to the intact and mildly degraded (22dB SNR) conditions already used in Experiment 2.
Experiment 3B: Domain-general vs. Language-specific Cues in Segmentation.

According to the results of Experiment 2 and to previous findings (e.g., Mattys et al., 2005; Valliant & Levitt, 1996), since lexical stress effects only emerge in degraded conditions, it is possible that when a resilient, general-domain, statistical cue like TPs, is still available in the stream, the impact of the former (which is a speech-specific cue) will be to restrain the units extracted from the stream to the ones supported by all available sources of information. Shukla et al. (2007) already proposed that the role of universal prosodic cues would be acting as a filter of the potential words set forth by general-domain statistical learning. If this were holding true for language-specific prosodic cues, it would be possible that lexical stress will also filter out the statistical units that are not compatible with it when the signal is physically degraded. However, on the basis of previous findings revealing that lexical stress is not able to drive the segmentation process when other cues are still available in the stream (Mattys et al., 2005; Valiant & Levitt, 1996), it is possible that lexical stress by itself is not able to drive speech segmentation, being unable to overrule the statistical cue.

The present experiment evaluated this possibility, namely the relative weighting of TPs and primary word stress, through three input-intelligibility conditions (intact speech, and mildly and strongly degraded listening conditions), using four between-participants conditions according to the location of the stressed syllable. In the TP–unstressed condition, only statistical information was available in
the stream; it actually corresponded to the single cue version of Experiment 1; in the 1st syl-stressed condition, the first syllable of some exemplars of TP-words was lengthened by 100 ms while the other (unstressed) syllables of that TP-words were reduced by 50 ms each; in the 2nd syl-stressed condition, stress was located on the penultimate (second) syllable of TP-words (which corresponds to the Portuguese lexical stress default pattern); and in the 3rd syl-stressed condition, stress was located on the last syllable of the TP-words.

If the role of primary stress was similar to the one of a universal prosodic cue we would expect to find a pattern of results similar to the one observed in Experiment 1, namely, a performance gain in the condition in which the prosodic and the statistical segmentation outputs are compatible. However, as already mentioned according to the results of Experiment 1 and to previous findings (e.g., Mattys et al., 2005; Valliant & Levitt, 1996), lexical stress effects are particularly observed in degraded conditions. Thus, we would find an impact of primary stress only in degraded conditions, with its effect being maximized in the strongly degraded condition at study. Most likely, it will be in that situation only that listeners will parse statistical cohesive units obeying to the default stress pattern in their native language, namely TP-words stressed on their 2nd syllable.

3.6. Method

3.6.1. Participants

One hundred and twenty-eight undergraduate psychology students at the University of Lisbon participated in the experiment for a course credit. All were
monolingual EP speakers, with no reported history of speech or hearing disorders. Among them, 43 were randomly assigned to the intact speech condition (12 to the TP–unstressed, 11 to the 1st syl-stressed, 7 to 2nd syl-stressed, and 13 to the 3rd syl-stressed condition), 47 were assigned to the mildly noise (22dB SNR) condition (9 to TP–unstressed, 11 to 1st syl-stressed, 14 to 2nd syl-stressed, and 13 to 3rd syl-stressed condition), and 38 to the strongly noise (10db SNR) condition (9 to TP–unstressed, 10 to 1st syl-stressed, 10 to 2nd syl-stressed, and 9 to 3rd syl-stressed condition).

3.6.2. Material

All speech stimuli were synthesized using the same method of the previous experiments of the present chapter. The same AL of Experiment 1 was also used.

For the familiarization phase, four synthesized versions of the AL were created. Each version included the same sequence of syllables, divided into three 7-minutes listening blocks (rendering 21-minutes) as in Experiment 2.

TP-unstressed of the AL version corresponded to the single cue version of Experiment 2. In the three stressed versions of the AL, the acoustic correlate of lexical stress that was manipulated was the duration of the AL syllables. Stressed syllables were lengthened by 100 ms while unstressed ones were reduced each by 50 ms. However, the manipulation of acoustic correlate of lexical stress (i.e., duration of the AL-syllables) here adopted was similar to the one with coarticulatory cues presented in Experiment 1 of Chapter Two (cf. Fernandes et al., 2007). Only some (approximately one third of the exemplars) of the TP-words within the AL stream presented a stress pattern. These stressed TP-words were located as close as possible in the three stressed conditions of the AL, the only difference between them being
the location of the stressed syllable: for the 1st syl-stressed condition, the lengthened syllable was the first; for the 2nd syl-stressed condition, it was the second; and for the 3rd syl-stressed condition, it was the last syllable. For illustration, an orthographic translation of a sample of the speech stream in each version is presented in Table 5.

The two degraded signal conditions with white noise superimposed at 22dB SNR (i.e., mildly degraded condition) and 10dB SNR (i.e., strongly degraded condition) were created as in Experiment 1 and for the forced-choice test the same material of Experiment 2 was used (see previous section 3.3.2.).

<table>
<thead>
<tr>
<th>STRESS LOCATION CONDITIONS (available cues)</th>
<th>SPEECH STREAM</th>
<th>PART-WORDS</th>
<th>PART-WORDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP – unstressed</td>
<td>...#fu-ba-#ki-la-bu-#ka-fu-bi-#ba-bu-ku-#bu-ka-la-#fu-fi-bu-#…</td>
<td>ba-#ki-la</td>
<td>ka-la-#fu</td>
</tr>
<tr>
<td>1st syl Stressed</td>
<td>...#fu-ba-#KI-la-bu-#ka-fu-bi-#ba-bu-ku-#bu-ka-la-#FU-fi-bu-#…</td>
<td>ba-#KI-la</td>
<td>ka-la-#FU</td>
</tr>
<tr>
<td>2nd syl Stressed</td>
<td>...#fu-ba-#KI-#bu-#ka-fu-bi-#ba-bu-ku-#bu-#KA-la-#fu-fi-bu-#…</td>
<td>ba-#KI-#LA</td>
<td>KA-la-#fu</td>
</tr>
<tr>
<td>3rd syl Stressed</td>
<td>...#fu-ba-#KA-la-bu-#ka-fu-bi-#ba-bu-ku-#bu-#KA-la-#fu-fi-bu-#…</td>
<td>BA-#ki-la</td>
<td>ka-#LA#fu</td>
</tr>
</tbody>
</table>

3.6.3. Procedure

Identical to the one of Experiment 2 (see previous section 3.2.3.).
3.7. Results and Discussion

In the mixed ANOVA ran on participants’ TP-word choices with part-word type (part-words 3#12, part-words 23#1) as within-subject factor, signal quality (intact, 22db SNR, 10dB SNR) and lexical stress pattern (TP–unstressed, 1\textsuperscript{st} syl-stressed, 2\textsuperscript{nd} syl-stressed, 3\textsuperscript{rd} syl-stressed) as between-subjects factors, neither the main effect of lexical stress pattern nor its interaction with the other factors at study were significant [$F_{s} < 2.1, p > .10$]. The only significant effect found was the main effect of signal quality [$F(2, 116) = 5.9, p < .005, MSe = 8,72; \eta_{p}^{2} = .09$], with ALL performance declining with signal degradation [$F(1, 116) = 11.7, p < .001; MSe = 8,72, \eta_{p}^{2} = .09$]. Thus, contrary to what was found in contrast to what was found in Experiment 2, when the prosodic information was not a universal cue but a language-specific one, as was the case in the present experiment, no impact on listeners’ TP-words choices was found. This is in accordance with previous findings on the unreliability of prosody in speech segmentation when other cues are still available (e.g., Vallian & Levitt, 1996).

We next checked whether the ALL performance of listeners was affected by TP-level of TP-words (see Figure 5).

According to the mixed ANOVA with TP-words type (high-; low-TP-words) as within-subject factor, and signal quality, as well as lexical stress pattern as between-subject factors, the main effect of signal-quality was significant [$F(1, 116) = 6.1, p < .005, MSe = 8,74; \eta_{p}^{2} = .09$], as was the main effect of TP-words type [$F(1, 116) = 6.3, p < .01, MSe = 5,34; \eta_{p}^{2} = .05$]. Indeed, overall listeners had better performances for high- than for low-TP words. However, this TP-gradient effect was
modulated by signal-quality \[ F(2, 116) = 3.5, \ p < .05, \ MSe = 5.34; \ \eta^2_p = .06].

Importantly, the three way interaction was also significant \[ F(6, 116) = 3.0, \ p < .01, \ MSe = 5.34, \ \eta^2_p = .14\]. No other significant effect was found \[ Fs < 1\].

Figure 5: Performance pattern (proportion of AL-word responses, in percentage) broken-down by TP-level (High-TP; Low-TP) of AL-words, according to Signal condition (Intact; mildly degraded - 22dB SNR; strongly degraded - 10dB SNR) and Lexical Stress location (TP – No stress; 1st Syllable; 2nd Syllable; 3rd Syllable) in Experiment 3B. Vertical bars denote standard error of the mean on each condition. Chance level corresponds to 50%.

Average ALL performance broken down by TP-level is presented in Table 6.

In order to specifically analyze the TP-gradient effect on each stress pattern, we evaluated separately each input-intelligibility condition at study (see Figure 5).

With intact signal no significant effect was found [all \( Fs < 1.5, \ ps > .10\)]. All groups performed at a similar level, with no differences for low- and high-TP-words, independently of the absence/presence of an acoustic marker of primary stress and of
its location in TP-words (see Table 6 and Figure 5). All groups presented an above-chance overall performance [TP–unstressed: \( t(11) = 3.5, p < .005 \); 1st syl-stressed: \( t(10) = 2.3, p < .05 \); 2nd syl-stressed: \( t(6) = 2.0, p = .05 \); 3rd syl-stressed: \( t(12) = 3.9, p < .005 \)], choosing significantly more often the TP-words than the part-words as the “lexical units” of the new language. Thus, with intact signal, listeners were able to use statistical information to correctly parse the TP-words from the speech stream, independently of the stress pattern of these parsed stimuli.

Table 6: Average proportions of TP-Word responses (in percentage) in Experiment 2B, separately for each input-intelligibility condition (intact; mildly degraded; strongly degraded) and lexical stress location (TP-unstressed; 1st syllable; 2nd syllable; 3rd syllable), considering the TP-level of the AL-words (high- vs. low-TP). Standard error of the mean for each condition is presented in parentheses.

<table>
<thead>
<tr>
<th>Lexical stress location</th>
<th>Intact</th>
<th>Mildly degraded</th>
<th>Strongly degraded</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High-TP</td>
<td>Low-TP</td>
<td>High-TP</td>
<td>Low-TP</td>
</tr>
<tr>
<td>TP unstressed</td>
<td>67.6</td>
<td>58.4</td>
<td>60.5</td>
<td>59.2</td>
</tr>
<tr>
<td></td>
<td>(4.17)</td>
<td>(4.31)</td>
<td>(4.82)</td>
<td>(4.98)</td>
</tr>
<tr>
<td>1st syllable</td>
<td>61.1</td>
<td>57.1</td>
<td>59.6</td>
<td>43.9</td>
</tr>
<tr>
<td></td>
<td>(4.36)</td>
<td>(4.51)</td>
<td>(4.36)</td>
<td>(4.51)</td>
</tr>
<tr>
<td>2nd syllable</td>
<td>59.5</td>
<td>66.7</td>
<td>66.3</td>
<td>51.6</td>
</tr>
<tr>
<td></td>
<td>(5.47)</td>
<td>(5.65)</td>
<td>(3.87)</td>
<td>(3.99)</td>
</tr>
<tr>
<td>3rd syllable</td>
<td>60.7</td>
<td>63.2</td>
<td>63.7</td>
<td>55.1</td>
</tr>
<tr>
<td></td>
<td>(4.01)</td>
<td>(4.15)</td>
<td>(4.01)</td>
<td>(4.15)</td>
</tr>
<tr>
<td>Average</td>
<td>62.2</td>
<td>61.3</td>
<td>62.5</td>
<td>52.5</td>
</tr>
</tbody>
</table>

In the mildly degraded (i.e., 22dB SNR) condition, the main effect of TP-level of AL-words was significant [\( F(1, 43) = 12.8, p < .001, MSe = 5.83, \eta_p^2 = .23 \)]. No main effect of stress pattern [\( F(3, 43) = 1.3, p > .25 \)] was found. The interaction
between the two factors did not reached the conventional level of significance \(F(3, 43) = 1.9, p = .10, MSe = 5.42, \eta^2_p = .12\) but the size of this effect was moderated (Cohen, 1988). In fact, in line with the previous findings of Experiment 1 (see chapter one; cf. Fernandes et al., 2007), and as illustrated in Figure 5, when only statistical information was available in the stream, listeners chose both high- and low-TP-words above chance as the “words” of the new language \(t(8) = 2.2, p < .05; t(8) = 2.1, p < .05, \text{respectively}\), with no differences in the extraction of high- and low-TP-words \(F < 1\). In sharp contrast, for listeners exposed to the AL with the presence of an acoustic marker of primary stress, the TP-gradient effect was significant \(F(1, 43) = 17.5, p < .005\), reflecting the fact that for those listeners, only high-TP-words were extracted above chance [on average on 63% of the trials (SE = 2.0); \(t(37) = 6.0, p < .0005\)]. Performance for low-TP-words did not differ from chance [on average, 52% (SE = 1.9); \(t(37) = .3, p > .10\)], and was significantly poorer than in the TP– unstressed condition \(F(1, 46) = 4.0, p = .05\). This pattern of results suggests that although listeners performance is not globally affected by the presence of lexical stress patterns (no main effect of stress pattern was found) lexical stress plays some role in speech segmentation when the signal is mildly degraded, possibly constraining the extraction of statistical segmentation outputs to the ones with higher support. When lexical stress was available in the stream, independently of the stress pattern (see Table 6), only high-TP-words were correctly selected above chance as the plausible words of the AL. This is not surprising, since although by default in Portuguese stress is located at the penultimate syllable, the other two stress patterns are also legal in that language.

In the strongly degraded condition (i.e., 10dB SNR), the modulator impact of
lexical stress on the extraction of statistical outputs is even clearer, as demonstrated by the significant interaction between stress pattern and TP-level \( F(3, 34) = 5.7, p < .005, MSe = 4.069; \eta^2_p = .33 \), while no significant main effect was found \( Fs < 1.2, p > .10 \). Listeners exposed to only statistical cues available were able to correctly extract above chance level both high- and low-TP-words \( t(8) = 1.6, p < .05; t(8) = 3.6, p < .01 \), respectively; see Table 6] at similar performance levels \( F(1, 34) = 1.1, p > .25 \). Furthermore, as already observed in the mildly degraded condition, it was only in the single cue condition (i.e., TP-unstressed), low-TP-words were still correctly extracted from the stream. When stress cues were available, listeners were not able to extract the low-TP-words, independently of the TP-word stress pattern \( t(28) = -.5, p > .10 \).

With this strongly degraded signal, listeners exposed to a primary stress pattern diverging from the default one in their native language (i.e., to 1st and 3rd syl stressed conditions) performed overall at chance [on average, 48.6% (SE = 3.67) and 53.7% (SE = 3.87), respectively; \( ts < 1, ps > .10 \)]. In contrast, listeners exposed to AL with lexical stress on the second syllable (i.e., the default lexical stress pattern in EP) were still able to correctly choose the high-TP-words on 59.4% (SE = 4.58) of the trials \( t(9) = 2.4, p < .05 \), a performance level similar to the one found when only statistical information was available \( F = 1 \). However, in the 2nd syl stressed condition, low-TP-words were not extracted from the stream, which promoted the observation of a TP-gradient effect \( F(1, 34) = 12.6, p =.001 \) due to the exclusively extraction of high-TP-words.

The availability of a lexical stress pattern corresponding to the default stress pattern in the native language of the listeners did not promote any global benefit in
ALL, instead it narrowed the selection of statistical segmentation outputs to the ones with stronger support from the conjunction of the two segmentation cues available in the stream, enabling the exclusive extraction of these strongly supported units, namely of high-TP-words obeying to the default stress pattern in Portuguese, which were extracted as efficiently when the signal was quite distorted (i.e., 10dB SNR) as when it was intact ($F < 1, p \approx 1$).

The present results suggest that lexical stress effects in speech segmentation only emerge when listeners have to process a physically degraded signal, and that this holds true even if another resilient cue is also available in the stream (i.e., TPs computation). Since statistical information is very resistant to signal degradation, the ALL impairment observed in conditions in which stress cues were also available can only be attributed to an impact of language-specific prosodic information on segmentation. In fact, in stressed conditions, while with phonetically intact speech, both high- and low-TP-words were correctly extracted from the stream at a level similar to the one found when only statistical information was available in the stream, with mildly degraded signal, a lexical stress effect emerged, narrowing the extraction of AL units: low-TP-words, which with intact speech were correctly parsed from the stream, were no longer selected as possible words of the new language in the mildly degraded condition, resulting in much poorer performance than with intact speech [intact vs. 22 dB: $F(1, 116) = 10.7, p = .001$].

The strong degradation (i.e., 10dB SNR) maximized the strength of primary stress effects and hence only statistical outputs that obeyed to the default stress pattern in Portuguese were selected as plausible words of the new language. Consequently, statistical learning was inhibited in stressed conditions that did not
obey to the default pattern. For both 1st and 3rd syl-stressed conditions, while with intact signal both low- and high-TP-words were correctly chosen as the units of the AL, with mildly degraded signal only those with high-TPs were still extracted, and with the strongly degraded signal no statistical outputs were extracted at all. This progressive degradation of performance is supported by the linear trend of noise found in the ALL performance of listeners in these conditions \([F(1, 116) = 6.9, p < .01]\).

This pattern of results is not the one expected if we conceive that the role of prosodic cues in speech segmentation is one of filtering out statistical byproducts. Had this been the case, we would have found an overall above-chance performance with no TP-gradient effect for listeners familiarized with an AL in which TP-words were stressed on the penultimate syllable, since all statistical segmentation outputs (both high- and low-TP-words) were congruent with listeners’ native language default stress pattern. Instead, only statistical outputs with the strongest evidence (i.e., high-TP-words, obeying to the default stress pattern) were extracted from the stream. Indeed, with strongly impoverished signal, when stress and statistical cues were available in the stream, only listeners in the 2nd syl-stressed condition had an above-chance performance, and exclusively for high-TP-words, and this performance was as good with the degraded signal as with the intact one.

### 3.8. General Discussion

Recent work within the speech segmentation field (e.g., Mattys et al., 2005; Shukla et al., 2007) has been devoted to the study of the integration of different
sources of information, in accordance with computational simulations (e.g., Christiansen & Curtin, 2005) and behavioral data (e.g., Mattys et al., 2005) demonstrating that speech segmentation does not correspond to the sum of the independent byproducts of different available segmentation cues. Instead, as proposed by Mattys and colleagues (Mattys, 2004; Mattys et al., 2005; Mattys & Melhorn, 2007) segmentation cues are differentially weighted and speech segmentation is largely the product of the available sources of information and of the listening conditions.

Fernandes et al. (2007; see section 2.4. of chapter two) also proposed that the weighting of the available sources of information might also depend on two aspects: their domain generality and their particular role in language processing. First, both domain-general cues, like TPs, and universal speech-specific cues, like prosodic contours, are used since the very beginning of language acquisition (e.g., Dehaene-Lambertz & Houston, 1998; Thiessen & Saffran, 2003), thus occupying a fundamental role. However, speech-specific cues are probably processed in an encapsulated manner as part of the modularity speech perception system and hence are highly independent of “horizontal faculties” (Fodor, 1983), which may confer them a higher weighting in speech segmentation than general-domain cues. Second, within the speech-domain, while universal cues (such as intonation prosodic boundaries) are physiologically based (Grice, 2006; Shukla et al., 2007; Werner & Keller) and thus available in any natural language, language-specific cues depend on the particular language and hence intervene latter in speech segmentation. While universal cues could maintain an important role in sublexically-driven speech segmentation processing, the weighting of language-specific cues probably depends
on their word-boundary predictability in a specific language (cf. Mattys et al., 2005). Therefore, cross-linguistic differences would be particularly observed on the weighting of these language-specific cues.

In the present chapter we used two ALL settings to investigate the impact of physical noise on the weighting of these three types of sublexical cues in speech segmentation. In Experiment 2 we investigated the weighting of domain general TPs computation and universal prosodic cue by marking acoustic edges (through syllable lengthening and falling pitch) in the AL stream. Prosody was not only as resilient to physical degradation of the signal as the domain general TPs computation, but it was also a powerful segmentation cue, able to drive the segmentation processes even when TPs were incongruently available in the stream, and this held true in any listening condition. When TPs and prosodic right-edges suggested the same segmentation hypotheses a performance gain was observed.

This result adds to Shukla et al.’s (2007) proposal on the predominance of universal prosodic information over the statistical learning mechanism. While both general domain and speech-specific universal cues are resilient to physical noise on the grounds of their fundamental and precocious role in segmentation (Dehaene-Lambertz & Houston, 1998; Saffran et al., 1996a), language speech-specific cues are probably processed in an encapsulated manner (Fodor, 1983; 2000), being primarily considered and hence occupying a preponderant role over general-domain statistical cues.

Remarkably, in physically degraded conditions, a primary stress effect also emerged, assisting statistical learning. On the one hand, listeners exposed to the AL with TP-words acoustically marked on their penultimate syllable (corresponding to
Chapter 3. Universal and Speech-specific cues in ALL

he default stress pattern in their native language) considered these stimuli the “lexical units” of the AL, discarding part-words. On the other hand, listeners exposed to the AL with statistics and universal prosody available in the stream were penalized in their ALL performance when TP-words (byproducts of segmentation processing driven by both TPs and universal prosody) were confronted with stimuli with lower TPs but that were acoustically marked as paroxytones (i.e., part-words 23#1). However, even in this case, listeners were still able to choose above chance the TP-units as the probable words of the new language.

Therefore, listeners’ performance in the ALL test phase was modulated by the type of stimuli with which statistical outputs (i.e., TP-words) were confronted with (i.e., part-words). This demonstrates that during speech segmentation processing (even when high-level information is not available, as in ALL settings), listeners are sensitive to the different available cues and to the multiple possible byproducts of those cues, even being able to consider one cue (i.e., universal prosodic information) more reliable than the others (see also Shukla et al., 2007; Experiment 4). Thus, listeners’ performance in the forced-choice test is not the product of an all-or-none adoption of a segmentation procedure. Instead, it is influenced by the reliability of the segmentation cues that underlie the presented potential “words”. This is in line with Mattys et al. (2005) proposal suggesting that the role of any speech segmentation cue is a graded rather than an all-or-none phenomenon.

Shukla et al. (2007) proposed that universal prosody would act as to filter the output of statistical computations, suppressing the TP-words that were incompatible with prosody segmentation hypotheses. If the only function of prosody would be to allow or suppress the units extracted on a TP-basis as a “higher” filter of statistical
outputs, probably no redundancy gain would be observed when TPs and universal prosodic cues suggested the same parsing (i.e., in the congruent cues condition). On the contrary, in Experiment 2 a redundancy gain was observed both with intact and physically degraded signal. Additionally, universal prosody was also able to drive the segmentation process even when no statistical output was congruent with it and hence no TP-word was compatible with the universal prosodic segmentation procedure.

The role of prosody in speech segmentation is rather different when, instead of a universal cue (as in Experiment 2 of the present chapter), the available prosodic cue is primary stress, namely a language-specific cue, as in Experiment 3B. As a matter of fact, in that case, the availability of congruent stress cues with TPs did not endorse a general redundancy gain neither enabled the selection of all statistical outputs that were congruent with it. Had that been the case and we would have found the best performance in the condition in which TP-words were stressed on their second syllable, which corresponds to the default stress pattern in listeners’ native language. In fact, three findings suggest that language-specific prosody does not act as a filter of statistical segmentation outputs. First, no impact (either negative or positive) of stress pattern was found with intact speech: independently of the lexical stress pattern of the stimuli, all listeners were able to correctly extract the statistical segmentation outputs from the stream, and they did so at the same performance level. Second, stress pattern effects only emerged with degraded signal, in line with Mattys and colleagues’ proposal (Mattys, 2004; Mattys et al., 2005), and were maximized in the strongly degraded (i.e., 10dB SNR) condition. Third, when both statistical and stress cues were available in the stream and the strong degradation of the signal
enabled stress cues to operate as efficiently as the statistical learning mechanism, only units parsed from the stream that were highly compatible with both cues - paroxytone TP-words - were considered as reliable outcomes of segmentation. Neither the statistical byproducts (high- and low-TP-words), nor the stress byproducts (stimuli with stress located on their penultimate syllable: part-words for listeners in the 1\textsuperscript{st} and 3\textsuperscript{rd} syl-stressed conditions; and TP-words for listeners in the 2\textsuperscript{nd} syl-stressed condition) were considered as probable words of the new language as long as they were not jointly supported by all the available cues (TPs and primary stress). In other words, only those units jointly supported by both available cues (paroxytone TP-words) were considered as “lexical items”.

Segmentation outputs considered by listeners as probable “units” of an unknown language do not seem to be the product of one preponderant segmentation cue acting as a filter over the outputs of another. Instead, a more parsimonious account would suggest that the outcome of speech segmentation is largely the result of the conjunction of the available cues, both in accordance with their nature and their weighted reliability. As long as a reliable cue is available in the speech stream (e.g., in intact speech, coarticulation as presented in Experiment 1 of chapter two of this thesis; in any listening condition, universal prosodic cues: Shukla et al., 2007 and Experiment 2 of the present chapter), its byproducts are largely considered as reliable potential words. If these units are also supported by a lower-weighted cue, such as the general domain TPs, a redundancy gain will probably be observed. Indeed, Fernandes et al. (2007; see Experiment 1) demonstrated that when coarticulation and TPs were congruent, with intact speech, an additive effect (for both high- and low-TP-words) allowed the optimization of speech segmentation.
processing (see Experiment 1, chapter two). Accordingly, in Experiment 2, when TP-words were acoustically marked by a prosodic edge, their correct parsing was also optimized. However, if the available cues are weakly reliable in speech segmentation, as lexical stress – the “last segmentation resource heuristic” according to Mattys et al., 2005 (see also Valiant & Levitt, 1996) and general domain TPs computation, only the units extracted from the stream that are strongly supported by both cues in conjunction will be considered as highly reliable products of segmentation. In that case, a redundancy gain (i.e., a quantitative gain in listeners’ performance) will not be observed.

In fact, word primary stress has an important role only within a well-defined period during language acquisition. Nine-month-olds consider it as a strongly reliable speech segmentation cue (e.g., Curtin et al., 2005; Johnson & Jusczyk, 2001; Jusczyk et al., 1999a, 1999b; Thiessen & Saffran, 2003). However, at 10.5 month-old, infants already consider other segmentation cues more reliable than primary stress (Jusczyk, 1999), with this prosodic cue maintaining its lower weighting in adulthood (Mattys, 2004; Mattys et al., 2005). Furthermore, lexical stress is an abstract property not always acoustically realized. General domain cues are also lower weighted, and after 8-month-old, infants already consider speech-specific cues more reliable to speech segmentation than TPs computation (Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003), which is also observed in adulthood (Fernandes et al., 2007). Thus, when lower-weighted cues are available in the stream, only units strongly supported by the conjunction of these cues will be considered.

This idea is not new. As already suggested by Christiansen and colleagues’ computational work (Christiansen et al., 1998; Christiansen & Curtin, 2005), a cue
that insure a deeper encoding of structural regularities of the input enables the reliance on more subtle aspects of the input for making correct predictions. Thus the integration of different cues does not necessarily promote a quantity gain (i.e., more units parsed from the stream) but it can promote a qualitative gain (i.e., the correct parsing and deeper encoding of highly supported units), and hence the correct extraction of the byproducts of the conjunction (i.e., intersection) of the available sources of information will minimize errors and unwanted over-generalizations. Indeed, Curtin et al. (2005; see also Creel et al., 2006) already suggested that lexical stress could benefit the learner by changing the representational landscape, providing a qualitative gain on information that learners can use in speech segmentation. Thus, the role of lexical stress could be one of modulating confusability in learning.

It is important to note that the study presented in this chapter does not advocate that lexical stress has no role in speech processing. As a matter of fact, adult listeners are sensitive to lexical stress (for electrophysiological evidences, see: Böcker et al., 1999; Cunillera et al., 2006; Friedrich et al., 2004). Yet, that does not provide direct evidence regarding its possible role in speech segmentation. Additionally, since lexical (or word primary) stress is an abstract property not always acoustically realized, the possible effectiveness of a primary stress segmentation procedure in the context of a full mature speech perception system is obviously reduced. Importantly, lexical stress seems to have a role in lexical access, in particular in some free-stressed languages. While in English (Cutler, 1986; but see Slowiaczek, Soltano, & Bernstein, 2006) no suprasegmental mismatching effects were found in words processing, in Dutch (Cutler, & van Donselaar, 2001; Donselaar, Koster, & Cutler, 2005) and Spanish (Soto-Faraco, Sebastián-Gallés,
Cutler, 2001) suprasegmental matching seems to facilitate word recognition and suprasegmental mismatch inhibits the recognition of words that are segmentally compatible with the spoken input.

In EP a considerable amount of words are minimal stress pairs, namely words that share all segmental information but differ in lexical stress location. Although many of them are strongly associated, differing on their grammatical category (/\textit{duvidt}/ and /\textit{vidt}/; the \textit{doubt} and he/she \textit{doubts}), there are also minimal pairs that have no kind of (syntactic / semantic) relation (/\textit{pës\text{"e}w}/ and /\textit{pës\text{"e}w}/; \textit{pension} and they \textit{think}). Thus, as already suggested by Soto-Faraco et al. (2001) for Spanish, in EP, stress pattern information could also reduce the population of competing word candidates during spoken word processing. Further work should test this hypothesis.

Finally, it is worth mentioning that the present results cannot be accounted for raw frequency differences between TP-words and part-words. The frequency of occurrence is potentially important in speech processing for both infants (Brent & Cartwright, 1996) and adults (e.g., Dahan & Brent, 1999), and can explain some ALL results (Vallian & Levitt, 1996). Nevertheless, former work has shown that statistical learning can occur even when the raw frequency of the AL units is controlled (Aslin et al., 1998). In the experiments presented in this chapter, the observation of a TP gradient when primary stress was available in the stream (in the degraded conditions of Experiment 3B) would not be expected by a learning process mediated by the raw frequency of the AL stimuli. Moreover, the learning pattern observed in Experiment 2 in the incongruent cues conditions (both the significant below-chance level performance in the incongruent cues-1st syl condition with intact and physically degraded signal, and the absence of learning in the incongruent cues -
2nd syl condition with intact speech) shows that listeners did not rely (or at least not only) on the absolute frequency of the items (for a similar proposal, see section 2.4. of chapter two). Had listeners used only frequency of AL items to analyze the speech stream and no differences would be observed between the different segmentation cues conditions at study.

In summary, the ALL patterns observed in the present study have shown that the nature of the available segmentation cues, and in particular their domain-generality and their role in specific languages, differently affects the outcome of speech segmentation processing. Both general-domain and universal speech-specific cues are highly resilient to physical degradation of the signal, and the latter has a preponderant role in segmentation over the former, probably due to its encapsulated processing. The role of sublexical language-specific segmentation cues is highly dependent of listening conditions (Fernandes et al., 2007; Mattys et al., 2005). Importantly, speech segmentation processing is not an all-or-none phenomenon. Listeners are able to adopt a particular segmentation procedure on the grounds of the cues reliability, but their ALL performance demonstrates that less reliable cues are not simply discarded from the beginning of processing. On the contrary, the outputs based on these cues are also considered by listeners. Thus, speech segmentation does not seem to depend on one predominant cue acting as a “higher” filter of the outputs of another lower-weighted cue. Instead, it is the product of all available sources of information, according to their weighting, and of the listening conditions (cf. Mattys et al., 2005) that delineates the future of segmentation outputs in speech processing.
Chapter 4.

Cognitive Noise is also Noise:

The Impact of Attentional Resources on Statistical Information and Coarticulation as Speech Segmentation Cues:\(^{14}\).

The nature of available information sources and listening conditions modulate the relative weight of the cues used in speech segmentation (Mattys, White & Melhorn, 2005; Fernandes, Ventura, & Kolinsky, 2007). However, until now, the study of listening conditions was confined to intact and physical degraded signal comparisons. In Experiment 3, we investigated the impact of cognitive noise on the weighting of two sublexical cues: transitional probabilities (TPs) and coarticulation. In incidental artificial language learning, while coarticulation processing was largely unaffected by a severe reduction of attentional resources, TPs computation was penalized in a graded manner. This is in sharp contrast to the pattern found with physical noise in previous studies. Thus, if we aim a realistic perspective of how listeners deal with multiple available segmentation cues in natural settings, we need to study them in conjunction, and also in listening conditions in which noise is no longer confined to physical degradation.

4.1. Introduction

Mattys and colleagues (Mattys, 2004; Mattys et al., 2005) proposed the first integrated theoretical approach on speech segmentation field (see previous section 1.4.).

---

\(^{14}\) The experiment presented in this chapter were integrated in a manuscript in preparation (Fernandes, Kolinky, & Ventura, a).
Recent studies have attempted to evaluate and clarify in a more fine-grained manner not only the possible interactions between the various sources of information represented at different tiers according to Mattys and colleagues (e.g., Mattys & Melhorn, 2007), but also between segmentation cues that were considered by these researchers as belonging to the same tier.

4.1.1. Statistics and Coarticulation: the impact of physical noise

In Experiment 1 of the present thesis (see chapter two), using an ALL learning experiment, we examined the relative impact of two sublexical cues as a function of signal quality: coarticulation, a subsegmental acoustic-dependent cue, and TPs between adjacent syllables, a segmental statistical cue, both being represented at the same middle tier in Mattys et al.’s (2005) model.

In that experiment, we have observed that the impact of each one of these two segmentation cues varies in a graded manner as a function of signal quality. In line with Mattys’ previous data (Mattys, 2004; Mattys et al., 2005), with phonetically intact speech, coarticulation was a powerful segmentation cue, overruling statistical information when both cues were put in conflict in the speech stream. Nevertheless, coarticulation seems much more affected by the presence of noise than other cue types, either lexical stress (Mattys, 2004) or statistical information (Fernandes et al., 2007). Indeed, being themselves largely insensitive to the physical degradation of the signal, TPs overrode coarticulation in strongly degraded speech. When both cues were congruently available in the stream, an additive effect allowed the optimization of the speech segmentation process. But integration was affected and hence the redundancy gain largely reduced as signal degradation increased. Thus, the
weighting change of the available segmentation cues does not occur in a linear fashion. Instead, it largely depends on both the unavailability of coarticulatory cues and the high resilience of statistical information to physical noise superimposition.

This differential weighting may partly be accounted for by the structural grain of the cues (see section 2.4 of chapter two; Mattys et al., 2005). Indeed, the use of cues defined at a lower structural level depends on the opportunity to process fine-grained, low-level acoustic properties, which are more easily masked by physical noise than higher-level units like syllables (involved in TPs). Nevertheless, cue reliability cannot be reduced to perceptual salience per se. If this were the case, it would always be easier to extract information from highly salient syllables than from perceptually less salient sub-syllabic (segmental or sub-segmental) structures, and coarticulation would not have overruled statistical information in phonetically intact speech.

In addition, although Mattys et al.’s (2005; Mattys, 2004) proposal emphasized the importance of listening conditions in the weighting of various speech segmentation cues, until now this impact was only evaluated through physical degradation of the signal (i.e., physical noise). Mattys et al. (2005) already suggested that listening conditions go beyond physical quality. Under this view, impoverishing listening conditions while maintaining intact the physical quality of the signal would also affect the kind of cues predominantly used in segmenting speech.

4.1.2. The impact of cognitive noise

A situation of cognitive noise could be achieved by reducing the participant’s attentional resources available to extract the AL words. In fact, several studies
suggest that statistically-driven segmentation can occur in the absence of attentional allocation to the AL stream. Human infants (Saffran et al., 1996a) and non-human primates (Hauser et al., 2001), who cannot be formally instructed to attend to the speech stream, are sensitive to the available TPs. In addition, human adults are able to use the statistical information conveyed by the speech stream even in incidental passive listening (Toro et al., 2005) and when they have to perform a concurrent drawing task (Saffran et al., 1997). But this does not imply that word segmentation based on TPs occurs automatically, regardless of the available attentional resources. First, in particular for human infants, the speech signal has a privileged status since a very precocious phase (Vouloumanos & Werker, 2004; 2007). Second, simply instructing to focus attention on another diverting task may not be sufficient to prevent attention-switching to the simultaneously presented AL stream. To ensure that attention is actually diverted from the AL stream, it is necessary to use another concurrent task of high processing load that engages full attention on it (e.g., Lavie, 2005). In high attention-load conditions, Toro et al. (2005) found a dramatic impairment in ALL. As a matter of fact, participants were able to use TPs in mere incidental learning (a Passive-Listening condition) and in conditions of moderate attentional load, when their attention was diverted to a RSVP task in a relatively easy condition, namely with a SOA of 750 ms. Yet, when the primary diverting task had to be performed on an auditory stream (a separate one or the same as the AL stream), or when the RVSP used a shorter SOA (500 ms), ALL performance dropped to chance level (see also Turk-Browne et al., 2005, for a visual statistical learning situation). Thus, statistical speech segmentation seems severely affected by the reduction of the attentional resources available for processing.
Experiment 4. The impact of cognitive noise on the relative weighting of speech segmentation cues.

To the best of our knowledge, until the present, only Toro et al.’s study (2005) has directly examined the impact of attentional resources on speech segmentation, and not with the aim of understanding such effects as reflecting the impact of cognitive noise on the relative weighting of various speech segmentation cues. In the Experiment 4 of the present study, we evaluated this impact as regards the weighting of two sub-lexical segmentation cues: statistics and coarticulation.

Based on the findings of the Experiment 1 presented in chapter two of this thesis, we already know that these two sources of information are differentially sensitive to physical noise, with statistical information (i.e., the extraction of TPs) being highly resilient to physical degradation and coarticulation particularly sensitive to it. Whether modulation of the weighting of these two segmentation cues by cognitive noise resembles the one observed with physical noise was evaluated in the present experiment.

To this aim, we adopted an ALL paradigm (see section 1.3.2. of chapter one on the advantages of this paradigm).

In order to evaluate the weighting of statistical information (i.e., TPs of adjacent syllables) and coarticulation, as in Experiment 1 (presented in chapter two) the AL was presented in three (between-subjects) conditions differing by the number and congruence of the segmentation cues available in the speech stream: a single cue condition, in which only the TPs between adjacent syllables could assist
segmentation into TP-words; a **congruent cues condition**, in which statistics and coarticulation suggested the same segmentation hypotheses; and an **incongruent cues condition** in which coarticulation suggested that the AL “lexical” stimuli were the ones that had the lower TPs, namely, the part-words that cross TP-words boundaries.\(^{15}\)

Two (between-subjects) conditions of cognitive noise were used. In both conditions, the AL was presented in an incidental context, simultaneously with a visual stream of pictures, as in Toro et al. (2005; Experiment 2). We did not use a concurrent auditory task in order to guarantee that the pattern of ALL could not be due to physical degradation of the AL stream, based on its simple masking by a concurrent auditory stream. In the **weak cognitive noise** condition, participants were passively but simultaneously exposed to the AL and to the visual stream of pictures, while in the **strong cognitive noise** condition, participants had to perform a repetition detection task (presented as their main task) on the visual stream. In addition, in the strong cognitive noise condition the auditory AL stream was presented as a potential distractor.

If modulation by cognitive noise of the weighting of the two segmentation cues studied here resembles the one observed with physical noise, we would observe results similar to those found in Experiment 1 of the present work (*see* chapter two) with physical noise, namely resilience of statistical information allied with a reduction of coarticulation’s reliability in the strong cognitive noise condition. Actually, if TPs were found to be the most resilient segmentation procedure, not only to physical but also to cognitive noise, any interpretation of coarticulation’s sensitivity

---

\(^{15}\) Note that part-words, although occurring in the AL stream as well as TP-words, in statistical terms, these are constituted by syllables that belong to different TP-words.
to noise superimposition based only on physical masking of acoustic properties would be dismissed.

However, there is no theoretical reason to consider that the weighting change would be similar with cognitive and physical noise, much on the contrary. In fact, the differential weighting of coarticulation and statistical cues may also depend on another basic factor: their domain generality. Indeed, the ability to track TPs appears to be precocious (e.g., Kirkham et al., 2002; Thiessen & Saffran, 2003), rapid (e.g., Saffran et al., 1996a), involuntary (e.g., Saffran, et al., 1997; but see Toro et al., 2005) and age-independent (e.g., Saffran et al., 1997). This statistical computation seems to involve a domain-general learning mechanism\textsuperscript{16}, as demonstrated by the fact that TPs are also extracted in non-linguistic material such as musical tones (Saffran et al., 1999) and visual sequences (Fiser & Aslin, 2001; Kirkham et al., 2002; Turk-Browne et al., 2005), with a possible phylogenetic origin (e.g., Hauser, et al., 2001).

On the contrary, coarticulatory information is a speech-specific cue, which is probably processed as encapsulated information (Fodor, 1983, 2000) of a modular speech perception system (\textit{for neuroanatomic evidences, see} e.g. Dehaene-Lambertz et al., 2006; Vouloumanos, Kiehl, Werker, & Liddle, 2001). Under this view, on the one hand coarticulation processing would be a mandatory, automatic operation, that could not be “\textit{switched off}” and that would be relatively unaffected by a reduction of the available attentional resources, as the one implemented in the present experiment. On the other hand, the statistical general learning mechanism, although very resilient

\textsuperscript{16} It is not yet clear whether statistical learning reflects one general process (e.g., Saffran, Pollack, Seibel, & Shkolnik, 2007), or multiple specialized processes (e.g., Conway & Christiansen, 2005, 2006), but it does appear to be subject to one unifying principle, namely a tendency toward “reduction of uncertainty” (cf. Gibson, 1991; Gúmez, 2002).
to extraneous, physical, noise is a general-domain mechanism, hence probably not encapsulated. In addition, statistical learning does not seem resilient to the reduction of attentional resources, as Toro et al.’s (2005) data suggest. Thus, with strong cognitive noise we may actually expect to observe a mirror image of the pattern presented with strong physical noise. If this were the case, one would observe coarticulation to be more resilient to cognitive noise than the statistical information afforded by TPs. In particular, in the strong cognitive noise condition, one may then expect coarticulation to drive the segmentation process even when statistics run against it, as it is the case in the incongruent cues condition.

Another important aspect outlined in Mattys and colleagues’ proposal is that the impact of noise on the weighting of the available sources of information for speech segmentation is a graded rather than an all-or-none phenomenon. Indeed, cognitive noise, as any impoverishment of listening conditions, could also have this graded impact in the available speech segmentation cues. Whether this was the case in Toro et al.’s data (2005) cannot be checked, since these authors kept the TPs of all AL words constant (they were always 1). In the present experiment, we were able to evaluate this potential impact of cognitive noise in a more fine-grained manner since we used AL words with different levels of TPs. As in Experiment 1, the AL was constituted by six trisyllabic AL-words, three with very high TPs (high-TP-words, ranging from 0.75 to 1), and three with lower TPS (low-TP-words ranging from 0.50 to 0.58). Nevertheless, all the AL words had higher TPs (on the average, 0.68) than the part-words (on the average, 0.38). With strong cognitive noise, when only statistical information is available (i.e., in the single cue condition), observation of a TP gradient effect, revealed by better performance for high-TP-words than for low-
TP-words, would nicely fit with Mattys et al.’s (2005) proposal, according to which the impact of noise is a graded phenomenon.

4.2. Method

4.2.1. Participants

Eighty-six undergraduate psychology students at the Universities of Lisbon and Évora participated in the experiment for a course credit. All were monolingual EP speakers, with no reported history of speech or hearing disorders. Among them, 41 were randomly assigned to the weak cognitive noise condition (13 in the single cue condition, 14 in the congruent cues condition, and 14 in the incongruent cues condition), and 45 to the strong cognitive noise condition (15 in each one of the three cue conditions), corresponding to the six experimental groups according to a 2 (cognitive noise: weak vs. strong) x 3 (cues: single; congruent; incongruent) experimental design. The AL gradient (high- vs. low-TP-words) was a within-subjects factor.

4.2.2. Material

**AL Stimuli**

To achieve realistic coarticulation, we used naturally produced utterances rather than synthesized speech, and created concatenated vs. coarticulated versions of the AL stimuli (*for the detailed method, see section 2.2.2. of chapter two*). Based on
the pretests of Experiment 1, we already know that the use of naturally produced rather than synthesized utterances does not impact on ALL performance.

All natural speech stimuli corresponded to the one used in Experiment 1 presented in chapter two of this thesis, and the material of the familiarization and forced-choice test phases corresponded to the one of Experiment 1 (see section 2.2.2. for details, and for an illustration of the AL-material used see Table 1 presented in chapter two).

Note however, that contrary to what happened in Experiment 1, in the present experiment no white-noise was superimposed. The only difference in the signal material regarded the type of information available in the speech stream during the AL-familiarization phase. Therefore, in the single cue condition, only the concatenated version of the AL was used: each of the CV syllables that constituted the AL inventory was recorded individually (thereby avoiding coarticulation between syllables) and then concatenated to form the AL stream. Thus, the only segmentation cue available was the TPs between syllables. In the other two versions, the stimuli that constituted either the TP-words (in the congruent cues condition) or the part-words (in the incongruent cues condition) were recorded in their coarticulated forms. These coarticulated portions substituted their concatenated exemplars of the input with the same raw frequency in these two conditions; thus, the same number of coarticulated exemplars occurred in the congruent and incongruent cues conditions. The only difference between the congruent and the incongruent conditions regarded the stimuli that appear in their coarticulated form: the TP-words in the congruent cues condition and the part-words in the incongruent cues condition. These coarticulated exemplars were located as closely as possible in the same position in
these two versions of the AL. Thus, these two versions were constituted by both coarticulated and concatenated versions.

The forced-choice test phase’s material, presented to all participants, corresponded to the one used in Experiment 1, and it was produced by concatenating the three syllables that constituted each word or part-word (see Appendix I).

**RSVP Stimuli**

For the RSVP situation, 12 pictures of three semantic categories (4 pictures per category, presented in Appendix II) were selected from Snodgrass and Vanderwart (1980). The pictures were converted into bmp files and rotated exemplars (30° to the right or to the left) of each picture were created with *Irfanview* (available at [www.irfanview.com](http://www.irfanview.com)). In the familiarization phase, the RSVP occurred simultaneously to the AL presentation. In order to have a similar duration as the blocks of the AL, three RSVP blocks (with stimuli blocked according to semantic category) of 832 trials each were created. Within each block, the four pictures appeared in both rotations and with the same frequency of occurrence. Repetitions of the same picture (independently of the type of rotation) occurred in approximately 25% of the trials. Visual stimuli were presented centrally at the same spatial location for 250 ms each, with an ISI of 250 ms, in a sequence order (SOA = 500 ms). The RSVP was displayed on a HP Compaq 7500 computer monitor of 17”. The order of the three visual blocks was randomized for each participant.

**Reading Comprehension Test**

Participants in the weak cognitive noise condition were presented with a
neutral text, i.e., an encyclopedic definition of Thermodynamics (see Appendix III), after the AL-familiarization and the forced-choice test phases. It was presented on two double-spaced pages, in Times New Roman font (size 12; see Appendix III).

4.2.3. Procedure

Participants were tested individually or in groups of two in a sound-attenuated room. Auditory stimuli presentation in the familiarization phase was done with *Windows Media Player*. Both in the familiarization phase and in the forced-choice test, all the auditory stimuli were presented at a comfortable level through *Sennheiser HD 280 Professional Silver* headphones. For the RVSP situation and for the forced-choice test phase, presentation, timing and data collection were controlled by *E-Prime 1.1* (Schneider et al., 2002a, 2002b).

All participants were informed that the experiment had two phases. They were told that in the first phase they would hear an AL and simultaneously would see rotated pictures, presented one at a time on the computer screen, through three blocks. Nothing was said about the characteristics of the AL or about the second (forced-choice test) phase of the experiment. Participants were seated approximately 50cm away from the computer screen.

Both groups of cognitive noise were presented with the AL in an incidental situation. In the *weak cognitive noise* condition, participants were told that the experiment was part of a study about reading comprehension, being aimed at evaluating any possible effect of passive exposition to neutral auditory and visual stimuli on reading comprehension of a subsequent text. They were also informed that
the text they would read later on had no direct relation to the auditory and visual stimuli that would be presented in the first phase. Thus, their task in the first phase was to passively listen and watch the stimuli that would be presented. These instructions aimed at forcing participants to allocate their attention to both the visual and auditory streams, in a nonetheless incidental passive learning situation (see Turk-Browne et al.’s 2005 discussion).

In the strong cognitive noise condition, participants were told that the experiment was part of a study about attention, and that it was aimed at evaluating any possible interference effect of hearing an AL on visual detection performance. Thus, participants’ attention was diverted to the visual stream, and the AL was presented as an irrelevant (and even potentially interfering) distractor. They were asked to press the “1” key of the keyboard with their preferred hand, as soon as they detected a repetition between two consecutive visual stimuli, independently of their orientation. Correctness of this visual detection task was emphasized.

For all participants, a 5-minutes break was given after the first and the second block of this first, familiarization phase. Then all participants performed the same forced-choice test. They were told that, although they were not asked to explicitly attend to the AL presented in the first phase, they could have captured some knowledge about the “words” that constitute it, and that was going to be evaluated in this phase. They were asked, on each trial, to indicate which of the two presented strings sounded more like a word of the language that they had heard in the first phase. The procedure for this forced-choice test phase was identical to the one of Experiment 1 (see section 2.2.3. of chapter two).

Order of presentation of test trials was randomized for each participant, and
order of presentation of the stimuli within a trial (i.e., word – part-word or part-word
– word) was counterbalanced within each group.

The four practice trials provided in Experiment 1, prior to the test, were also
used in the present experiment. Feedback on the correctness of the responses was
only provided for these practice trials.

After the forced-choice test, the groups presented with the strong cognitive
noise condition were dismissed, and the groups presented with the weak cognitive
noise condition red silently the text for a period of 5 min., after which they were
given also 5 min. to write down their answers to four questions about the text.

4.3. Results

4.3.1. Detection of Repetition in RSVP

In order to ensure that the three groups of participants exposed to the AL in
the strong cognitive noise condition were subject to a similar reduction of the
available attentional resources, we evaluated their performance in the RSVP situation
regarding the percentage of correct detections and false alarms (see average values in
Table 7).

Percentages of correct detections were arcsin transformed and analyzed
according to the mixed ANOVA with cue condition (single cue; congruent cues;
incongruent cues) as between-subjects factor and block (1 to 3) as within-subjects
factor. No main significant effect was found [cue condition: $F < 1$; block: $F(2, 84) =
3, p > .05, MSe = 0.647$], nor interaction [$F < 1$]. The mixed ANOVA with the same
factors run on the percentages of false alarms (also arcsin transformed) revealed a similar pattern, with no significant effect or interaction [all $F_s < 2$, $p_s > .10$].

Table 7: Performance pattern (mean percentage of Correct Detections – CD – and false alarms – FA) for the three AL-conditions, according to the number and congruence of segmentation cues available (single cue; congruent cues – congruent -; incongruent cues – incongruent), in the RSVP task, according to visual block (block1; block 2; block 3) and $d'$ scores in Experiment 4. Standard error of the mean for each condition is presented in parenthesis.

<table>
<thead>
<tr>
<th>AL condition</th>
<th>RSVP Task</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>block 1</td>
<td>block 2</td>
<td>block 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CD</td>
<td>FA</td>
<td>$d'$</td>
<td>CD</td>
</tr>
<tr>
<td>Single Cue</td>
<td>57.6</td>
<td>5.8</td>
<td>1.9</td>
<td>61.0</td>
</tr>
<tr>
<td></td>
<td>(4.1)</td>
<td>(1.0)</td>
<td>(0.1)</td>
<td>(4.5)</td>
</tr>
<tr>
<td>Congruent</td>
<td>60.7</td>
<td>4.2</td>
<td>2.0</td>
<td>60.1</td>
</tr>
<tr>
<td></td>
<td>(2.7)</td>
<td>(0.5)</td>
<td>(0.1)</td>
<td>(3.9)</td>
</tr>
<tr>
<td>Incongruent</td>
<td>58.5</td>
<td>2.8</td>
<td>2.2</td>
<td>62.8</td>
</tr>
<tr>
<td></td>
<td>(3.8)</td>
<td>(0.3)</td>
<td>(0.1)</td>
<td>(4.4)</td>
</tr>
</tbody>
</table>

False alarms and hits (i.e., correct detections) were combined to calculate $d'$ scores separately for each subject in each cue type and block condition (see Table 7). Coherently to what was observed separately on correct detection and false alarms, the ANOVA performed on the $d'$ scores (with cue condition and block as factors) revealed no main significant effect or interaction [all $F_s \leq 1$]. Moreover, within each cue condition the average $d'$ score was significantly greater than zero, with scores of 2.07 (SE = 0.17) in the single cue condition [$t(14) = 12.27$, $p < .0001$], 2.05 (SE = 0.1) in the congruent cues condition [$t(14) = 20.13$, $p < .0001$], and 2.28 (SE = 0.12) in the incongruent cues condition [$t(14) = 18.17$, $p < .0001$]. Thus, participants were able to discriminate the situation in which a repetition occurred from the situation in which it did not, and they did so at the same success level in the three cue conditions.
The fact that the three groups in the strong cognitive noise condition presented a similar performance in the RSVP situation ensures that any effect of the cue condition on the ALL cannot be due to attentional load differences between these groups.

4.3.2. ALL performance

The percentage of AL-word choices was computed for each participant. In Table 8, the average ALL performances are presented broken down by TP level (high- vs. low-TP-words).

These data were entered into a mixed ANOVA with cognitive noise (weak vs. strong) and cue condition (single cue; congruent cues; incongruent cues) as between-subject factors, and TP level (low- vs. high-TP-words) as a within-subject factor.

The main effect of cue condition was significant \( F(2, 80) = 46.31, p < .0001, MSe = 6.93 \), with congruent cues leading to better performance than both the single cue \( F(1, 80) = 9.18, p < .005 \) and the incongruent cues \( F(1, 80) = 89.05, p < .0001 \) conditions. Furthermore, the incongruent cues condition, which led participants to prefer coarticulated part-words over TP-words, also differed from the single cue condition \( F(1, 80) = 39.96, p < .0001 \).

Neither the effect of cognitive noise \( F(1, 80) = 2.17, p > .10, MSe = 6.93 \) nor the effect of TP-level \( F < 1 \) were significant, but their interaction was \( F(1, 80) = 3.967, p < .05, MSe = 6.65 \). Indeed, a significant difference between the two conditions of cognitive noise was found in the extraction of low-TP-words \( F(1, 80) = 6.7, p = .01, MSe = 7.51 \), with strong cognitive noise leading to the worst extraction of these AL units. On the contrary, the high-TP-words were extracted at a
similar level in both noise conditions \( F < 1 \). This suggests a graded impact of cognitive noise on statistical computation.

Table 8: Average proportions of TP-Word responses (in percentage), separately for each cue condition (single cue; congruent cues; incongruent cues) and cognitive noise condition (low vs. high load), considering the TP- gradient of the AL words (high- vs. low-TP-words) in Experiment 4. Standard error of the mean for each condition is presented in brackets.

<table>
<thead>
<tr>
<th>COGNITIVE NOISE</th>
<th>Low Load</th>
<th>High Load</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High-TP</td>
<td>Low-TP</td>
<td></td>
</tr>
<tr>
<td>Single Cue</td>
<td>57.2 (3.6)*</td>
<td>62.8 (3.2)**</td>
<td>65.5 (3.3)**</td>
</tr>
<tr>
<td>Congruent Cues</td>
<td>65.8 (3.1)**</td>
<td>73.4 (3.7)**</td>
<td>63.7 (3.5)**</td>
</tr>
<tr>
<td>Incongruent Cues</td>
<td>44.0 (3.8)(*)</td>
<td>45.2 (3.6)(*)</td>
<td>41.1 (3.9)*</td>
</tr>
<tr>
<td>Average</td>
<td>55.7</td>
<td>60.5</td>
<td>56.8</td>
</tr>
</tbody>
</table>

One-way t tests (in comparison to chance level, which corresponds to 50%):

(*) \( p \leq .10 \)

* \( p < .05 \)

** at least \( p < .01 \)

Underlined scores refer to scores significantly below chance level, namely a preference for coarticulated part-words.

Although neither the interaction between cue condition and cognitive noise \( [F<1] \) nor the three-way interaction \( [F(2, 80) = 1.34, \ p > .25] \) were significant, it is worth noting that the effect of cue condition was modulated by TP level only in the strong cognitive noise condition \( [F(2, 42) = 3.38, \ p < .05] \) but not in the weak one \( [F<1] \). Indeed, as revealed by planned comparisons, using the Bonferroni-corrected alpha rate of .017, with strong cognitive noise a significant TP-gradient was found only when listeners were presented with purely statistical information \( [F(1, 42) = 7.53, \ p < .01] \). No differences were found between the performance for high- and low-TP-words for listeners presented with congruent and incongruent cues \( [Fs < 1] \).
Thus, with strong cognitive noise, high-TP-words were learned significantly better than low-TP-words, in the single cue condition. In addition, while in the single cue condition the high-TP-words were correctly extracted from the continuous stream whatever the level of cognitive noise (i.e., were chosen more often as the “words” of the AL than part-words, see the t-tests results in Table 8), performance for low-TP-words was at chance in the single cue condition with strong cognitive noise [$t < 1$]. Accordingly, in the strong cognitive noise condition it was exclusively for low-TP-words that participants exposed to congruent cues performed better than those exposed to the single cue condition [$F(1, 42) = 9.58, p < .005$].

This pattern of results thus suggests that the statistical computations linked to the extraction of TPs are penalized by a reduction of attentional resources, and that this effect is graded, affecting performance on low-TP-words but not on high-TP-words.

Cognitive noise seems to affect less and for sure differently the use of coarticulation cues. In fact, with weak cognitive noise, performance in the incongruent cues condition only tended to differ from chance level for both high- and low-TP-words (see Table 8). This may be due to the fact that with soft reduction of the attentional resources both types of cues were still sufficiently available and hence competed with each other, which was not the case anymore when attentional resources were severely reduced. Indeed, when strong cognitive noise disabled participant to use the TPs efficiently (i.e., for both low- and high-TP-words), listeners did rely reliably on coarticulation cues (see Table 8).
4.4. Discussion

Former work has already shown the impact of listening conditions on the relative weighting of various speech segmentation cues (Mattys, 2004; Mattys et al., 2005). Nevertheless, until the present, this impact has only been evaluated by comparing intact speech with a physically degraded signal. Mattys et al. (2005) already suggested that listening conditions go beyond the physical quality of the signal. Under this view, impoverishing listening conditions while maintaining intact the physical quality of the signal would also affect the kind of cues predominantly used in segmenting speech.

In the present experiment, we used an incidental ALL setting to investigate the impact of cognitive noise (i.e., attentional load) on the weighting of two types of sublexical cues to speech segmentation: statistical information – TPs between adjacent syllables – and subsegmental information – coarticulation. In both the weak and strong cognitive noise conditions, participants were presented not only with the AL auditory stream, but also and simultaneously with a visual (RSVP) stream. While participants in the weak cognitive noise condition had no task to perform on these visual stimuli, those in the strong cognitive noise condition had to detect visual repetitions, a task presented as their main one.

In Experiment 1 (see chapter one of the present thesis) we already demonstrated that the segmentation process adopted by listeners when both statistics and coarticulation are available in the stream varies as a function of signal quality. Indeed, with phonetically intact speech, when only sublexical information is available in the stream, coarticulation is a powerful segmentation cue, able to drive
the segmentation process even when TPs (Experiment 1 of the present work) or lexical stress (Mattys, 2004; Mattys et al., 2005) act against it. However, coarticulation is much more affected by the presence of physical noise than other cue types, while TPs were highly resilient to this noise superimposition. Moreover, the weighting change of the available segmentation cues largely depended on both the unavailability of coarticulation and the high resilience of statistical information to physical noise superimposition.

We hypothesized that cognitive noise would also modulate the weighting of these two segmentation cues, but in a different way. Indeed, although both segmentation cues at study are available since an early age (e.g., Johnson & Jusczyk, 2001; Kirkham et al., 2002; Saffran et al., 1996a) and are used by adult listeners (e.g., Mattys, 2004; Saffran et al., 1996b), they differ as regards their domain generality. Coarticulation is speech-specific and hence probably processed in an encapsulated manner. Modules stipulate a mandatory, automatic processing in the presence of the “correct” information (Fodor, 1983, 2000), which is highly independent of “horizontal faculties” (Fodor, 1983), in particular of the available attentional resources. Thus, while this speech-specific acoustic-dependent cue is much more affected by physical degradation of the signal than statistical information, it may be far less sensitive to cognitive noise, since segmentation based on TPs relies on a general learning mechanism that is not speech-specific (e.g., Fiser & Aslin, 2001; Perruchet & Pacton, 2006; Saffran et al., 1999; Turk-Brown et al., 2005) and that has been shown to be affected by a strong reduction of the available attentional resources (Toro et al., 2005). In addition, following Mattys et al. (2005), we hypothesized that the impact of noise on the weighting of segmentation cue is a
graded rather than an all-or-none phenomenon. Extraction of the AL words with low TPs would thus be much more affected by cognitive noise than extraction of AL words with high TPs.

In agreement with this view, we have found a weighting change of these sublexical segmentation cues modulated by cognitive noise that tends to mirror the one observed with physical noise. With strong cognitive noise, namely when the participants’ attention was focused on the RSVP task, coarticulation was able to drive the segmentation process even when statistical information was also available in the stream. On the opposite, the ability to use statistical information alone (in the single cue condition) was severely affected by cognitive noise.

Toro et al. (2005) had already demonstrated that TPs computation can be severely compromised by a strong reduction of the available attentional resources. Our results add to this evidence by suggesting not only that statistical computation requires attentional resources, but that the impact of cognitive noise is a graded phenomenon. Indeed, with strong cognitive noise, when only statistical information was available in the stream, a TP gradient effect (i.e., better performances in the forced-choice task for TP-words with high than with low TPs) was observed. In addition, the learning effect was exclusive for AL words with high TPs, since with strong cognitive noise participants were unable to extract the TP words presenting the lowest TPs.

Nevertheless, observation of a statistical learning effect with strong cognitive noise, although impaired since restricted to high-TP-words, may seem at odds with Toro et al.’s (2005) data, according to which high-TP-words (since they only used a constant TP equal to 1) were not learned at all in a similar setting. This discrepancy
cannot be attributed to cross-study differences in the reduction of the attentional load by the RSVP task, since all the groups in the strong cognitive noise condition displayed a similar performance in the visual task that was remarkably similar to the one observed by Toro et al. (2005). Also, the observed ALL pattern cannot be attributed to a practice effect that could have allowed participants to become more and more efficient in dual task conditions, with one task (here, RSVP) becoming automatic, which would allow attentional resources to be allocated to the other task, namely here, ALL (see Shanks, Rowland, & Ranger, 2005, for a similar argument as regards sequence learning). Had it been the case, we should have observed an improvement in the visual task performance along the three blocks, which we did not, probably because we used different stimuli in these blocks. Rather, this discrepancy may be attributed to the much longer familiarization duration we used, actually three times longer (21 min.) than the one used by Toro et al. (7 min.). As acknowledged by these authors, it is thus possible that divided attention does not completely prevent, but only slows down, statistical learning. When taken together with Toro et al.’s data, the present results suggest that this is the case for AL words with high TPs. In Experiment 5 presented in chapter five of the present thesis we tested whether the amount of familiarization would have any impact on TPs computation, namely whether longer familiarization durations (21 minutes vs. 7 minutes) would allow low-TP-words to be extracted as well as high-TP-words in conditions of strong cognitive noise.

Interestingly, with passive exposition to visual stimuli (in the weak cognitive noise condition), participants were still able to extract both high- and low-TP-words, and no TP gradient was observed. Taken together with the fact that participants only
extracted high-TP-words in the strong cognitive noise condition, these results are coherent with former observations showing that human adults are able to use the statistical information conveyed by the speech stream even in incidental passive listening (Toro et al., 2005) and when they have to perform a concurrent but easy task (drawing: Saffran et al., 1997; easy RVSP task: Toro et al., 2005), but not when their primary diverting task is attentional demanding (Toro et al., 2005; Turk-Browne et al., 2005). It thus seems that at least some attentional resources are required for an optimal extraction of “lexical” units based on statistical computation. However, attention is generally conceived as, on the one hand a selective mechanism, and on the other hand as limited-capacity resources. Thus, we do not know, based in the present results and in Toro et al.’s, which one (or if both) attentional mechanism is responsible for an impairment on statistical speech segmentation in conditions of strong cognitive noise. This was also evaluated in chapter five of the present thesis in Experiment 6.

Although coarticulation is a more reliable cue than statistics in phonetic intact speech, as observed in Experiment 1 of the present work as well as in conditions of strong cognitive noise (as shown by the results of this experiment), it does not seem sufficiently powerful to overrule incongruent statistical information in a condition of weak cognitive noise, namely when participants were merely exposed to visual stimuli, without having to perform any task on them. This result pattern differs from the preponderance of coarticulation observed when intact speech is presented without any distracting visual information, as it was the case in the no noise incongruent cues condition of Experiment 1 (see chapter one). The absence of strong preponderance of any of the incongruent sublexical cues available in the stream in the weak cognitive
noise condition is similar to what was observed in Experiment 1 with mild physical noise. Thus, a soft reduction of the attentional resources available to speech segmentation through passive exposition to visual stimuli seems to have the same (at least quantitative) impact as a mild physical degradation of the auditory AL stream. In both cases, the product of the available incongruent cues is disrupted. As already commented on, in the present case this may be due to the fact that with soft reduction of the attentional resources both types of cues were still sufficiently available and hence competed with each other, which was not the case anymore when attentional resources were more strongly degraded. Indeed, when strong cognitive noise disabled participant to optimally extract TP-units from the stream, listeners did rely reliably on coarticulation when TPs were incongruent with that information.

The findings of this experiment also demonstrate that when two different but congruent speech segmentation cues are available in the stream, their integration can be achieved even in strongly degraded attentional conditions. Indeed, with strong cognitive noise a redundancy gain (better performance when both statistics and coarticulation suggested the same word-boundaries) was found for the AL words with low TPs. Thus, strong cognitive noise hinders the efficient use of statistical information when this is the only source of information that can help listeners to locate word boundaries (i.e., for low-TP-words in the single cue condition), but this information, if congruent with a more robust one (here coarticulation) can still be used by the processing system to optimize performance.

Finally, it is worth mentioning that the present results as well as the results of the previous experiments of the work presented in this thesis cannot be accounted for raw frequency differences between AL words and part-words. The frequency of
occurrence is potentially important in speech processing for both infants (Brent & Cartwright, 1996) and adults (e.g., Dahan & Brent, 1999). Former work has shown, however, that statistical learning can occur even when the raw frequency of the AL units is controlled for (Aslin et al., 1998). Here, the observation of a TP gradient (at least in the strong cognitive noise condition), as already observed in Experiment 3B with stress and TPs available in physical noise conditions, would not be expected by a learning process mediated by the raw frequency of the AL stimuli. Moreover, the learning pattern observed in the incongruent cues condition (both the significant below chance level performance with strong cognitive noise, and the absence of learning with weak cognitive noise) shows that listeners did not rely (or at least not only) on the absolute frequency of the items. Had listeners used only frequency of the trisyllabic items to analyze the speech stream, performance should have been similar in the three cues conditions.

In summary, the AL learning patterns observed in the present study have shown that cognitive noise affects differently the use of two cues that may assist segmentation of speech into words, coarticulation and TPs. Although this effect was gradual rather than all-of-none, strong cognitive noise affected more the use of statistical information than of coarticulation. The impoverishment of the statistically driven segmentation procedure by strong cognitive noise cannot be due to a general increase in task difficulty with attentional load, since in Experiment 1 we have demonstrated that sensory degradation (i.e., reducing the physical quality of the signal by white-noise superimposition), which also increased task difficulty, had a
different impact on the relative weighting of coarticulation and TPs as cues to speech segmentation.

The observation of an impact of listening conditions in speech segmentation that is not confined to signal degradation suggests that if we aim to obtain a realistic perspective of how listeners deal with multiple available segmentation cues in natural settings, we need to study them in conjunction, and also in different listening conditions. Therefore, noise should no longer be restricted to the physical degradation of the speech stream.
Chapter 5.

What is the impact of attention on Statistical Learning?

Amount of Familiarization and Intention to Learn in ALL\textsuperscript{17}.

In the present chapter, we investigated the impact of attention as a selection mechanism and as capacity-limited resources on speech segmentation based on transitional probabilities (\textit{TPs}).

In Experiment 5, we evaluated the extent to which the statistical learning mechanism operates efficiently in different AL familiarization time conditions (7-, 14-, 21-minutes), under low and high attentional load. Under high load, the longer the familiarization-time, the greater the advantage for the extraction of high-TP-words over low-TP-words. In Experiment 6, we evaluated what is the attentional mechanism responsible for the impaired statistically-driven speech segmentation observed for low-TP-words in Experiment 5. Although selectively attending to the AL stream led to an overall performance gain, this did not overrule the impact of attentional-resources load. Thus, TPs computation is attentional-resources’ dependent. Yet, it occurs even when the AL is not the focus of attention and scarce resources are available, suggesting some automaticity of statistically-driven speech segmentation processes.

5.1. Introduction

The intensely dynamic environment that surrounds us mainly involves stimuli

\textsuperscript{17} The experiments presented in this chapter are part of a manuscript in preparation (Fernandes, Ventura, & Kolinky).
occurring in temporal and/or spatial sequences. Thus, sequence learning is a fundamental process involved in many different cognitive skills. Among them, language is probably the most complex, and the role that elementary associative sequence learning may play in its development has begun to be explored in the last decade.

Saffran and collaborators (e.g., Saffran et al., 1996a, 1996b) have proposed a statistical learning mechanism responsible for listeners’ ability to discover “words” embedded in an AL (see sections 1.1.2. and 1.3. of chapter one).

TPs-based statistical learning actually shares many features with other sequence learning effects studied within the framework of IL (Perruchet & Pacton, 2006; see section 1.3.1. of chapter one). Surprisingly, however, while in the latter framework a vast bulk of research has been devoted to clarify the relation between automaticity, attention, and sequence learning; the relation with attention and hence the potential automaticity of statistical learning has received little direct study. The aim of the series of experiments presented in this chapter was to evaluate these relations as regards ALL effects. In particular, we examined if statistically-driven speech segmentation is an attentional-dependent mechanism, and, importantly, whether it depends on attention as a selection process or as a limited-resources capacity.

Indeed, (see section 1.5.2. of chapter one), since the beginning of the cognitive revolution the distinction between automatic and controlled processes casts largely on their reliance on attention. The traditional view (e.g., Posner & Snyder, 1975; Shiffrin & Schneider, 1977) conceived automaticity as an all-or-none concept. Thus, automatic processes (contrary to controlled ones) were by definition not
dependent on attention. Yet, the distinction between automatic and controlled process is far more complex than simply relying or not on attentional mechanisms. Indeed, Stroop interference, generally thought to be unintentional and uncontrollable, is not independent of attentional limitations (e.g., Kahneman & Chajzyck, 1983). Thus, automaticity is not an all-or-none construct but rather a graded one (Cohen et al., 2004; Logan, 1988; Logan et al., 1999; Moors & De Houwer, 2006).

In addition, attention is a very broad term (e.g., Cohen et al., 2004; Moors & De Houwer, 2006). Roughly, two mechanisms of attention can be considered: (i) a selection mechanism that avoids overloading of the information processing system by filtering task-relevant aspects of incoming information; and (ii) a resources mechanism, or “source of energy” (Moors & De Houwer, 2006) of limited capacity that can be flexibly allocated to any stage in the information-processing chain.

5.1.1. Statistically-driven speech segmentation and IL

In the framework of IL, several studies have addressed the possible role of attention as central resource. They mainly used the Serial Reaction Time Task (see Cleeremans et al., 1998) in dual-task conditions in which a concurrent secondary task (e.g., auditory tone counting; or symbol counting) performed on the same sequence as the one used in the Serial RT task is introduced to reduce the available attentional resources. Yet, results from such manipulations have been quite discrepant. On the one hand, Jiménez and Mendéz (1999) suggested that a division of attention barely affected IL, but that selective attention to the predictive dimensions of the stimuli is needed for learning to occur (see also Jiang & Chun, 2001, and, for a similar proposal regarding visual statistical learning, see Turk-Browne et al., 2005). On the
other hand, Shanks et al., (2005) suggested that learning under dual-task conditions impairs sequence learning.

Rowland and Shanks (2006a) suggested that these observations could be reconciled if selection and resources are considered as inextricably linked, with selection becoming increasingly focused when resources are scarce (i.e., when the relevant task becomes more demanding), as suggested by the *Attentional Load Theory* of Lavie (1995, 2000, for a review see Lavie, 2005). Considering this theory, Shanks and colleagues (Shanks et al., 2005; Rowland & Shanks, 2006a; 2006b) proposed that learning of multiple independent contingencies in a sequence learning task will occur provided the primary task, as a consequence of being undemanding, allows spillover of attention to a secondary set of contingencies. Learning of secondary sequence is not expected if the demands of the primary sequence are so high as to preclude the spillover of attention to the secondary set. In other words, IL would be largely affected by a reduction of central attentional resources.

However, whether attentional mechanisms impact ALL in a similar way as in implicit sequence learning remains to be determined.

### 5.1.2. The role of attention in statistical learning

What is already clear is that statistically-driven speech segmentation can occur without a specific effort from listeners to segment the continuous speech stream. First, the observation of TP-based segmentation in infants or even in non-human primates (Saffran et al., 1996a; Hauser et al., 2001) as well as of an age-independent ALL effect (Saffran et al., 1997) suggests that statistical learning may
be efficient since an early phase, and relatively independent of both cognitive
development and of intention to learn. Second, ALL was observed even when
participants were performing a primary visual color-drawing task, the AL being
presented as “background” nonsense auditory stream (Saffran et al., 1997). Since
participants were told neither that the stream consisted of an AL, nor that they would
be tested later in any way, the observation of an ALL effect in these conditions has
been interpreted as evidence that statistical learning proceeds incidentally (Saffran et
al., 1997) “as a byproduct of mere exposure” (Saffran et al., 1999).

Yet, showing that statistical learning is incidental, in the sense that it occurs
unintentionally when it is not part of the task requirements, does not mean that it is a
fully automatic or attention-free process (see e.g. Moors & De Houwer, 2006). For
example, although participants were probably less attentive to the background
auditory stream in Saffran et al.’s (1997) incidental condition than in an intentional
learning situation, namely when they are asked to attend to the stream and to actively
try to find embedded nonsense “words”, as it is often the case in ALL studies (e.g.,
Saffran et al., 1996b; Vroomen et al., 1998; Experiment 3), merely instructing them
to focus attention on another task may not be sufficient to prevent attention-
switching to the simultaneously presented speech stream. In situations of low
perceptual load (cf. Lavie, 1995) such as the one used by Saffran et al. (1997), any
attentional capacity not taken up by the processing of task-relevant stimuli (the color-
drawing task) might involuntarily spill over to the perception of task-irrelevant
information (e.g., Lavie, 1995; 2005). A more stringent test to diagnose the
automaticity of a mechanism would be one in which the critical processing is neither
required nor beneficial to the task intentionally performed by the participants.
(Tzelgov, 1997). Hence, according to this view, the best way to examine the impact of attention (both as a selection mechanism and as capacity-limited resources) on statistical learning would be to use a situation in which the statistical information is presented as secondary, irrelevant, and even potential harmful (and thus not selectively attended to) for another resource-consuming task that is presented as the main participants’ undertaking.

This is exactly the situation used by Toro et al. (2005), and the one used in Experiment 4 of the previous chapter in order to evaluate the weighting of coarticulation and TPs in cognitive noise conditions. While participants in an incidental “passive listening” condition (in Experiment 4, weak cognitive noise, and in the present experiment, low-attention load condition) were able to extract the statistical information conveyed by the stream, those presented with a high-attention load condition18 (in Experiment 4, strong cognitive noise) showed degraded performance. Indeed, no learning effect occurred when attention was diverted to a difficult, unrelated, task that had to be performed either in the same modality (on a separate auditory stream – in Toro et al.’s Experiment 1 – or on a different, non-predictive, feature – pitch changes – of the AL sequence itself – in Toro et al.’s Experiment 3), or in a different modality, namely on pictures presented at a rapid rate (SOA = 500 ms; Toro et al.’s Experiment 2). While the impairment observed when attention was diverted to another auditory task could stem from simple physical masking (in Toro et al.’s Experiment 1), from disruption of the temporal organization of the speech stream by irrelevant acoustic variations (in Toro et al.’s Experiment 3),

18 In Experiment 4, presented in the previous chapter of this thesis, we performed a similar attentional load manipulation to evaluate its impact on the relative weight of two speech segmentation cues. In that experiment, we had adopted a different terminology (i.e., cognitive noise) than the one used here (i.e., attentional load). This could seem at odds. Yet, the present terminology seems more appropriate to our current goal (to evaluate the exact impact of attention on speech segmentation driven by statistical cues) and is more in line with the terminology used in former studies examining this question (Toro et al., 2005).
or from the availability of incongruent speech segmentation cues (also in Toro et al.’s Experiment 3)\textsuperscript{19}, this could not have been the case with the visual task. In addition, when their attention was diverted to an easier visual task (in which the pictures were presented with a 750 ms SOA), participants were still able to extract the statistical regularities and did this with the same success as in passive listening. A straightforward interpretation of these results is thus that at least some attentional resources must be available in order to segment the speech stream on the basis of statistical information.

However, in Experiment 4 of the present work (see chapter 4 of this thesis) we have observed that in a strong cognitive noise condition (similar to the high-attention load of Toro et al.’s Experiment 2), in an AL familiarization condition of 21-minutes, when only TPs were available, listeners were still able to extract from the stream the AL-words with the highest TPs (i.e., high-TP-words), even though their performance for low-TP-words was at chance. Statistical learning seems to be an attention-dependent mechanism. However, several points were left unanswered by Toro et al.’s (2005) study, and these were also not approached on the previous experiment of the study presented in this thesis (see chapter four). First, although Toro and colleagues seem to prefer an inhibition interpretation (with the high-attention load task engaging full capacity in relevant processing and hence leaving no spare capacity for the perception of task-irrelevant stimuli, cf. Lavie, 2005), they also

\textsuperscript{19} Indeed, pitch change is one of the acoustic correlates of prosody; since in Spanish, the participants’ native language on Toro et al.’s, lexical stress can be located on any one of the last three syllables of a polysyllabic word, although random, the pitch changes did follow a legal pattern of that language. Thus, the prosodic information available in the AL stream might have not only competed with statistical information in the segmentation process, but, importantly, might have disrupted the correct statistically-based extraction of AL-words (see Results section of Experiment 3B of the chapter 3 of the present thesis). In line with this supposition, in the passive listening condition of Toro et al. (2005), listeners exposed to that AL stream (Experiment 3) seem to have performed more poorly than those presented with a stream in which no such acoustic information was available plus a simultaneous auditory stream (Experiment 1).
suggested that diverting attention from the AL may not completely prevent statistically-driven speech segmentation, but would only slow it down. Given that the familiarization phase was held constant at 7 minutes in Toro et al.’s study, this possibility was not explored. The idea that ALL performance may improve with longer familiarization times even when attention is diverted to a primary task is supported by Saffran et al.’s (1997) observation of a significant ALL improvement when listeners were exposed to a longer familiarization phase, i.e., to two sessions of 21 minutes each rather than to one 21 minutes session, while performing a primary visual color-drawing task. However, as already mentioned it seems plausible that Saffran et al.’s task was not very resource consuming. But the impact of amount of familiarization could also possibly explain the discrepancy between Toro et al.’s results and the ones found in Experiment 4 of the present study (in chapter four of this thesis). Note that in Experiment 4 of the present study we adopted an AL-familiarization phase of 21 minutes (three times longer than the one used by Toro et al.’s) using attention-loads similar to the ones of Toro et al. (2005).

5.1.3. An overview of the experiments presented in this chapter

In the present chapter, we examined, in Experiment 5, the extent to which the amount of AL-familiarization would influence ALL in conditions of low- and high-attention load similar to the ones used by Toro et al. (2005, Experiment 2) and already adopted in Experiment 4 of the present work with cognitive noise.

Second, although Toro et al. (2005) privileged the interpretation that at least some attentional resources must be available for statistical learning to be efficient, their study does not allow us to know exactly how attention impacts the extraction of
TPs. In particular, we do not know whether attention acts in a graded or in an all-or-none way. Indeed, as demonstrated in the present work (see the previous chapters of this thesis) statistical information is one of the possible sources of information that listeners use in speech segmentation (e.g., Saffran et al., 1996b). Mattys et al. (2005) proposed that the involvement of any segmentation cue, either lexically-driven or signal-derived, is a graded rather than an all-or-none phenomenon. Speech segmentation is largely the product of the differential weighting of the types of information available in the signal and of listening conditions. Yet, until now the impact of the impoverishment of listening conditions was only evaluated through physical degradation of the signal (i.e., physical noise). Mattys et al. (2005) already suggested that listening conditions go beyond physical quality. Under this view, impoverishing listening conditions while maintaining intact the physical quality of the signal would also affect the kind of cues predominantly used in segmenting speech. Furthermore, in line with the results found in Experiment 4, the impact of attention load (which may be conceived as a kind of cognitive noise) was also graded. However, we do not know if this would always be the case independently of the amount of familiarization of listeners with the AL. Note that although the answer to this question cannot be derived from Toro et al.’s data, because in their material the TPs between adjacent syllables were constant, with values of 1 for TPs within AL-words and of 0.33 for syllables spanning AL-word boundaries, in Experiment 4 we found a TP-gradient effect in a condition in which the amount of listeners’ familiarization with the AL largely differed from the one of Toro et al. (21-minuts in Experiment 4 of the present work; 7-minuts in Toro et al.). Therefore, in Experiment 5 and 6 of the present chapter, we examined whether there is a TP gradient by using
TP-words with either high or low TPs.

Third, and importantly, from Toro et al. (2005)’s study and the results of Experiment 4 of this work, we do not know if the severe impact of high-attention load on statistically-driven speech segmentation is only due to the drastic reduction of the available attentional resources, or if it (also) reflects a filtering cost. Indeed, in Toro et al.’s high-attention load condition and in our strong cognitive noise conditions of Experiment 4, the AL stream was presented as an irrelevant distractor to the only but resource-consuming task participants had to perform (e.g., the detection of picture repetition). Consequently, the AL stream was most probably not selectively attended to in this high-attention load (or strong cognitive noise) condition. Thus, both scarce attentional resources and/or lack of selective attention might have been responsible for the statistically-driven speech segmentation impairment observed in the high-attention load condition.

In order to disentangle the role of attention as limited-capacity resources and as a selection mechanism, in Experiment 6, we manipulated orthogonally the available attentional resources (low- and high-attention load conditions, as in Experiment 5) and the focus of participants’ attention, using incidental and intentional learning instructions.
Experiment 5. Exposition Time and Attentional Resources in Artificial Language Learning

The design of the present experiment was similar to the one of Toro et al. (2005, Experiment 2) and the one used in Experiment 4 (see chapter four).

Attention was diverted from the speech stream by using a RSVP of pictures with a SOA of 500 ms – picture duration was 250 ms and ISI was 250 ms. A visual task was used for guaranteeing that the pattern of ALL could not be due to physical degradation or disruption of the AL stream by the availability of incongruent segmentation cues.

In comparison to Toro et al.’s (2005) experiment, two main modifications were introduced. First, instead of using a fixed duration of 7 minutes in the familiarization phase, participants were exposed to familiarization phases of 7, 14, or 21 minutes (henceforth, 7-min, 14-min, 21-min). This allows us to evaluate whether attentional load inhibits or slows down statistical learning. Indeed, if attentional load prevents statistical computations, speech segmentation based on statistical learning would be seriously compromised in all familiarization duration conditions. But if attentional load merely slows down statistical learning, it is possible that with longer familiarization the processing system becomes able to extract the AL-units.

Second, based on our previous finding of a TP-gradient effect in a high-attention load (i.e., strong cognitive noise) condition in Experiment 4 of the present work we evaluated whether attentional load affects speech segmentation based on statistical learning in an all-or-none or graded way. To this aim, we adopted the same
AL used in the previous experiments of this work, which was constituted by two types of TP-words: half had high TPs (high-TP-words, with TPs from 0.75 to 1) and the other half low-TPs (low-TP-words, with TPs from 0.50 to 0.58); both types having the same raw frequency (see Appendix I). The observation of a TP-gradient, revealed by better performance for high- than for low-TP-words, would nicely fit with Mattys et al.’s (2005) proposal, according to which the impact of any type of noise (physical or cognitive) is a graded rather than an all-or-none phenomenon.

5.2. Method

5.2.1. Participants

Ninety-one undergraduate psychology students at the University of Lisbon participated in the experiment for a course credit. All were monolingual EP speakers, with no reported history of speech or hearing disorders and normal or corrected-to-normal vision. Among them, 45 were randomly assigned to the low-attention load (i.e., passive listening) condition: 17 were exposed to one block including 630 trisyllabic tokens (7-min condition), 15 to two blocks (14-min condition), and 13 to three blocks (21-min condition). The other 46 participants were assigned to the high-attention load condition: 14 in the 7-min condition, 17 in the 14-min condition, and 15 in the 21-min condition.

5.2.2. Material

All the speech stimuli were synthesized as in the previous experiments presented in this thesis (i.e., with MBROLA software; Dutoit et al., 1996).
The AL used in the previous experiments (see Appendix I) was here adopted. For the familiarization phase, three 7-min blocks of the synthesized stream by concatenating 105 tokens of each of the six words (1890 syllables, 630 tokens of “words”), with the criterion that two tokens of the same AL-word never occurred adjacently in the stream. The segmentation cue available in the stream was the statistical information, i.e., TP between adjacent syllables, with no other acoustic or segmentation cues to word boundaries.

The Visual Repetition Detection Task was similar to the one used in Toro et al.’s (2005; Experiment 2) and identical to the one of Experiment 4 (see section 4.2.2. of the previous chapter of this thesis; see Appendix II).

For groups exposed to only one block (i.e., 7-min exposition), one third of the participants were presented to each possible block. For groups exposed to two blocks (i.e., 14-min exposition) the blocks were randomly selected from the three possible ones. For the groups exposed to three blocks (i.e., 21-min exposition) the order of blocks was randomized for each participant.

The forced-choice test phase, performed by all participants, was identical to the one of Experiment 2 (see section 3.3.2 of chapter three). For the low-attention load (passive listening) condition the “neutral” text of Experiment 4 was also used (in Appendix III; see section 4.2.2. of chapter four).

5.2.3. Procedure

Participants were tested individually or in groups of two in a sound-attenuated room. Presentation of the auditory and RSVP material was done as in Experiment 4 (see section 4.2.3. of the previous chapter). All participants were told
that in the first phase they would hear an AL (no specific information about what is an AL was given) and simultaneously they would see rotated pictures, one at a time. Nothing was said about the ALL test. Participants were seated 50 cm away from the computer screen.

Low-attention load groups were told that this study was aimed at evaluating any possible effect of passive exposition to neutral auditory and visual stimuli on the reading comprehension of a subsequent text with no direct relation to the former material. Their task was to passively listen and watch the stimuli that would be presented (see Turk-Browne et al.’s 2005 discussion, and section 4.2.3. of the previous chapter).

High-attention load groups were informed that this study was aimed at evaluating any possible interference effect of hearing an AL on performance in a visual detection task. They were asked to press the “1” key of the keyboard with their preferential hand as soon as they detected a consecutive repetition of pictures, independently of their orientation. The correctness of the visual detection task was emphasized. Thus, their attention was fully diverted from the AL stream that was presented as an irrelevant (potential disrupting) distractor while they performed an effortful independent visual task.

After this first phase, all participants performed the same forced-choice test. Procedure in this test phase, and for the low-attention load (mild cognitive noise condition in Experiment 4) in the reading phase, was identical to the one of the previous experiment (see section 4.2.3.).
5.3. Results and Discussion

5.3.1. RSVP Task

Performance in the visual task was only collected for participants in the high-attention load condition. Percentages of correct detection of repetitions (henceforth, *hits*) and of false alarms (FAs) were arcsine transformed and analyzed according to the familiarization time condition. For the groups that were presented with more than one block, performance remained stable across blocks, namely both across the two blocks presented to the 14-min group [*hits*: $t(16) = .15, p > .10$; FAs: $t(16) = .71, p > .10$], and across the three blocks presented to the 21-min group [*hits*: $F(2,28) = .91, p > .10$; FAs: $F(2,28) = 1.13, p > .10$]. Thus, a similar attentional load was instigated in each visual block within each familiarization-time group.

Furthermore, the three familiarization time groups showed equivalent performances [$F < 1$], with 58% (SE = 3.8), 57% (SE = 3.4) and 60% (SE = 4.0) average hit scores in the 7-, 14-, and 21-min groups, respectively. No differences were observed in the false alarm rates either [$F < 1$], with 4% (SE = 0.7), 5% (SE = 0.9) and 5% (SE = 0.9) average FAs, respectively. Interestingly, these hits and false alarm rates are remarkably similar to those reported by Toro et al. (2005).

FAs and hits were combined to calculate $d'$ scores separately for each subject in each familiarization time condition. Coherently to what was observed in the separate analyses of hits and FAs, the ANOVAs performed on $d'$ scores revealed neither a significant effect of block within the 14- [$t(16) = -.48, p > .10$] and 21-min [$F(2,28) = 2.19, p > .10$, $MSe = 0.138$] familiarization time groups, nor a significant difference between the three familiarization time groups [$F < 1$]. Moreover, for each
familiarization time condition the average $d'$ score was significantly greater than zero, with scores of 2.27 (SE = 0.17) [$t(13) = 16.18, p < .0001$], 1.95 (SE = 0.16) [$t(16) = 11.20, p < .0001$], and 2.07 (SE = 0.17) [$t(14) = 12.28, p < .0001$] in the 7-, 14-, and 21-min groups, respectively. Thus, participants were able to discriminate the situation in which a repetition occurred from the situation in which it did not, and they did so at the same success level in the three familiarization time conditions.

These results ensure that any effect of familiarization time on the ALL pattern cannot be due to the requirement to process a different number of blocks in the visual task. Indeed, with substantial training (14- and 21-min) the RSVP task might have become more automatic, therefore allowing more attentional resources to be allocated to the AL. This seems highly improbable, given that performance on the RSVP task was remarkably similar, not only in all familiarization time conditions, but also through blocks within the 14- and 21-min familiarization condition.

5.3.2. ALL

Average ALL performance broken down by TP-level, and local one-sample $t$-test\textsuperscript{20} comparisons with chance level are presented in Table 9.

In order to evaluate the impact of the available attentional resources and of the amount of familiarization time on TPs computation, we ran a mixed ANOVA on raw scores. Attentional load condition (low vs. high) and familiarization time (7- vs. 14- vs. 21-min) were entered as between-subject factors, and TP level (high- vs. low-TP-words) as a within-subject factor.

\textsuperscript{20} All the single sample t-tests are unilateral, testing whether ALL performance was significantly above chance in the forced choice task.
Table 9: Average proportions of TP-Word responses (in percentage), separately for each familiarization time (7-; 14-; and 21-min) and attentional load (low vs. high load) condition, considering the TP level of the AL words (high- vs. low-TP-words) in Experiment 5. Standard error of the mean for each condition is presented in parentheses.

<table>
<thead>
<tr>
<th>Familiarization Time</th>
<th>ATTENTIONAL LOAD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low Load</td>
</tr>
<tr>
<td></td>
<td>High-TP</td>
</tr>
<tr>
<td>7-minutes</td>
<td>68.0 (2.8)**</td>
</tr>
<tr>
<td>14-minutes</td>
<td>62.6 (2.7)**</td>
</tr>
<tr>
<td>21-minutes</td>
<td>57.3 (3.7)*</td>
</tr>
<tr>
<td>Average</td>
<td>63.1</td>
</tr>
</tbody>
</table>

One-way (unilateral) t-tests (in comparison to chance level, which corresponds to 50%):

* $p < .05$

** at least $p < .01$

* scores at chance level, $t < 1$, $p > .25$.

The analysis revealed a significant main effect of attentional load [$F(1, 85) = 7.4, p < .01, MSE = 5.4$]: overall, participants exposed to the AL in the high-attention load condition had lower performance than participants presented with the low-attention load condition. This is in line with Toro et al.’s (2005) results. No other main effect was significant [familiarization time: $F < 1$; TP level: $F(1, 85) = 1.2, p > .10, MSE = 4.14$]. The only interaction that was significant was the three-way interaction between all the factors studied [$F(2, 85) = 8.5, p < .001, MSE = 4.14$]. In order to analyze the TP gradient and any modulation of it by familiarization time, we evaluated each attentional load condition separately (see Figure 6).

In the low-attention load condition, neither familiarization time nor TP level were significant [$F \leq 1$ in both cases]. Thus, longer familiarization times did not increase the ALL magnitude on the whole. Instead, in the low-attention load condition the interaction between familiarization time and TP level was close to
significance \(F(2, 42) = 2.9, p = .07, MSe = 4.88\). Indeed, familiarization time seems to modulate the extraction of high- and low-TP-words differently. For high-TP-words, performance was better in the shortest (7-min) relative to the longest (21-min) familiarization time condition \(F(1, 42) = 6.02, p = .01, MSe = 4.54\), while no difference was observed between either the 7- and 14-min conditions \(F(1, 44) = 1.64, p > .10, MSe = 4.54\) or the 14- and 21-min conditions \(F < 1\). Extraction of low-TP-words was not modulated by familiarization time \((all Fs < 2.1, ps > .10)\).

Thus, with low attentional load, when familiarization time is drastically shortened (7-min) the system seems to privilege the extraction of high-TP-words. Coherently, it was only with 7-min familiarization time that performance was
significantly better for high- than for low-TP-words \[F(1, 42) = 3.7, p = .05\]. With increasing familiarization times, the system is able to extract both high- and low-TP-words at similar performance levels \[F \approx 1 \text{ with both } 14- \text{ and } 21\text{-min}\], but this procedure seems to carry with it a cost: this equivalent extraction of both TP-word types seems to be done at the expense of performance on high-TP-words. However, it is worth noting that performance was significantly better than chance for both high- and low-TP-words in all familiarization time conditions (see Table 9). Thus, with low attentional load, ALL is observed for each TP level, even with a short familiarization time of 7-min.

Familiarization time affected performance quite differently and much more dramatically in the high-attention load condition. Indeed, in this condition, the main effect of TP level only tended towards significance \[F(1, 43) = 3.5, p = .06, \text{MSe} = 3.42\], with high-TP-words somewhat more correctly extracted from the stream than low-TP-words, but this effect was strongly modulated by familiarization time \[F(2, 43) = 6.7, p < .005, \text{MSe} = 3.42\]. As a matter of fact, it was only with 7-min familiarization that participants performed significantly better than chance for both high- and low-TP-words (see Table 9), with no difference between their extraction \[F(1, 43) = 2.7, p > .10, \text{MSe} = 105.6\]. With 14-min familiarization, only high-TP-words were recognized better than chance (see Table 9). Yet, the difference between the recognition rates for high- and low-TP-words did not reach significance \[F (1, 43) = 2.2, p > .10, \text{MSe} = 3.42\]. With 21-min familiarization, the difference was more contrasted: as with 14-min, only high-TP-words were extracted significantly better than chance from the stream, but here performance was significantly better for high- than for low-TP-words \[F(1, 43) = 12.6, p < .001, \text{MSe} = 3.42\].
Thus, in the high-attention load condition performance for high-TP-words was better in the longest (21-min) condition relative to both shorter familiarization time conditions [21- vs. 7-min: $F(1, 43) = 5.58, p < .05, MSe = 3.56$; 21- vs. 14-min: $F(1, 43) = 5.47, p < .05, MSe = 3.56$], in which it was equivalent [$F < 1$]. Increasing familiarization time had an opposite effect on the extraction of low-TP-words: for these, performance was significantly worse in the shortest (7-min) familiarization time condition relative to the other two familiarization time conditions [7- vs. 14-min: $F(1, 43) = 5.2, p < .05, MSe = 5.86$; 7- vs. 21-min: $F(1, 43) = 4.4, p < .05, MSe = 5.86$]), in which performance for low-TP-words was equivalent [$F < 1$].

It thus seems that the reduction of attentional resources in conjunction with exposition time allows the perceptual system to adopt a differential procedure that aims at maximizing success: extraction of the “words” that have a higher probability (defined according to TPs) of being “lexical items” of the AL, at the cost of the extraction of words with lower TPs.

In line with Toro et al.’s (2005) data, one may thus conclude that speech segmentation based on statistical learning is severely affected by attentional load. In addition, we showed that this impact is a graded rather than an all-or-none phenomenon, since it affected differently the extraction of high- and low-TP-words.

Actually, there was a significant interaction between attentional load and familiarization time for both high- and low-TP-words [$F(2, 85) = 6.03, p < .005; MSe = 4.05$; and $F(2, 85) = 3.85, p < .025; MSe = 5.49$, respectively]. However, while for high-TP-words the performance disadvantage of the high- in comparison to the low-attention load condition decreases with increasing familiarization time (only in 7-min condition, performance was significantly better in the low- than in the high-
attention load group \( F(1, 85) = 8.31, p < .005; 14\text{-}min: F(1, 85) = 2.0, p > .10 \), for low-TP-words the performance drop linked to attentional load *increases* with familiarization time, being significant with the longer \( F(1, 85) = 10.11, p < .005 \) with 14-min and \( F = 4.61, p < .05 \) with 21-min] but not the shortest \( F < 1 \) familiarization times).

It is remarkable that a statistical learning effect was observed in the high-attention load condition, which was the case at least for high-TP-words. Indeed, in this condition TPs computation was not required (the visual task performance did not depend on it) and was not beneficial to the visual task that participants were intentionally performing, since the AL stream was presented to them as a potential harmful distractor. This suggests that at least some statistical learning may occur unintentionally, an idea that is coherent with former observations that statistical learning can occur without a specific effort from listeners to segment the continuous speech stream (Saffran et al., 1996a; 1997; Hauser et al., 2001).

Nevertheless, we cannot infer from the present experiment, what exactly is the cause of the detrimental impact observed on speech segmentation driven by TPs, and especially on low-TP-words extraction in the high-attention load condition. This effect may stem either from the fact that attention was diverted from the speech stream towards the visual task, or from the fact that this task was resource-consuming. In other words, we still do not know whether the detrimental effect observed in the high-attention load condition is linked to attention as a selection mechanism or as limited-capacity resources. These two facets of attention have been proposed as critical for the operation of sequence learning (see on the one hand Jiménez & Méndez, 1999; Turk-Brown et al., 2005, and, on the other hand, Shanks
et al., 2005; Rowland & Shanks, 2006a). To examine this issue as regard statistically-driven speech segmentation, in Experiment 6, we manipulated orthogonally both the focus of attention (i.e., selective attention) and the available attentional resources.
Experiment 6. Intention to learn in ALL – disentangling the impact of attention as selection and as resources in ALL.

Turk-Browne et al. (2005) did already suggest that selective attention to the to-be-learned stream is required to gate visual statistical learning. Indeed, these authors presented participants with two statistical sequences differing in a physical feature, by randomly intermixing a stream of red geometric shapes with a stream of green geometric shapes. Each color stream was constructed from four possible triplets, just as in the studies of Saffran et al. (1996). However, Turk-Browne et al. asked participants to detect shape repetitions in only one of the colors. In a surprise forced-choice test, triplets from both color streams were then pitted against foil sequences of three shapes from the same color that had never appeared in succession. No learning effect was found for the sequence that, although presented to participants, was not the focus of their attention. Selective attention dependence was also reported by Jiménez & Méndez (1999) in a SRT dual-task condition. Indeed, they only observed learning of a secondary contingency (i.e., between the shape of the present stimulus and the spatial location of the next) when participants were specifically asked to attend to that dimension.

If top-down guidance of selective attention to the to-be-learned stream was also critical for ALL, it may be the case that statistical learning could still operate optimally even in conditions of severely reduced attentional resources, as long as participant’s selective attention was focused on the statistical sequence, here, on the
In order to evaluate this hypothesis, in the present experiment we tested four groups of participants in a 21-min AL-familiarization setting, manipulating orthogonally the available attentional resources (low vs. high) and learning instructions. These instructions corresponded to either an incidental learning condition similar to the one used in the previous experiment and in Experiment 4 (see chapter 4), or to an intentional learning condition in which participants were asked explicitly to learn the material and were oriented to its sequential nature.

Thus, the two intentional learning groups were informed from the beginning of the experiment that the AL comprises a limited number of “words” and that their knowledge on those “words” was going to be evaluated in a subsequent phase. In fact, both intentional learning groups were presented with a dual-task situation, since they were asked to attend simultaneously to the visual and to the AL informations. The only difference between the low-attention and high-attention load intentional learning groups was that, during the first, exposition, phase, no overt task was required in the low-attention load condition. These participants were simply asked to attend to both the visual and auditory information, being informed from the beginning of the session that they would have to perform two subsequent tasks related to those materials (rotational preference for the visual material, and knowledge of the AL-words, respectively). In this way, while in the exposition phase no effortful operation was required from the participants of the low-attention load intentional learning group; instructions nevertheless pushed them to allocate their attention on both the visual and the auditory streams. On the contrary, the high-attention load intentional learning group, which was presented with the same
instructions as regards the AL, was asked to perform the visual repetition detection task during the exposition phase.

If selective attention to the AL stream was crucial for efficient statistically-driven speech segmentation, the cost linked to the requirement to perform the primary resource-consuming visual task would be overcome in the intentional learning condition. More precisely, since we know from Experiment 5 that participants are unable to extract low-TP-words in a 21-min high-attention load setting (see also section 4.3.2. for the results in the strong cognitive noise condition), we would predict that focusing attention on that stream would specifically enhance their performance on these low-TP words in the high-attention load condition.

5.4. Method

5.4.1. Participants

Forty-eight undergraduate psychology students at the University of Lisbon participated in the experiment for a course credit; none of them had participated in Experiment 1. They were monolingual EP speakers, with no reported history of speech or hearing disorders and normal or corrected-to-normal vision. Among them, 22 were randomly assigned to the condition of low attentional load (13 in the incidental learning condition, and 9 in the intentional learning condition) and 26 to the condition of high attentional load (13 in each learning condition).

5.4.2. Material

The three 7-min blocks of AL and visual items and the 36 trials of the forced-
choice task of Experiment 5 were used in the present experiment.

5.4.3. Procedure

All participants were exposed to the three AL- and visual-blocks used on Experiment 5, and hence to a total familiarization duration of 21-min.

The same procedure as the one of the previous experiment was used in the incidental learning condition (both for low- and high-attention load groups). The only difference regarded the instructions given to participants in the intentional learning conditions.

The intentional learning groups were informed from the beginning of the experiment that the AL comprises a limited number of “words” and that their knowledge on those “words” was going to be evaluated in a subsequent phase. In fact, both intentional learning groups were presented with a dual-task situation, since they were asked to attend simultaneously to the visual and the AL informations.

However, for the low-attention load intentional learning group, during the first phase, no overt task was performed. They were simply asked to attend to both the visual and the auditory information presented in the first phase, since they would have to perform two subsequent tasks about those materials. One was about their knowledge of the AL-words. The other was about their rotational preference for the visual material, since during the first phase their visual preference would be biased by the continuous presentation of rotated pictures.

The high-attention load intentional learning group was informed about the subsequent test on their knowledge of the AL-words, and was asked to perform an overt task during the first phase, namely the visual repetition detection task. Thus, in
the first phase, they should press the “1” key of the keyboard with their preferential hand, as soon as they detected a consecutive repetition of pictures, independently of their orientation. Response correctness in the visual detection task was emphasized, as well as the importance of attending to the AL in order to correctly perform the second phase task. In order to guarantee that participants would, in fact, be in a condition of high attentional load instigated by correct performance in the visual task, they were motivated to correctly perform it, with participants presenting the best performance in this task receiving a local bookstore’s voucher of 10 euros.

After the first phase, all participants performed the same forced-choice test as the one used in Experiment 5.

5.5. Results and Discussion

5.5.1. RSVP Task

Performance in the visual task was only collected for participants in the high-attention load condition. Percentages of hits and FAs were arcsine transformed and analyzed according to a mixed ANOVA with learning condition (incidental vs. intentional) as between-subject factor and block as within-subject factor. For both hits and FAs, neither the main effects nor their interaction were significant [all $Fs \leq 1$]. Thus, both groups showed equivalent performances in the picture repetition task, with average hit scores of 57% (SE = 3.8) and 55% (SE = 2.6) and average false alarm scores of 5.2% (SE = 1.0) and 6.8% (SE = 1.2) in the incidental and intentional learning conditions, respectively. Both groups were thus submitted to a similar
reduction of attentional resources available to statistical learning, independently of learning instructions.

FAs and hits were combined to calculate $d'$ scores, separately, for each subject in each learning condition. Coherently to what was observed in the previous analyses, there was no significant differences on the $d'$ scores ($t(24) = 1.15, p > .10$). Moreover, the average $d'$ score was significantly greater than zero, with scores of 1.95 (SE = 0.17) for incidental learning [$t(12) = 11.30, p < .0001$] and 1.70 (SE = 0.13) for intentional learning [$t(12) = 12.69, p < .0001$]. Thus, participants were able to discriminate the situation in which a repetition occurred from the situation in which it did not, and they did so at the same success level, independently of attending (intentional condition) or not (incidental condition) the AL stream.

The fact that intentional and incidental learners in the high-attention load condition presented similar performances in the RSVP situation, with similar levels of discriminability, ensures that any difference found in ALL pattern between the two learning conditions cannot be attributed to attentional load differences. Was this the case and we would have found significant differences between intentional and incidental learners in their ability to discriminate between situations in which a visual repetition occurred from the one in which it did not.

5.5.2. ALL

Average ALL performance broken down by TP-level are in Table 10.

In order to evaluate the impact of both attentional load (low vs. high) and learning conditions (incidental vs. intentional), we ran a mixed ANOVA with these two between-subjects factors and TP level (high- vs. low-TP-words) as within-
subjects factor. The main effect of learning condition was significant \( F(1, 44) = 4.6, p < .05, MSe = 4.48 \), since overall performance was significantly better in the intentional than in the incidental learning condition (see Table 10). Thus, it seems that the intentional learning instructions allowed participants to focus their attention on the AL stream, thereby endorsing a general performance gain in comparison to the incidental learning condition in which participant’s attention was diverted from the AL stream.

Table 10: Average proportions of TP-Word responses (in percentage), separately for each learning condition (Incidental and Intentional) and attentional resources condition (low-attention vs. high-attention load), considering the TP-level of the AL words (high- vs. low-TP-words) in Experiment 6. Standard error of the mean is presented in parentheses.

<table>
<thead>
<tr>
<th>Learning Condition</th>
<th>Attentional resources</th>
<th></th>
<th></th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low-attention Load</td>
<td>High-TP</td>
<td>Low-TP</td>
<td></td>
</tr>
<tr>
<td>Incidental</td>
<td></td>
<td>57.3 (3.6)</td>
<td>62.8 (3.3)</td>
<td>67.1 (3.6)</td>
</tr>
<tr>
<td>Intentional</td>
<td></td>
<td>65.4 (4.4)</td>
<td>70.4 (3.7)</td>
<td>67.1 (4.2)</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>61.3</td>
<td>66.6</td>
<td>67.1</td>
<td>54.7</td>
</tr>
</tbody>
</table>

However, the benefit afforded by focusing attention on the AL stream did not overrule the cost induced by the reduction of the available attentional resources. Indeed, as illustrated in Figure 7, there was a significant interaction between attentional load and TP level \( F(1, 44) = 8.7, p = .005, MSe = 6.77 \) that was not modulated by learning condition \( F < 1 \). As a matter of fact, a TP gradient effect (with worst performance for low- than for high-TP-words) was observed only in the high-attention load condition \( F(1, 44) = 9.5, p < .005 \). Actually, in this condition, only high-TP-words were extracted from the stream significantly better than chance.
\((t(25) = 6.3, p < .001\), while performance was at chance for low-TP-words \((t(25) = 1.7, p = .10\). In the low-attention load condition, no TP-gradient effect was found at all, since low- and high-TP-words were both extracted from the stream above chance level \([t(21) = 6.3, p < .001 \text{ and } 3.7, p < .005, \text{ respectively}]\) and at similar performance levels \([F = 1.4, p > .10]\).

Figure 7: Performance pattern (proportion of AL-word responses, in percentage) broken-down by TP level (High-TP-words; Low-TP-words) of AL words, according to attentional resources condition (Figure 7a: High-attention Load; Figure 7B: Low-attention Load) and learning condition (Incidental learning; Intentional learning) in Experiment 6. Vertical bars denote standard error of the mean on each condition. Chance level corresponds to 50%.

Thus, the observed cost on the extraction of the AL-units with lower TPs was exclusively found in the high-attention load condition. Moreover, orienting participants’ attention to the AL stream (in the intentional learning condition) did not overrule the cost instigated by a severe reduction of the available resources, since the same statistical learning impairment, with poor performance for low-TP-words, was
observed in both intentional and incidental high-attention load conditions.

In line with the ALL pattern observed in the 21-min familiarization condition of Experiment 5 and of Experiment 4 of the previous chapter (i.e., in single cue - strong cognitive noise condition), the present pattern of results show that the efficient operation of the statistical learning mechanism is not a matter of selective attention, but one of attentional resources. When a severe reduction of the available attentional resources is instigated by the need to perform another resource-consuming task, as in the high-attention load conditions of both experiments, statistically-driven speech segmentation operates in an impaired fashion. The present experiment shows in addition that this impairment occurs independently of the fact that participants are selectively attending to the AL (as in the intentional learning condition) or diverting their attention from it (as in the incidental learning conditions of both experiments). Thus, the TP-gradient effect observed in the high-attention load condition is a collateral damage of reduced available resources. In order to operate efficiently, the statistical learning mechanism requires (at least some) attentional resources. When this requisite is not fulfilled, TPs are still computed but in an impaired manner, leading to the inability of correctly extracting from the stream the AL-units with the lower TPs.

In summary, this experiment revealed three important facts: (i) learning conditions in which participants’ attention is directed to the AL stream promote an overall performance gain in statistically-driven speech segmentation; (ii) the reduction of the available attentional resources caused by the requirement to perform another resource-consuming task has a significant but graded impact on TPs computation, leading to the observation of a TP gradient effect, with impaired
extraction of low-TP-words; (iii) although attentional load seems to affect performance slightly more severely in the incidental than in the intentional learning condition, in conditions of scarce attentional resources it is not possible to eliminate this negative impact by merely orienting participants’ selective attention to the auditory stream.

5.6. General Discussion

The idea of a statistical segmentation procedure that is sensitive to temporal contingencies of the speech stream and that operates largely independently of awareness and control (and hence in an automatic fashion) is attractive, since this mechanism would allow the effortless acquisition of such a powerful cognitive ability as language, without the need to call on conscious (and controlled) strategies or processes.

In the present chapter, we evaluated the relation between attention(s), as a selection process and as a central resource, and statistical speech segmentation. Two previous studies have directly evaluated this relation.

Saffran et al. (1997) reported effective statistical speech segmentation in both adults and children in conditions in which attention was not explicitly directed towards the AL stream: participants were presented with “background” AL while performing a primary visual color-drawing task. However, although these findings suggest that statistical segmentation is incidental, they are not conclusive as to the possible role of attentional resources in speech segmentation by statistical learning. Merely instructing people to focus attention on a certain task is not sufficient to prevent distractor interference. In situations of low perceptual load (Lavie, 1995;
2005), any capacity not taken up in perception of task-relevant stimuli could involuntarily “spill over” to the perception of task irrelevant distractors (here, AL). Consequently, although Saffran et al. (1997) suggest that the incidental situation of their task prevented participants from focusing their attention on the speech stream, there is no actual guarantee that attentional resources were fully reduced for the TPs computation by simply performing a free drawing task.

Using a more stringent manipulation of attentional load, Toro et al. (2005) observed that reducing the available attentional resources had a detrimental impact on statistical learning: when attention was diverted to a difficult visual task (the detection of repetition of rapidly presented pictures), speech segmentation fell to chance level.

In the present chapter we aimed to address in a fine-grained manner the role of attention in statistically-driven speech segmentation, and hence the potential automaticity of this statistical learning mechanism. In particular, in this chapter we considered three goals: (i) to evaluate in a more fine-grained manner whether speech segmentation is inhibited or slowed down by a reduction of the available attention; (ii) to evaluate if attentional load has an absolute or a graded impact on speech segmentation driven by TPs; and (iii) to disentangle the role of attention as a selective mechanism and as a central resource in TPs computation.

In Experiment 5, we used the same design as the one used in Toro et al.’s (2005; Experiment 2), namely a visual concurrent task with a 500 ms SOA, but introduced one procedural and one material modification. First, instead of using a fixed duration of 7 minutes in the familiarization phase, as did Toro et al., in our experiment different participants were exposed to familiarization times of 7, 14, or
21 minutes. Second, instead of using a fixed TP of 1, we used AL words with high- and low-TPs (as already used in Experiment 4; see chapter four of this thesis).

The main conclusion to be drawn from that experiment is that although speech segmentation by statistical learning is able to proceed to some extent with minimal requirements on attentional mechanisms, its fully efficient operation depends on attention.

Indeed, in the high-attention load condition, independently of the amount of exposition to the AL, participants always succeeded in extracting the high-TP-words from the stream. However, it was only with 7-min familiarization that they were also able to extract the low-TP-words. With familiarization time, performance for high-TP-words actually increased, while performance for low-TP-words decreased, as though the cognitive system adapts itself to the reduction of attentional resources by focusing on the most salient “lexical” units.

This result’s pattern is very different from the one observed in the low-attention load condition. Here, the system seems to privilege the extraction of high-TP-words only when familiarization time is drastically shortened (namely, with 7-min). However, even in this condition, performance was above chance level for both low- and high-TP-words, as it was in the other familiarization time conditions. In addition, with increasing familiarization times, the system is able to extract both high- and low-TP-words at similar performance levels, although this procedure seems to be done at the expense of performance on high-TP-words.

Thus, attentional load affects statistical computation but in a graded, rather than in an all-or-none manner, since it modulates differently the extraction of high- and low-TP-words. Reducing attentional resources and/or diverting attention from
the AL does not completely prevent statistical learning, but slows down the extraction of the most salient units, while it progressively inhibits the extraction of the less salient ones. Indeed, the performance drop linked to attentional load, decreased with increasing familiarization time for high-TP-words, but increased with familiarization time for low-TP-words.

Our results cannot be attributed to the fact that with substantial training (14 and 21 minutes), participants simply became more efficient at task sharing: the visual (RSVP) task might have become automatic, therefore allowing more attentional resources to be allocated to the speech stream. This seems highly improbable given that performance on the RSVP task was remarkably similar, not only in all familiarization time conditions, but also through blocks within the 14- and 21-min familiarization time conditions. Moreover, had it been the case, we should have observed a stronger overall effect of attentional load in the longest familiarization time condition (i.e., 21 min). Instead, it was in this familiarization time condition that the impact of the reduction of the available resources was more severe, with only high-TP-words being correctly extracted from the stream, leading to the observation of a TP gradient effect. Importantly, the present results also suggest that the TP-gradient effect observed in Experiment 4 in strong cognitive noise condition when only statistical information was available in the AL-stream can not be due to a weaker instigation of attentional load than the one of Toro et al. (2005).

In Experiment 6, we evaluated independently the role of two attentional mechanisms: attention as a selection mechanism, and as capacity-limited resources. Attending to the AL and being informed about its structure and about the need to extract “lexical” items from it had a general benefit on ALL performance. This adds
to previous evidence from studies comparing intentional and incidental ALL conditions (e.g., Saffran et al., 1996b; 1997).

Nevertheless, focusing attention on the AL stream did not overrule the radical cost induced by a severe reduction of attentional resources. Indeed, in the high-attention load condition, only high-TP-words were extracted from the stream significantly better than chance, and this TP gradient did not depend on learning condition (incidental or intentional). Thus, even when participants were focusing their attention on the AL stream, statistical computations suffered a dramatic impairment when attentional resources were scarce. This suggests that statistical learning is an attentional resources dependent mechanism, as had been demonstrated in former implicit learning studies for sequence learning (Rowland & Shanks, 2006a; 2006b; Shanks et al., 2005).

However, intentional learning may involve other factors besides selective attention, such as hypothesis-testing, encoding strategies, and even a reallocation of attentional resources to an optimal operation of the processes underlying the task to perform. Nonetheless, the results found in the high-attention load condition of Experiment 6 suggest that a redistribution of attentional resources is unlikely. Indeed, independently of the learning condition (incidental or intentional learning) both high-attention load groups have achieved similar performances in the resource-consuming visual task. Thus, both intentional and incidental learners suffered the same drastic reduction of the available attentional resources.

The main impact of attention on speech segmentation driven by statistical learning thus seems to be a matter of capacity resources rather than of selection. Experiment 5 shows in addition that the impairment caused by scarce attentional
resources cannot be fully overruled by providing participants with more opportunity to learn, since longer familiarization times only favored the extraction of the most salient units (high-TP-words) of the AL stream.

It is important to emphasize that the impact of attentional load cannot be reduced to difficulty of the task *per se*, as already suggested in the previous chapter of this thesis (*see* section 4.4. of chapter four). Indeed, in Experiment 1 (*see* chapter two) we observed a very high resilience of the statistical segmentation procedure to physical noise (i.e., white noise superimposed on an AL stream, cf. Fernandes et al., 2007). Indeed, statistical information (i.e., TPs) was able to drive the segmentation process at similar levels in conditions of white noise superimposition (at 10 dB and 22 dB SNR) and in conditions of intact signal (in which the AL stream was presented with no noise superimposed). Moreover, in the strongly degraded condition (i.e., 10 dB SNR), statistical information overruled coarticulation when both cues were put in conflict in the AL speech stream. In the present chapter, as already shown in Experiment 4 (*see* chapter 4), we observed on the contrary a high sensitivity of the statistical segmentation procedure to the reduction of attentional resources, which might be considered as a kind of cognitive noise. The observation of different patterns of ALL as a function of the nature of the degradation instigated (i.e., physical or cognitive) rules out general task difficulty as an alternative account for the effects observed in the present study. Although manipulations of both physical (Fernandes et al., 2007; *see* chapter two) and cognitive (in the previous and in the present chapter of this thesis) noise increase general task difficulty, they clearly have independent and different impacts on statistically driven speech segmentation processes. The evaluation of the impact of attention on statistical learning addressed
in the present chapter sheds some light regarding the automaticity of this mechanism. The present results suggest that statistical learning (i.e., TPs computation) is at least partly an automatic process. Indeed, according to current approaches (e.g., Cohen et al., 2004; Logan et al., 1999; Moors & De Houwer, 2006; Tzelgov, 1997), not only different processes vary in the degree to which they rely on attention, but also automatic processes are conceived as being potentially sensitive to resource limitations (note that even Shiffrin & Schneider, 1977, already stated that interference between two automatic processes is inevitable when they use overlapping nodes in an incompatible fashion). In line with Tzelgov’s (1997) and Bargh’s (1992) proposals, an automatic process is one that is autonomous, running both without monitoring (it can be incidental) and, importantly, even when it is not part of participants’ task requirement. Thus, statistical learning seems to fulfill these criteria. In fact, in the present study, speech segmentation driven by statistical information was observed even in conditions in which processing the auditory stream was not required and not beneficial (being presented as potentially harmful) to performance in an independent task. In that sense, although statistically-driven speech segmentation is attentional-resources’ dependent, and thus is not a purely “stimulus-driven” mechanism (cf. Moors & De Houwer, 2006), it may nevertheless be conceived as automatic.

The present results may also help to shed some light on the nature of the processes that are responsible for statistical learning. Indeed, the observation of a TP gradient effect and the differential effect of attentional load as a function of TP level is difficult to reconcile with fragment-based models of implicit learning. For example, according to the PARSER model (Perruchet, 2005; Perruchet & Pacton,
2006; Perruchet & Vinter, 1998, *that will be approached in more detail in chapter 7 of the present thesis*), the knowledge acquired in statistical learning tasks might result from simple learning algorithms that give rise to little more than explicitly memorized short fragments or “chunks”, with sensitivity to statistical structure being a by-product of this process (*for a discussion see* Bonatti, Peña, Nespórl, & Mehler, 2006; Perruchet, Peereman, & Tyler, 2006; Perruchet, Tyler, Galland & Peereman, 2004). The chunks are formed from the outset on a random basis, as a natural consequence of the capacity-limited attentional processing of the incoming information, with each chunk corresponding to one attentional focus. As a consequence of their temporal proximity (in language or in any other temporal sequence), chunks become the constituents of a new representational unit, and their future depends on the laws governing associative learning and memory. If the association between primitives that form a percept is not repeated, the internal representation created by this percept rapidly vanishes, as a consequence of both natural decay and interference with similar material processing. If the same percept reoccurs, the internal representation is progressively strengthened. Thus, the probability of repeatedly selecting the same group of syllables is higher if these syllables form intra-word rather than between-words components. This fairly simple learning algorithm provides an alternative mechanism that can explain some of the data showing that listeners are sensitive to statistical information in speech segmentation (e.g., Perruchet et al., 2004).

The observation of a TP-gradient effect had already been demonstrated by Saffran et al. (1996b) in an intentional ALL condition. It was also observed here in both Experiments 5 and 6 (and also in the previous experiments 3B and 4 in
particular conditions): listeners showed better performance on AL-words with the higher TPs (from 0.75 to 1) than on AL-words with the lower TPs (from 0.37 to 0.50). PARSER failed to reproduce this aspect of the empirical data (Perruchet & Vinter, 1998). However, Perruchet and Pacton (2006) argued that taking into account the effect of interference promoted by forgetting (since the sequential material used in statistical learning is typically generated by recombining a reduced number of primitives, e.g., syllables) is sufficient to make chunk strength sensitive to TPs. Yet, the sensitivity to TPs observed by Perruchet and Pacton (2006) was based on the different raw frequencies of the TP-words: half had a higher frequency of occurrence than the others. Here, we observed a TP gradient effect even though high- and low-TP-words had the same frequency of occurrence in the AL, as was already the case in Saffran et al.’s (1996b) study. This poses some problems to Perruchet and colleagues’ interpretation.

The differential impact of familiarization time on high- and low-TP-words (as in Experiment 5 of the present chapter) also suggests that statistically-driven speech segmentation cannot be confined to chunk formation. PARSER is actually compatible both with the occurrence and with the absence of an overall performance gain with familiarization time. Indeed, if statistical learning was governed by general laws of associative learning and memory, the longer the exposition time, the stronger the AL-units extracted and the higher the interference will be. Therefore, the longer the familiarization time, the stronger the AL units extracted, but at the same time, the higher the interference between different AL units extracted from the stream. Between these positive (i.e., strengthening) and negative (i.e., interference) impacts, learning will not be linearly incremented. However, if an effect of familiarization
was observed, PARSER would predict that it would be general: both high- and low-
TP-words, which had the same raw frequency, would have the same cost/benefit of
familiarization time. This was clearly not the case here. Thus, the present results
suggest that participants’ sensitivity to statistical information is probably based on
distributed knowledge that is represented as subsymbolic information (Cleeremans et
al., 1998; Cleeremans & Jiménez, 2002).
Chapter 6.

The Metamorphosis of the Statistical Segmentation Output:
On-line measures and Lexicalization during Artificial Language Learning\textsuperscript{21}.

The ALL paradigm is a promising tool in spoken word segmentation. However, previous studies did not attain two central aspects of ALL: direct evidence of on-line statistical segmentation; and the potential lexical status of statistical learning outputs. With these aims, in the present chapter, we combined ALL settings with conventional experimental techniques. In Experiment 7, a variant of the word-spotting task was adopted, demonstrating that adult listeners are able to use on-line statistical information. In Experiments 8 and 9, we evaluated whether the output of ALL exhibits a lexical competition signature revealed by the inhibitory priming effect of novel neighbors (e.g., *cathedruke*) on lexical decisions to real words (e.g., *cathedral*). Listeners actually treated statistical learning outputs as potential words, but only when segmentation cues were congruent, as was the case in Experiment 8. In Experiment 9, no effect of lexicalization was observed with incongruent cues, although listeners’ ALL performance in the forced-choice test differed from chance, suggesting a dissociation between AL-items configuration and AL-items lexical engagement. These results suggest that current models of spoken-word recognition need to incorporate the role that congruency between segmentation cues play in both speech segmentation and word learning.

\textsuperscript{21} The experiments presented in this chapter were integrated in a manuscript in preparation (Fernandes, Kolinsky, & Ventura, in press).
6.1. **Introduction**

Combining lexical-driven mechanisms to any one of the signal-derived cues may be helpful to word segmentation. This, however, holds true only when the speech signal can map onto stored representations. Remarkably, that seems to be the case early on language acquisition (e.g., Bortfeld et al., 2005; Newman, Ratner, Jusczyk, Jusczyk, & Down, 2006; Swingley & Aslin, 2007). Thus, a compromise between lexically-driven and signal-derived informations seems to be established throughout linguistic development.

In this context, the ALL paradigm (Saffran et al., 1996a; 1996b) has proved to be useful regarding the role of sublexical cues in speech segmentation (see section 1.3. of chapter one of the present thesis). During the familiarization phase of an ALL study, listeners are usually exposed to a continuous artificial language (AL) stream constituted by a limited repertoire of concatenated syllables. In many studies, statistical information is the only available cue that can help listeners to locate word boundaries, since the TP between adjacent syllables is higher within AL words (henceforth, *TP-words*) than between them. This situation partly mimics natural speech, in which generally syllables that follow one another within a word have a higher TP than syllables that straddle word boundaries.

Demonstration that listeners can parse TP-words from the continuous speech stream through a statistical segmentation procedure is generally provided by their preference for TP-words over either *nonwords* (constituted by the syllables of the AL with TP 0, since they did not occur adjacently in the familiarization phase) or *part-words* (constituted by syllables that occurred in the stream as adjacent parts of
different TP-words and hence presenting lower TPs) in a subsequent two alternative forced-choice test.

The reliability of ALL is demonstrated by numerous replications throughout linguistic development (e.g., Johnson & Jusczyk, 2001; Saffran et al., 1996a; Saffran et al., 1996b; Vroomen et al., 1998) in different linguistic backgrounds. TP extraction is furthermore considered as a basic process in language acquisition. Indeed, even if there were reliable acoustical cues to word boundaries, some minimal lexical knowledge would be a prerequisite for their efficient use (Perruchet & Vinter, 1998). This is not the case of TP computation, which does not require even minimal lexical knowledge and hence could act as the “first window into the regularities of [infants’ native language]” (Thiessen & Saffran, 2003, p. 716), with “a pivotal position in the acquisition of not only words, but also other word boundary cues” (Mattys et al., 2005, p. 493). In other words, for infants the statistical outputs could act as “stem cells” for lexical acquisition.

For sure, TPs computation plays a less fundamental role in language processing by adults, in comparison to infants. Nevertheless, adults are still sensitive to statistical information such as conditional probabilities, not only in speech processing (e.g., Saffran et al., 1996a; Vroomen et al., 1998), but also in other sequence learning tasks (e.g., Boyer et al., 2005; Conway & Christiansen, 2005; Fiser & Aslin, 2001; Perruchet & Pacton, 2006). Indeed, since the intensely dynamic environment that surrounds us mainly involves stimuli occurring in temporal and/or spatial sequences, statistical learning would enable adults to rapidly structure novel sequences of events into emergent units (i.e., into sets of temporally or spatially contiguous events). More generally, the process of reduction of uncertainty (Gibson,
1991) drives learners to seek invariant structure in the stimuli array (Gómez, 2002).

However, two main aspects at the core of ALL effects still remain obscure. First, no direct demonstration of on-line statistical speech segmentation has been provided up to now. Second, whether the statistical segmentation outputs are actually lexicalized is still unknown, at least for adults. Shedding light on these aspects in the context of a full mature speech perception system was the aim of the experiments described in the present chapter.

As regards the first point, it is worth noting that the ALL effect revealed by listeners’ above-chance preference for TP-words over AL nonwords provides only indirect evidence of speech segmentation processes, since it is based on an off-line measure derived from performance in a subsequent test. Except for some reaction time (RT) studies on adults’ sensitivity to artificial phonotactic constraints\textsuperscript{22} that have demonstrated that participants rapidly acquire novel phonotactic regularities through the laboratorial listening experience (Onishi, Chambers, & Fisher, 2002; Warker & Dell, 2006; see also McLennan & Luce, 2004; McLennan, Luce & LaVigne, 2004), on-line behavioral measures have seldom been collected in ALL, and, to our knowledge, never as regards TPs. In addition, both forced-choice tests and RT studies used isolated, already segmented stimuli and thus do not ensure that test stimuli indeed correspond to listeners’ segmentation by-products, since the AL potential “lexical-units” were given beforehand.

As regards the second point, neither the forced-choice tests nor the aforementioned RT studies inform us about the degree of lexicalization of the statistical segmentation outputs. Once segmented from a continuous speech stream, listeners

\textsuperscript{22} These artificial phonotactic constraints are not the same as adjacent TPs (generally used in ALL settings). However, both are defined by a conditional probability between two (adjacent or non-adjacent) linguistic units, although in the former case between segments (e.g., diphones) rather than syllables.
might continue to treat those outputs as relative coherent sound sequences with high internal probabilities, but with no particular status regarding their native language. Alternatively, TP units may be treated as world-like units, namely as potential “words” available for lexical integration.

6.1.1. The nature of statistical segmentation outputs

In infants, two studies suggested that statistical learning generates novel word-like units, being rapidly available for integration in their linguistic knowledge. Saffran (2001) showed that, immediately after an AL familiarization period with only statistical information available, eight-month-olds only demonstrated preference for TP-words over part-words when these were embedded in linguistic familiar frames (e.g., I like my…[AL stimulus]), but not when embedded in either nonsense linguistic frames (e.g., zy fike ny…[AL stimulus]) or in nonlinguistic ones (e.g., in sequences of tones). In an ALL setting, Estes et al. (2007) also observed that 17-month-olds were only successful in a subsequent object-label learning task when labels corresponded to TP-words, but not when these were either nonwords or part-words. It thus seems that for infants word segmentation and word learning are directly linked in ALL, as they are in natural language. Indeed, Newman et al. (2006) showed that children who displayed better speech segmentation abilities during the first year of life (i.e., between 6 to 12 month of age) reached higher scores in language measures (e.g., in expressive vocabulary) the following year, when they were 2-years-old. Whether the output of statistical segmentation also becomes fully

---

23 Here the term ‘word’ loosely refers to phonological representations that are available for subsequent language processing (see Brent & Cartwright, 1996).
lexicalized for adults is far from clear. Indeed, contrary to infants, adult listeners already use multiple speech segmentation cues and have acquired a consolidated mental lexicon.

Recent work on word learning by adults has shown a dissociation between, on the one hand, the acquisition of a novel *item configuration* (Leach & Samuel, *in press*), namely of the word’s input form (phonological and orthographic) and associated meaning and syntactic role(s), and, on the other hand, its *lexicalization* (Gaskell & Dumay, 2003a; 2003b) or *lexical engagement* (Leach & Samuel, *in press*), namely the item’s involvement in lexical dynamics. In fact, learning new words does not always produce fully functional lexical representations: certain information about new words may be represented, but these representations may not achieve some criterion needed for behaving as “true” lexical items (Dumay & Gaskell, 2007; Dumay, Gaskell, & Feng, 2004; Gaskell & Dumay, 2003; Leach & Samuel, *in press*).

While in adults item configuration may be attested through recognition tests (using either off-line or on-line measures), according to Dumay and Gaskell (2007; Dumay et al., 2004; Gaskell & Dumay, 2003) and to Leach and Samuel (in press), lexical engagement is best attested by measuring the ability of a lexical representation to affect the activation of other representations. Since lexical items’ dynamic behavior is in essence reflected by their engagement in lexical competition, Gaskell and Dumay, 2003a; 2003b) conceived a strong proof of lexicalization. Through a phoneme-monitoring task, Gaskell and Dumay (2003a) first familiarized adults with lists of auditory pseudowords that strongly overlapped with existing words, like *cathedruke* that was derived from *cathedral*. Using both lexical decision
and pause detection tasks, they then checked whether these novel items (e.g., *cathedruke*) inhibit recognition of already existing lexical items (*cathedral*) that are members of the same cohort of candidates since they share their phonological onsets (cf. Marslen-Wilson, 1987). After familiarization through 12 presentations of each novel item, such an inhibitory effect progressively emerged over the course of five days (Gaskell & Dumay, 2003a; Experiment 2). With 36 presentations of the novel items, a similar effect was also observed one week after familiarization, with no intervening new exposure period (Gaskell & Dumay, 2003a; Experiment 3). The inhibitory priming effect also occurred after a 12-hour interval, as long as it included one night’s sleep (Dumay & Gaskell, 2007), suggesting that lexicalization possibly requires an *incubation period*.

Exposure to the phonological form of novel items that are close variants of existing words appears to be the only necessary and sufficient condition for lexicalization to occur. Indeed, the phoneme-monitoring task used by Gaskell & Dumay (2003a; 2003b; Dumay & Gaskell, 2007) only provides a purely phonological context. Similarly, in visual word recognition exposing participants to novel orthographic neighbors (e.g., *banara*) in a typing task, in which they were only instructed to correctly type those items, had an inhibitory impact on subsequent semantic classification of existing lexical items like *banana* (Bowers, Davis, & Hanly, 2005). In addition, providing a semantic context to spoken novel neighbors (i.e., in a sentence verification task, with novel items being assigned a semantic category and a meaning) does not seem to result in either faster or deeper lexicalization than providing listeners with purely phonological contexts, like in phoneme-monitoring (Dumay et al., 2004; for similar evidence with 1.5-year-olds,
see Swingley, 2007).

6.1.2. An overview of the experiments presented in this chapter

In the experiments presented in this chapter, we adopted an ALL approach combined with experimental designs used in research on natural language processing, in order to both obtain evidence of on-line statistical speech segmentation and evaluate whether statistical segmentation outputs acquire some lexical status in the adult speech system.

In Experiment 7, we checked whether adult listeners were actually able to use statistical, TP-based, information in an on-line segmentation task in which no a priori segmented “lexical” solution was provided. To this aim, listeners were first familiarized with an AL with only statistical segmentation cue (TPs) available, and then had to perform a variant of the word-spotting task (Cutler & Norris, 1988) in which they were required to detect any AL word embedded in four-syllabic nonsense strings.

In Experiments 8 and 9, we evaluated whether statistical segmentation outputs are lexicalized, adopting the approach of Gaskell and Dumay (2003a; 2003b) within an ALL setting. In both experiments, two types of segmental information were available in the stream during the familiarization phase, namely the statistical information afforded by TPs between the adjacent syllables of the AL (these TPs being always higher within the trisyllabic TP-words than between them) and the wordlikeness of the trisyllabic TP-words. The latter was estimated by taking into account both the onset overlap between TP-words and real words of the listeners’ native lexical entries (as was the case for novel words in Gaskell and Dumay’s
studies) and phonotactic information computed on the listeners’ native language lexicon.

Since former work has shown that adults listeners tend to integrate the output of statistical learning with knowledge of their native language acquired prior to the experimental task, be it knowledge of coarticulation (Fernandes et al., 2007), of prosodic properties (Shukla et al., 2007; Vroomen et al., 1998), or of phonotactic probabilities (Onnis, Monaghan, Richmond, & Chater, 2005; see also Perruchet, Tyler, Galland, & Peereman, 2004), we expected both of these cue types (i.e., the AL statistical information and phonotactics) to play some role in ALL. As regards lexicalization, it has been suggested that it may be easier for the system to develop lexicalized representations for novel items that are close variants of existing words than to build such representations from scratch. As a matter of fact, with novel items that had no such close relatives in the lexicon, training without meaning (via phoneme-monitoring) led to very weak lexicalization effects, even after almost 50 training trials, while training with meaning (either picture association, or full stories) produced much stronger evidence of lexicalization (Leach & Samuel, in press). However, in order to check whether wordlikeness per se would be able to account for lexicalization of AL stimuli, we manipulated the congruency of the two types of cues, using congruent cues in Experiment 8 and incongruent cues in Experiment 9.

More specifically, in Experiment 8, participants were familiarized with one of two ALs, each one being constituted by trisyllabic TP-words that differed from real words (hereafter, base-words) by their last syllable, after the base-words’ uniqueness point (UP). Thus, TPs and the wordlikeness of TP-words suggested the same segmentation outputs. In Experiment 9, participants were familiarized with one of
two new ALs constituted by the same phonological repertoire as the one used in Experiment 8, but this time it was the AL part-words that differed from the base-words after the UP. Thus, in Experiment 9, while statistical information of the AL suggested that TP-words were the “units” of the new language, the wordlikeness of AL stimuli pointed to the part-words as the potentials “units”.

In the second phase of both Experiments 8 and 9, listeners performed a conventional ALL test (a two alternative forced-choice test) plus a lexicalization test, namely a lexical decision task on base-words. Since previous work has suggested that the full engagement of novel items in lexical competition is rather slow (Gaskell & Dumay, 2003a), listeners were presented with these two tests twice, once immediately after familiarization, and the second time one week later, with no intervening new AL familiarization period.
Experiment 7. On-line statistically-driven speech segmentation

In order to evaluate directly, on-line listeners’ adoption of statistical speech segmentation, in Experiment 7, we combined the ALL setting with a variant of the word-spotting task designed by Cutler & Norris (1988).

In the word-spotting task (see also section 1.2.2. of chapter one of this thesis), listeners are asked to detect any real word that is embedded in nonsense (pseudo-continuous) strings, which gives the task some “ecological validity” (McQueen, 1996). Crucially, word-spotting has often been used to analyze the role of various cues in segmenting natural speech, like metrical prosody (e.g., strong syllables, in English: Cutler & Norris, 1988; lexical stress, in Finish: Vroomen et al., 1998), phonotactics (McQueen, 1998), the PWC (Norris et al., 1997), vowel harmony (Vroomen et al., 1998) and syllable onsets (Dumay et al., 2002).

In the present experiment, listeners were first familiarized with an AL with only TPs available in the stream as cues to word-boundaries. In a second phase, they performed a variant of the word-spotting task, in which they were asked to detect any AL “word” embedded in four-syllabic nonsense strings. The critical strings included either a TP-word or a part-word; each one of these two AL stimulus types could be embedded either at the beginning (henceforth, initial carriers) or at the end (henceforth, final carriers) of the four-syllabic carrying string. For example, the TP-word /fufibu/ appeared once in /fufibuRõ/ and once in /Rõśfufibu/. As illustrated, in addition to including either a TP-word or a part-word, critical strings included a
residual CVC or CCV nonsense syllable. The choice of residual syllables that had a different phonological structure than the one of the AL syllables (that were always CV) aimed at guarantying that participants would be able to carry out the spotting task, since it was previously demonstrated that they have difficulty to explicitly identify AL stimuli (e.g., Sanders, Newport, & Neville, 2002). While the use of such residual strings leaves room for strategies in spotting performance, such strategies would however be unable to account for differences (in RT and/or in accuracy) between TP-words and part-words, since both types of AL stimuli were constituted by the same phonological repertoire and had occurred in the stream during the first, familiarization phase. If listeners actually used the statistical information provided by TPs in this on-line segmentation task, they would both spot more often and identify more rapidly TP-words than part-words.

6.2. Method

6.2.1. Participants

Fourteen undergraduate psychology students at University of Lisbon participated in the experiment for a course credit. All were monolingual EP speakers, with no reported history of speech or hearing disorders.

6.2.2. Material

All speech stimuli were synthesized as in the previous experiments, using MBROLA software (Dutoit et al., 1996) and an EP female diphone database.
The same AL previously used was also adopted in the present experiment (see section 2.2.2. of chapter two and Appendix I).

For the familiarization phase, the same three 7-minutes listening blocks (rendering a total familiarization time of 21-minutes) used in the single cue version of Experiment 2 (see section 3.3.2. of chapter three of this thesis).

For the word-spotting phase, the six TP-words and six part-words (the same part-words used in the previous experiments of this work; see Appendix I) were embedded into four-syllable nonsense strings (see Appendix IV). The four-syllabic nonsense strings were synthesized by concatenating the three syllables of each AL stimulus and a residual CVC or CCV syllable. Those residual syllables were used in all carrier types, namely in four-syllabic nonsense strings including either a TP-word or a part-word as well as in filler items, that did not contain any AL stimulus.

Since in ALL studies the repertoire of “words” is always reduced, as it is in the present experiment, to gain statistical power on the subsequent analyses of word-spotting performances, each one of the 12 AL stimuli (6 TP-words and 6 part-words) was presented in two different carriers (i.e., target-bearing strings), with each string presented in different blocks, guarantying a long lag between the two presentations of the same AL-stimuli. Furthermore, the AL stimulus was in a different position within the two carriers: it was embedded at the onset in initial carriers (e.g., /fubifu/ in /fubifuRõ/) and at the end in final carriers (e.g., /fubifu/ /Rõfubifu/), rendering 24 carriers. None of the residual strings contained any real word.

Two experimental lists of 12 test-items each (6 TP-words- and 6 part-words-bearing strings; see Appendix IV) were constructed, one with initial carriers, the other with final carriers. Each list also included 24 filler items that did not contain
any AL stimulus (or any EP word), but had the same structure as carriers, namely a trisyllabic CV sequence plus a CVC or CCV syllable.

A 16-items practice list preceded each experimental list. In each practice list, four target-bearing items included an embedded Portuguese-word (e.g., /muzikə/ - music, in /muziktrə/ and in /trəmuzikə/). The remaining items did not contain any real word or any AL-syllable.

6.2.3. Procedure

Participants were individually tested in a sound-attenuated room and informed that the experiment had two phases. In the familiarization phase, all stimuli were presented at a comfortable listening level through Sennheiser HD 280 Professional Silver headphones, using Windows Media Player and Sound Blaster Audigy 2 ZS Audio. Participants were instructed to listen to an AL and try to find out what words constituted it. No information about structure, phonology or length of the AL-words was given. The three 7-minutes listening blocks were separated by two 5-minutes breaks. After the AL-familiarization, in the word-spotting phase, stimuli were also presented through headphones, with presentation, timing and data collection controlled by E-Prime 1.1 (Schneider et al., 2002a, 2002b) with an ISI of 6.5 seconds. Participants were told that on each trial they would hear a warning tone followed by a nonsense string and in some of the strings a word from the AL heard in the first phase would be embedded. They were also informed, before each experimental list, about the position of the potential AL-word within the string (at the beginning for Preceding List and at the end for Following List), as done by McQueen.
et al. (1994; Experiment 2) and Norris et al. (1995). Their task was to attempt to spot the AL-words embedded in the nonsense strings, and to say them aloud as rapidly and accurately as possible. They were asked to answer on all trials: If they did not spot any AL-word they had to say “nada” (nothing). The vocal response triggered a voice key connected to a button box (PST SRB 200A), with vocal latencies measured from the onset of bearing sequences to the triggering of the voice key by participants’ response. These were also registered with a Sony TCM-200DV recorder and noted by the experimenter. Before each experimental list, the corresponding practice list was provided to familiarize participants with the task. They were informed that the practice-targets were real-words, and feedback was only provided for these practice lists.

Both experimental lists were presented to all participants, with order of Preceding and Following list counterbalanced. Order of test-trials within each list was randomized for each participant.

6.3. Results and Discussion

Spotting responses and RTs were analyzed separately, with subjects as the only random factor (Raaijmakers, 2003; Raaijmakers, Schrijnemakers, & Gremmen, 1999).

6.3.1. Spotting Responses

Participants correctly discarded filler items, with 99.85% correct “nothing” responses, on average.
Responses on carriers were analyzed to determine the proportion of spotted trials, and to check next which, among them, corresponded to fully (trisyllabic) or partially (disyllabic) correct identifications of TP-words and part-words.

Strategic procedure could underlie participants’ performance and hence they could correctly spot both types of AL stimuli merely by chance. Indeed, participants were informed that some strings contained an AL stimulus at a particular edge (chance level: 50%), and their response could be constituted by one, two, or three syllables (i.e., chance level: 33.33%). Consequently, 16.67% (i.e., 0.50 x 0.33) corresponds to the probability of correctly detecting a trisyllabic AL stimulus by chance.

Participants spotted, on the average, 77.97% of the TP-words. Among these responses, 60.71% (SE = 5.9) were fully correct trisyllabic TP-words, a result significantly above chance level \((t(13) = 7.48, p < .001)\). Although 17.26% (SE = 4.7) of the responses were disyllabic parts of TP-words (with an average TP of .70), this performance did not differ from chance level \((t(13) < 1, p > .10)\).

Participants spotted 58.32% of the part-words; 33.33% (SE = 7.7) were fully correct trisyllables, which is above chance level \((t(13) = 2.15, p = .05)\). As already observed for TP-words, for part-words the disyllabic responses (24.99% of the responses; SE = 6.3) were at chance level \([t(13) = 1.43, p > .10]\). These disyllabic responses were of two types, since part-words included two syllables of the same TP-word and one syllable of an adjacent TP-word. Identification of two syllables belonging to the same TP-word were more numerous than responses formed by two syllables belonging to different TP-words [15.47 vs. 9.52%, respectively; \(t(13) = 2.10, p = .05\)], as expected on the basis of the TPs of these disyllabic fragments, since
the average TP of the former type was higher than the one of the latter (.75 and .10, respectively).

In agreement with these observations, the 2x2x2 ANOVA on spotting performance, with AL stimulus type (TP-words; part-words), AL stimulus embedded position (initial vs. final carriers) and response type (trisyllabic; disyllabic) as within-subjects factors, revealed a main effect of AL stimulus type \( F(1, 13) = 16.41, p < .005, MSe = 0.037 \], with TP-words spotted more frequently than part-words. There was also a main effect of response type \( F(1, 13) = 13.47, p < .005, MSe = 0.1914 \], and the interaction between these two factors was significant \( F(1, 13) = 6.76, p < .05, MSe = 0.071 \]. Indeed, trisyllabic fully correct identifications of entire AL stimuli were more numerous for TP-words than for part-words \( F(1, 13) = 14.52, p = .005 \], while there was no difference between the two AL-stimulus types as regards disyllabic responses \( F < 1 \). No other effect or interaction was significant (all \( Fs < 2, ps > .10 \).

This pattern of results suggests the use of TPs as a segmentation cue, with TP-based trisyllabic word-like units as its product. The extraction of these units occurred on-line without any a priori presentation of the plausible AL “lexical units”. As a matter of fact, only trisyllabic responses for both types of AL stimuli were above chance level, with TP-words spotted more often than part-words.

This is the first direct evidence that listeners are able to correctly extract from a longer sequence the trisyllabic stimuli that corresponded to the AL “words” in the familiarization phase, even if they were never presented with these AL stimuli as already segmented “lexical” units. Furthermore, even fragment responses obeyed to the TP gradient, since identification of two syllables belonging to the same TP-word
were more numerous than responses formed by two syllables belonging to different TP-words.

6.3.2. RT data

Raw RTs, recorded from sequence onset, were adjusted, in order to be measured from AL stimulus offset by subtracting the duration either of the AL stimulus (for initial AL stimulus carriers) or of the complete four-syllable string (for final AL stimulus carriers), as usually done in word-spotting studies (cf. McQueen, 1996). Since only trisyllabic responses corresponded to a fully correct identification of AL stimuli, and were also the ones spotted above chance, they were the only response type considered in this RT analysis.

The mean RTs observed in the present experiment were quite slow in comparison to conventional word-spotting studies. Indeed, even when adjusted RTs were considered, average RTs were well above 1000 ms. This probably reflect the fact that, contrary to conventional word-spotting studies, in the present experiment listeners were asked to detect AL “words” that could only be parsed from the carrier strings on the basis of statistical (TPs) information, and hence that did not benefit from associated orthographic, semantic, or syntactic information.

Much more important is the fact that in a situation in which AL units were not presented ever as already segmented stimuli, participants correctly detected the trisyllabic stimuli much faster when these corresponded to “words” of the new language (i.e., to TP-words) than when they were part-words.

As a matter of fact, the 2x2 ANOVA on RTs with AL stimulus type and embedded position as within-subjects factors revealed a significant main effect of AL

226
stimulus type \[ F(1, 13) = 4.81, p < .05, MSe = 167728 \], with responses to TP-words on average 240 ms faster than to part-words (see Figure 8), which is line with the pattern of spotting responses found.

Figure 8: Mean Adjusted RT (in ms; from AL-stimuli offset) for trisyllabic responses in spotting TP-words and part-words, separately for initial and final AL-stimuli-bearing sequences (in the Preceding vs. Following lists, respectively), in Experiment 7. The bars denote the standard error of the mean in each condition.

The main effect of embedded position was also reliable \[ F(1, 13) = 9.12, p < .01, MSe = 142528 \], with faster responses to AL-stimuli embedded in initial than in final carriers. While frequently observed in word-spotting, although not in a consistent direction (with different effects for mono- and disyllabic words, e.g., Dumay et al., 2002; Norris et al., 1997), this effect is irrelevant to the aim of the present experiment, since it did not interact with AL stimulus type \[ F < 1 \].

Thus, the RT data are in line with the pattern of spotting responses. While both types of AL stimuli occurred in the AL stream during the familiarization phase and were only distinguishable on the basis of their TPs, listeners were more accurate
and faster at spotting TP-words than part-words. This lends support to the notion that listeners are able to use statistical speech segmentation cues on-line: using TPs, they have acquired new items configurations that they are able to parse on-line in nonsense pseudo-continuous strings.

Still, whether the by-products of the output of this on-line statistical segmentation process become involved in lexical dynamics remains far from clear. This is what we tried to evaluate in Experiments 8 and 9, combining an approach similar to the one of Gaskell and Dumay (2003a; 2003b).
Experiment 8. The status of statistical segmentation

output: When different segmental congruent cues are available in the AL-stream.

For Experiment 8 we created two new ALs. In each of these ALs, both the TPs between the adjacent syllables of the AL and the wordlikeness of the TP-words in the listeners’ native language were available in the stream and were congruent, since they suggested the same “word” boundaries. Indeed, TP within TP-words was higher than between them, and TP-words were also much more “word-like” than part-words.

As a matter of fact, TP-words strongly overlapped from onset with lexical entries in the listeners’ native language, here, European Portuguese. This manipulation of wordlikeness is thus similar as the one used in Gaskell and Dumay’s (2003 a; b) studies. For example, the TP-word /ɡəzəmə/ used in the familiarization phase was created from the base-word “gasoso” /ɡəzəzu/ (meaning “gaseous” in Portuguese). Wordlikeness was further estimated by taking into account phonotactic probabilities computed on the listeners’ native language lexicon, since this information constitutes a unique and important factor of wordlikeness (Bailey & Hahn, 2001; Frisch, Large, & Pisoni, 2000) to which listeners are highly sensitive (e.g., Mattys et al., 2005; McQueen, 1998; Vitevitch & Luce, 1998; 1999; 2005; for a review see also Auer & Luce, 2005). Here, we evaluated the TPs between the adjacent syllables that constituted the AL repertoire by considering TPs within real
words in EP, as TPs of diphones in listeners’ native language are represented prelexically, independently of lexical information (Pitt & McQueen, 1998; see also Vitevitch & Luce, 1998; 1999; 2005).

Congruent sources of information, namely the availability of different cues suggesting the same segmentation hypotheses, has a beneficial impact on speech segmentation, promoting a redundancy gain in ALL in comparison to a single-cue condition (e.g., Fernandes et al., 2007; Vroomen et al., 1998). We thus expected to find a strong ALL effect, suggesting that AL statistical information and wordlikeness of TP-words jointly assisted speech segmentation.

However, whether the segmentation outputs were lexicalized could only be observed in a subsequent lexicalization test, here a lexical decision task. If the output of AL segmentation were actually lexicalized, it would have a lexical competition signature, inhibiting the recognition of already existing lexical entries belonging to the same cohort. Thus, having been familiarized with the AL\textsubscript{A}, listeners would have been presented with a TP-word like \textipa{\textipa{f5zom5}} (embedded in the AL continuous stream) and it would inhibit the recognition of the base-word “gasoso” \textipa{\textipa{f5zozu5}} in the subsequently presented lexical decision task used as lexicalization test. But this would only hold true for primed base-words, namely for those related to the TP-words, not for unprimed ones, such as the EP word “gaveta” \textipa{\textipa{f5yvetu}} (meaning drawer in Portuguese), to which no TP-word in AL\textsubscript{A} was phonologically related in the familiarization phase. The reverse would be true for listeners exposed to AL\textsubscript{B}.

As illustrated in Table 11, to check for such a specific effect, we used a between-subjects design, with half of the participants presented with what we call the AL\textsubscript{A}, in which TP-words were related to the base-words of list A (see Appendix V)
but not to those of list B, and vice-versa for the other participants presented with we
call the $\text{AL}_B$. This ensures that any ALL effect (in the forced-choice test) across the
two ALs could only be due to statistical learning and allowed estimating the speed of
processing of each base-word with (primed condition) and without (unprimed
condition) the influence of a latent novel competitor (the AL word).

Table 11: Examples of materials used (in phonologic form, according to IPA) in the AL familiarization phase (AL-
Fam.), two-forced choice (ALL-test), and lexicalization test (Lexical Decision – LD on base-words) in Experiment 8, separately
for the two artificial languages (ALs: $\text{AL}_A$ and $\text{AL}_B$). The “#” defines word boundaries according to the statistical segmentation
cues (AL-TPs and phonotactics); the “-” represents concatenation. Shared phonological onsets between TP-words and base-
words are underlined.

<table>
<thead>
<tr>
<th>AL-Fam. phase</th>
<th>ALL-test</th>
<th>Lexicalization Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>(21-min exposition)</td>
<td>(forced choice test)</td>
<td>(LD on base-words)</td>
</tr>
<tr>
<td>ALs</td>
<td>TP-word</td>
<td>Part-word</td>
</tr>
</tbody>
</table>
| $\text{AL}_A$ | Primed | Unprimed | \begin{tabular}{l}
$\ldots$\#ge\-zo\#me\#pi\-ra\-
\end{tabular} & \begin{tabular}{l}
ge\-zo\-me \\
mu\#ge\-zo \\
ge\ozu \\
ge\evete \\
\text{ru\le\te} \\
\text{re\vina} \\
\end{tabular} |
| | Unprimed | Primed | \begin{tabular}{l}
$\ldots$\#ge\-ve\-lu\#re\-vi-
\end{tabular} & \begin{tabular}{l}
ge\-ve\-lu \\
ku\#ge\-ve \\
ge\ozu \\
ge\evete \\
\text{ru\le\te} \\
\text{re\vina} \\
\end{tabular} |

Since a pretest (see 6.4.2. section) had shown that the two lists of base-words led to similar lexical decision latencies and performances in participants that were not presented with any AL beforehand, in the present case any accuracy and/or latency difference between the two lists of base-words observed within each group of participants after familiarization to an AL would necessarily reflect a change in their processing due to the lexicalization of the outputs of the statistical segmentation
process (i.e., of the TP-words) that overlapped with the base-words. This inhibitory effect would be the “lexical footprint” of these novel items (Gaskell & Dumay, 2003a).

Furthermore, listeners also performed the ALL test and the lexicalization test one week later, with no intervening new familiarization period. Previous work has shown that while a low rate of familiarization to novel items presented in isolation (e.g., 12 times each, Gaskell & Dumay, 2003a) leads to a slow lexicalization process, massive exposure to such novel items endorse immediate as well as long-lasting lexicalization effects (Gaskell & Dumay, 2003b). Whether the same holds true for massive exposure to items presented in a continuous stream was examined here, since each TP-word was present 189 times in the AL stream.

6.4. Method

6.4.1. Participants

Thirty-two undergraduate psychology students at the University of Lisbon participated in the experiment for a course credit. None of them participated in the other experiments. Sixteen were familiarized with AL\textsubscript{A} and the other 16 with the AL\textsubscript{B}.

6.4.2. Material

Participants were exposed to different materials in the three phases that constituted the first experimental session of this Experiment. All the AL-material was created based on real EP words (i.e., the Base-Words).
**Base-Words**

Two lists of 10 trisyllabic (CV.CV.CV) real European Portuguese words constituted the base-words selected in order to create the two ALs and to evaluate the possible lexicalization of the TP-words.

These base-words were selected from the Porlex database (Gomes & Castro, 2003). All of them were six phonemes long, with UP occurring at or earlier than the fourth phonemic position. Word frequency was based on Corlex (available at: [http://www.clu.ul.pt](http://www.clu.ul.pt)) and log transformed for normalization of the distribution. The selected base-words were paroxytones and their second and third syllables onset were also possible word onsets in European Portuguese. Thus, neither lexical stress nor phonotactic violations could be used for segmentation.

<table>
<thead>
<tr>
<th></th>
<th>Block A</th>
<th>Block B</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Log Frequency</strong></td>
<td>3.51 (0.38)</td>
<td>3.79 (0.36)</td>
<td>-0.54</td>
</tr>
<tr>
<td><strong>Neighborhood density</strong></td>
<td>1.5 (0.17)</td>
<td>1.7 (0.26)</td>
<td>-0.65</td>
</tr>
<tr>
<td><strong>TP (1st and 2nd syl.)</strong></td>
<td>5.0 (2.15)</td>
<td>5.0 (2.17)</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>UP</strong></td>
<td>3.7 (0.15)</td>
<td>3.7 (0.15)</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>Duration (ms)</strong></td>
<td>885.20 (16.10)</td>
<td>885.40 (13.04)</td>
<td>-0.01</td>
</tr>
</tbody>
</table>

As illustrated in Table 12, the two lists of base-words (A and B; see Appendix V) were matched on word onset (and whenever possible on the first syllable), mean log frequency, neighborhood density (i.e., number of words that differed from them by 1-phoneme substitution, addition, or deletion, while
preserving relative position of the segments; e.g., Vitevitch & Luce; 1998; 1999; based on *Porlex* database, Gomes & Castro, 2003), mean TP between the first and the second syllable in European Portuguese, and mean duration.

**Lexical Decision Pretest**

To ensure that the base-words of the two lists were a priori equivalent, we checked whether they would lead to similar lexical decision performances when no prior familiarization phase to any AL is used. To this aim, we tested an independent group of 17 participants in lexical decision. No difference was found between the two lists on either accuracy \(t(16) < 1, p > .10;\) with mean accuracy of 80 and 82% for lists A and B, respectively or response latency \(t(16) = 1, p > .10;\) average RT of 1166 and 1144 ms for lists A and B, respectively. Thus, any accuracy and/or latency difference between the two lists of base-words observed after familiarization to the AL would necessarily reflect a change in their processing due to the lexicalization of the outputs of the statistical segmentation process (i.e., of the TP-words) that overlapped with them.

**AL Material**

Each one of the two ALs was based on one of the two lists of base-words, with AL\(_A\) based on list A and AL\(_B\) on list B. Each AL was constituted by 10 trisyllabic TP-words. These diverged from the base-words by their last CV syllable, after the fourth phoneme and thus after the UP.

Each group of listeners was familiarized with one of the two ALs, in a between-subjects (AL\(_A\); AL\(_B\)) counterbalanced design. For participants presented
with AL\textsubscript{A}, each TP-word (e.g., /\textipa{fiv\textipa{e}ku}/; /\textipa{g\textipa{e}zom\textipa{a}n}/) was phonologically related to one base-word of list A (primed base-words, e.g., “fivela” /\textipa{fiv\textipa{e}l\textipa{n}/, meaning “buckle”, and “gasoso” /\textipa{g\textipa{e}zoz\textipa{u}/ meaning “gaseous”, respectively), but no TP-word was related to any base-word of the other list (unprimed base-words; e.g., “finito” /\textipa{fin\textipa{t}/, meaning “finite”, and “gaveta” /\textipa{g\textipa{v\textipa{e}t\textipa{u}/ meaning drawer). Conversely, for those presented with AL\textsubscript{B}, each TP-word was phonologically related to one base-word of list B, but none was related to list A. Hence, across the two AL familiarization groups, each list of base-words occurred in each (primed vs. unprimed) condition.

The third syllable of each TP-word was created guarantying that: (i) the resulting stimuli was not a real word in EP; (ii) the phonological sequence was phonotactically legal in EP words; and (iii) it occurred in EP words more frequently in final than in initial position [based on \textit{Porlex}, Gomes & Castro, 2003; \( F (1,18) = 29.5, p < .001 \)]. This last criterion was chosen due to material constraints. For having available two balanced types of segmental information (i.e., from AL and from listeners’ native language), we add this positional syllabic information, since in the AL, TP-words had not only higher TPs than part-words but also occurred three times more often. Thus, in the AL stream, two types of segmental information were available: the statistical information of the AL (i.e., TPs between adjacent syllables and raw frequency of the trisyllabic AL stimuli), and (according to the strong overlap from onset of TP-words and Base-Words and the relative position frequency of the last syllable of each TP-word in EP).

For each AL, 10 trisyllabic part-words were selected, each consisting of the last syllable of one TP-word and the first two syllables of another (that occurred
adjacently in the stream during the familiarization phase). All AL stimuli (TP-words and part-words of the two ALs) are presented in Appendix VI.

**Porlex Data**

In order to guarantee that the two ALs were matched on statistical segmental information in listeners’ native language, we computed the average TP between the syllables that constituted each AL stimulus (TP-words and part-words) in European Portuguese words (based on Porlex database; Gomes & Castro, 2003), separately by stimulus type (TP-words and part-words) and AL (AL\(_A\) and AL\(_B\)). According to the mixed 2x2 ANOVA, neither the main effect of AL \([F(1, 36) = 1.9, p > .10]\) nor the interaction between the two factors were significant \([F < 1]\). Importantly, the main effect of stimulus type was reliable \([F(1, 36) = 11.0, p < .005]\), since TP-words had higher TPs in European Portuguese than part-words.

**Naming & Delayed Naming Pretest**

This was also confirmed by naming and delayed naming pretests on the AL stimuli (TP-words and part-words of the two ALs) ran on an independent group of 19 participants. They were presented, on each trial, with an auditory AL stimulus, which they should first name immediately and next name after an acoustical signal (occurring either at 1600, 1800, 2000, 2200, and 2400 ms after the offset of the stimulus). For the immediate naming condition, only the main effect of AL stimulus type (TP-words vs. part-words) was significant \([F(1, 18) = 45.9, p < .0001]\): TP-words were named faster than part-words, \([ALs: F < 1; AL condition x AL stimulus type: F(1, 18) = 3.6, p = .10]\). In delayed naming, no significant effects were found
[i.e., ALs: $F(1, 18) = 1.8, p > .10$; AL stimulus type: $F(1, 18) = 1.3, p > .10$; AL condition x AL stimulus type: $F < 1$].

**AL stream**

The ALs (AL\textsubscript{A}; AL\textsubscript{B}) were synthesized using the same method as in Experiment 1. Each AL was constituted by 10 TP-words. In the AL stream, TP between adjacent syllables was always higher within (TP of 1) than between (TP of 0.33) TP-words and TP-words occurred three times more often than part-words.

For the **AL-familiarization phase**, for each AL, three 7-minutes blocks (rendering 21-minutes) of the synthesized stream were created by concatenating, on each block, 63 tokens of each of the ten TP-words (1890 syllables, 630 tokens of TP-words per block), with the criterion that two tokens of the same TP-word never occurred adjacent in the stream. Only segmental informations (i.e., TPs between adjacent AL syllables and wordlikeness of AL-stimuli) were available in the stream, with no other acoustic or segmentation cues to word boundaries.

For the two **ALL-tests** (i.e., two-alternative forced choice test), as regards each AL, the three syllables of each AL-stimulus (TP-words and part-words) were synthesized and concatenated.

**Lexical Decision Material**

The lexical decision task was constituted by 120 trials. Twenty were critical trials corresponding to the (10 x 2 lists of) base-words and the remaining 100 trials were trisyllabic fillers. There were 40 word fillers, with varied phonological structures and lexical stress in each possible position. The 60 nonwords used as
fillers were created by changing one phoneme (or syllable) of European Portuguese real words of varying phonological structures. This ensured a low proportion of base-words related with the AL to which participants were familiarized (i.e., 16.67%) and a proportion of 50% of the trials requiring a “word” response. In the lexical decision task, no AL stimulus was presented. A practice list of 10 items with the same characteristics of the experimental list was also generated for participants’ familiarization with the task.

The stimuli used in lexical decision task were recorded in a soundproof booth (using M-Audio Fire Wire 410 and Adobe Audition 1.5 Program on a Carillon Audio Systems, Pentium IV PC) by an EP female native speaker using M-Audio Nova-Class A FET microphone and sampled at a rate of 22.05 kHz and 16-bit conversion. Editing of the digitized versions of the stimuli was made with Adobe Audition 1.5 and Praat 4.3.04 (available at http://fon.hum.uva.nl/praat/).

6.4.3. Procedure

Participants were tested individually or in groups of two in a sound-attenuated room. All auditory stimuli were presented through headphones and the AL-familiarization phase was similar to the one of Experiment 7. For the ALL test and lexical decision task, presentation, timing and data collection were controlled by E-Prime 1.1 (Schneider et al., 2002a; 2002b).

All participants were informed that there were two testing sessions, with the second session one week after the first. No other information about the second session was given. They were told that they were participating in an investigation
about word learning and that the first session was constituted by three phases (see Table 12).

The AL familiarization phase was procedurally identical to the one of Experiment 7. After the AL familiarization phase, participants were presented with the ALL-test (two-alternative forced-choice test), according to the AL to which they were previously exposed. The structure of the two ALL-tests (for AL_A and AL_B) was identical, and formally similar to the one described in Experiment 7. Each trial started with a warning tone, followed by two trisyllabic strings, separated by 500 ms of silence, presented through headphones. One of these strings was a TP-word and the other a part-word. For avoiding needless repetitions of AL-stimuli, each TP-word was only paired with each one of the two part-words with which it shares syllables (20 trials) and with two other part-words with no relation with it (other 20 trials), rendering 40 trials. Accuracy was emphasized, although participants were told to always provide an answer. The test began with the four practice trials used in Experiment 7.

Order of trials was randomized for each participant, and order of presentation of stimuli within trials was counterbalanced within each group and experimental session.

All participants performed the same timed lexical decision task. No information about any possible relation between the first two phases and this one was given. Order of trials’ presentation was pseudo-randomized for each participant in each experimental session, obeying to three criteria: the first ten trials were always filler (warm-up) trials; critical trials (Base-Words) did not occur in consecutive order; and no more than three consecutive trials required the same (yes/no) response.
Participants were asked to perform a lexical decision, as quickly and accurately as possible, on each auditory stimulus presented, using their index-fingers to press one of two buttons of the (PST SRB 200A) button box: “yes” responses were given with the right index-finger. Response latencies were measured from stimulus onset, with a response deadline of 3000 ms and 1500 ms inter-trial interval.

Before the experimental list, participants were familiarized with the task in a practice block, with feedback on correctness and speed provided only for practice trials.

After the lexical decision task, participants were reminded about the second session and were dismissed. Overall, the first session lasted about 50 minutes.

In the second experimental session, participants were informed that this was related to the previous first session with the only difference that they would not hear the AL again, but would perform the other two tasks (ALL-test and the lexicalization test). The second session lasted about 20 minutes.

6.5. Results and Discussion

6.5.1. ALL performance: two alternative forced-choice test

Proportions of TP-word choices (as an index of ALL), computed for each participant, is presented in Figure 9, separately for the two moments of test.

In order to evaluate if there was any difference between the two ALs (AL_A and AL_B) and/or between the two moments of testing (immediate; after 1 week) as regards the performance level reached in the two alternative forced-choice test, we ran a mixed 2x2 ANOVA on raw TP-choices, with AL as between-subjects factor
and moment of testing as within-subject factor. No significant effect or interaction was found [AL: $F(1, 30) = 3.8, p > .05$; moment of testing: $F(1, 30) = 1.9, p > .10$; AL x moment of testing: $F(1, 30) = 3.1, p > .05$]. Thus, the similar magnitude of ALL found with both ALs ensures that the statistical structure, instead of any idiosyncrasy about AL phonological repertoire, was the factor responsible for statistical learning.

Statistical learning did occur, indeed, as shown by the high proportion of TP-choices, that reached, on average, 84.8% (SE = 2.7) immediately after familiarization, and 80.6% (SE = 2.4) one week after, both of these results being well above chance level [$t(31) = 12.8$, and $t = 12.5$, respectively, $p < .0001$ in both cases]. Remarkably, immediately after familiarization, 17 out of 32 participants reached at
least 90% of TP-choices, which was still the case of 12 participants one-week later. This high performance level is probably due to the congruence of the two types of available segmental cues, the TPs of the AL material and the word-likeness of TP-words.

6.5.2. Lexicalization test: inhibitory priming in lexical decision

RTs were measured from stimulus to response onsets. Errors were removed and analyzed separately. For each participant, correct RTs longer or shorter than the mean plus or less 2.5 SD were removed from further analyses (less than 4% of the data excluded).

In order to evaluate if the output of segmentation has acquired some lexical status, we first submitted correct RTs to two mixed 2x2 ANOVAs, separately for each moment of testing, with AL (AL_A; AL_B) as between-subjects factor and base-words list (list A; list B) as within-subject factor, using participants as the only random variable (Raaijmakers, 2003; Raaijmakers et al., 1999). Effect sizes were analyzed according to partial eta square ($\eta^2_p$) values (Cohen, 1988).

Mean RTs for Block A and Block B of the Base-Words are presented separately for each AL and moment of testing in Table 13.

Immediately after the AL-familiarization period, a clear pattern was observed (see Table 13 and Figure 10). Indeed, while no significant main effects were found [AL: $F(1, 30) = .04, p = .83$; Base-Words Block: $F(1, 30) = .14, p = .71$], the interaction between the two factors was significant [$F(1, 30) = 9.0, p < .005$. MS$e = 24$

24 The magnitude of the effect sizes, according to $\eta^2_p$, was considered in line with Cohen c(1988), thus only $\eta^2_p > .14$ corresponds to a large effect, revealing differences that are significantly reliable. Cohen d was not used to estimate the effect sizes because it is inappropriate when repeated measures are considered.
3798; $\eta^2_p = .23$. Planned comparisons examined the effect of Base-Words’ Block on each AL. As illustrated in Figure 10, in both ALs, an inhibitory effect was observed for the list of base-words that overlapped with the the TP-words. Indeed, listeners exposed to AL_A presented longer latencies for the primed base-words A (phonologically related with TP-words of the AL_A) than for the unprimed base-words B (with no relation with TP-words of the AL_A) [$F(1, 30) = 5.7, p < .025$], while on the contrary listeners exposed to AL_B presented longer latencies to primed base-words B than to unprimed base-words A [$F(1, 30) = 3.4, p = .05$]. Coherently, the inhibitory effect, namely the latency difference between primed and unprimed base-words, was significantly greater than zero for both AL [$t(15) = 2.8, p = .01$] and AL_B [$t(15) = 2.0, p = .05$].

| Table 13: Mean RTs (measured from target onset) for Base-Words A and B (BWd A, BWd B) and the Priming Effect (P.E., the difference in lexical decision latencies between primed and unprimed conditions) separately for each AL-condition (AL_A and AL_B) and moment of testing (Immediately and Post-1 week) in Experiment 8. Standard errors of the mean in each condition are presented in parenthesis. |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|
|                                | Immediately     |                | Post 1 Week     |                |
|                                | AL_A            | AL_B           | AL_A            | AL_B           |
| Unprimed                       | Primed          | Unprimed       | Primed          | Unprimed       |
| 1092.7 (38.15)                 | 1040.8 (40.76)  | 51.90          | 1001.5 (36.70)  | 929.8 (44.94)  |
|                               | Primed          | Unprimed       | Primed          | Unprimed       |
| 1058.2 (38.15)                 | 1098.6 (40.76)  | 40.40          | 1030.1 (36.70)  | 1114.9 (44.94) |
|                               | Unprimed        | Primed         | Unprimed        | Primed         |
|                               | Primed          | Unprimed       | Primed          | Unprimed       |
|                               | 1092.7 (38.15)  | 1040.8 (40.76) | 51.90          | 1001.5 (36.70) |
|                               | Primed          | Unprimed       | Primed          | Unprimed       |
|                               | 1058.2 (38.15)  | 1098.6 (40.76) | 40.40          | 1030.1 (36.70) |

Thus, immediately after familiarization, the TP-words, by-products of statistical speech segmentation, seem to be already integrated in the mental lexicon, since they are involved in lexical competition, as revealed by the cost they promoted
on the recognition of already existing lexical items that were members of the same cohort.

The errors analysis did not reveal any significant effect [AL: \( F(1, 30) = 0.19, p > .10; MSe = 0.02045 \); base-words list: \( F(1, 30) = 1.08, p > .10, MSe = 0.01174 \); interaction between these factors: \( F < 1 \)], showing that there was no speed/accuracy tradeoff in the responses to the two lists of base-words. For AL\(_A\), average error rates reached 9% (SE = 2.0) and 11% (SE = 4.0) in the base-words lists A and B, respectively; and for AL\(_B\) it reached 11% (SE = 3.9) and 10% (SE = 2.0) for lists B and A, respectively.

Interestingly, the mixed 2x2 ANOVA ran over the lexical decision latencies of the post-one week test, with AL (AL\(_A\) vs. AL\(_B\)) as between-subjects factor and base-words list (list A vs. list B) as within-subject factor, revealed the same significant interaction between AL and base-words list as the one observed one week
before \(F(1, 30) = 42.7, p < .0001, MSe = 2294; \eta^2_p = .59\]. As a matter of fact, listeners exposed to AL\(_A\) responded more slowly to the primed base-words A than to the unprimed base-words B [for AL\(_A\): \(F(1, 30) = 17.9, p < .001\)], while on the contrary listeners exposed to AL\(_B\) responded more slowly to primed base-words B than to unprimed base-words A \([F(1, 30) = 25.11, p < .001]\). No other significant effect was found \([AL: F(1, 30) = 3.5, p = .07; \eta^2_p = .10; base-words list: F < 1; \eta^2_p = .01]\).

Thus, the cost promoted on base-words recognition by lexical competition from novel “lexical” items corresponding to the output of segmentation of the AL is long-lasting. Post-one week, with no further exposition to the AL, the output of statistical segmentation still presents a lexical footprint, penalizing the processing of already existing items that strongly overlap with them \([for AL_A: F(1, 30) = 17.9, p < .001; for AL_B F(1, 30) = 25.11, p < .001]\).

Once again, the error analysis did not reveal any significant effect \([all Fs < 1.5, ps > .20]\), with average error rates of 9% (SE = 2.5) and 10% (SE = 3.1) to Base-Words’ Blocks A and B, respectively, for AL\(_A\); and of 8% (SE = 3.1) and 7% (SE = 2.5) for Blocks B and A, respectively, for AL\(_B\).

Since previous results suggested that the inhibitory priming effect might reinforce over time (e.g., Gaskell & Dumay, 2003), we evaluated if the inhibitory effect observed in the recognition of primed base-words (in comparison to unprimed ones) was strengthened between the two testing moments \(see Figure 11\). The mixed 2x2 ANOVA ran on this priming effect, with AL as between-subjects factor and moment of testing as within-subject factor, revealed a significant main effect of moment of testing \([F(1, 30) = 7.9, p < .01, MSe = 2081.1; \eta^2_p = .21]\): the average
priming effect was 46.15 ms right after familiarization, while it reached 78.3 ms one week later. No other effect or interaction was found [all $F_s \leq 1; \eta_p^2 < .03$].

The present pattern of results thus generalizes to adult listeners Saffran’s (2001) and Estes et al.’s (2007) findings obtained with infants. The output of statistical learning is rapidly integrated into the mental lexicon, even in the context of a mature perception system. In addition, the lexical engagement of segmentation outputs is not only long-lasting, but is even consolidated between the first (immediate) and second (post-one week) moments of testing, which is in line with Dumay and Gaskell’s (2007) proposal.

However, since in the present experiment, the two types of information available in the AL stream (i.e., TPs and wordlikeness of the AL stimuli) suggested the same segmentation hypotheses, we still do not know if the observed lexicalization effect was promoted by the availability of the two congruent speech
segmentation cues or uniquely by the wordlikeness of TP-words. That knowledge of phonotactic probabilities was used by the participants of the present experiment in ALL is suggested from their high performance level in the forced-choice test, which is in accordance both with former ALL studies manipulating this factor (Onnis et al., 2005; Perruchet et al., 2004) and with studies that have showed a redundancy gain with congruent segmentation cues (e.g., Fernandes et al., 2007; Vroomen et al., 1998). However, none of these studies evaluated lexical engagement. To what extent wordlikeness per se and/or cue integration may have favored the lexicalization of the new item configurations was evaluated in Experiment 9.
Experiment 9. The status of the output of the segmentation procedure adopted by listeners: When *incongruent cues* are available in the AL-stream.

In the present experiment, the two types of segmental information studied in Experiment 8 (namely, TPs and wordlikeness of the AL stimuli) were also available in the AL stream, but were incongruent since they suggested incompatible segmentation outputs. As illustrated in Table 14, whereas the TPs between syllables suggested that TP-words were the “lexical items” of the new language, this time it was the part-words that were the AL stimuli that overlapped from onset with real words of the listeners’ native language.

Table 14: Examples of materials used in AL-familiarization phase (AL-Fam.), ALL-test, and Lexicalization test (Lexical Decision – LD on Base-Words) on Experiment 9, regarding the two Artificial Languages (ALs). The “#” defines word boundaries according to the AL-TPs; the “-” represents concatenation. Shared phonological onsets between Part-words and Base-Words are underlined. All material is presented in phonological form, according to IPA.

<table>
<thead>
<tr>
<th>AL-Fam. (21-mins exposition)</th>
<th>ALL-test (forced choice test)</th>
<th>Lexicalization Test (LD on Base-Words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL₂ₐ</td>
<td>TP-word</td>
<td>Part-word</td>
</tr>
<tr>
<td>AL₂ₐ</td>
<td><img src="image1.png" alt="Example" /></td>
<td><img src="image2.png" alt="Example" /></td>
</tr>
</tbody>
</table>

On the basis of previous results obtained with incongruent cues in ALL settings, we hypothesized that listeners’ native language information would be
considered as more reliable than AL statistical information, at least with an intact speech signal, as was the case here. Indeed, using an AL in which both TPs and another acoustic-phonetic cue, namely coarticulation, were available, Fernandes et al. (2007) showed that, with intact speech signal, coarticulation was able to drive the segmentation process, even when TPs suggested other, incongruent segmentation hypotheses. The same pattern had been previously observed, with intact signal, in 8-month-old infants (Johnson & Jusczyk, 2001).

In the present experiment with incongruent cues, if listeners preferred native language information over AL statistical information, the output of their segmentation process would correspond to the stimuli that, although presenting lower TPs in the AL, strongly overlap from onset with real words of listeners’ native language. In other words, in this case, listeners would prefer the part-words as the probable “words” of the new language, discarding the TP-words, a preference that would be reflected by a significant below-chance performance in the two-forced choice test. The lexicalization test (also used in Experiment 8) allowed evaluating whether the output of the segmentation procedure adopted by listeners (according to their performance on the ALL test) would still be integrated in their lexical environment.

6.6. Method

6.6.1. Participants

Twenty-four undergraduate psychology students (12 familiarized with new AL\textsubscript{A2} and the other 12 with the new AL\textsubscript{B2}) at the University of Lisbon participated in
the experiment for a course credit. None of them participated in the previous experiments.

6.6.2. Material

The base-words and AL stimuli were those used in Experiment 8, with the only difference that the TP-words of the present experiment corresponded phonologically to the part-words of Experiment 8, and vice-versa (see Table 14 for examples of the material used).

Each one of the two ALs (AL_{A2} and AL_{B2}) was constituted by ten TP-words (based on the AL-statistical information were the “lexical-units” of the new language). The ten part-words were trisyllabic stimuli constituted by the last two syllables of one TP-word and the first syllable of another, which strongly overlapped from onset until the fourth phoneme with real EP words (i.e., the Base-Words). Thus, part-words had a higher wordlikeness than TP-words (see Appendix VI).

**AL stream**

The two new ALs (AL_{A2} and AL_{B2}) were synthesized using the same method of the previous experiment. Each AL was constituted by 10 TP-words (the Part-words of the previous experiment). The TP between adjacent AL syllables was always higher within (TP of 1.00) than between (TP of 0.33) TP-words, as well as their raw frequency (TP-words occurred three times more often than part-words: each TP-word occurred 189 times, while each part-word occurred 63 times).
For the *AL-familiarization phase*, for each *new* AL, three 7-minutes blocks of the synthesized stream were created as in Experiment 8. The only segmentation cues available were the *incongruent* segmental informations, with no other acoustic or segmentation cues to word boundaries.

The *ALL-tests* were constructed as in Experiment 8, and the *lexicalization test* (i.e., lexical decision task) was the one used in the previous experiment.

### 6.6.2. Procedure

Procedure and structure of the two experimental sessions (immediately and post-one week) were identical to the previous experiment.

### 6.7. Results and Discussion

6.7.1. *ALL Performance: two alternative forced-choice test*

Average proportions of TP-word choices (as an index of ALL), computed for each participant, is presented in Figure 12, separately for the two moments of test.

As in Experiment 8, we first evaluated if there was any difference between the two ALs (AL$_{A2}$ vs. AL$_{B2}$) and/or between the two moments of testing (immediate; after 1 week) as regards the performance level reached in the two alternative forced-choice test. This was not the case, since the mixed 2x2 ANOVA ran on raw TP-word choices, (with AL as between-subject factor and moment of testing as within subject-factor) showed no significant effect or interaction [all $Fs \leq 1$].
Figure 12: Mean percentage of TP-word choices in the ALL-test for each participant, according to the moment of testing: immediately after the AL-familiarization phase (i.e., immediate) and one week after with no intervening AL-familiarization period, in Experiment 9. Chance level corresponds to 50%.

Performance was below the chance level (i.e., below 50%), reaching 41.1% (SE = 3.6) immediately after familiarization and 43.5% (SE = 3.6) one week later [t(23) = -2.5, p < .01; and = -1.8, p < .05, respectively], with 13 out of the 24 participants presenting a performance clearly below chance level (≤ 45% of TP-words choices) in the two tests. Thus, listeners discarded the TP-words and considered more often the part-words as the “lexical units” of the new language. This is in line with former results suggesting that at least with intact speech (cf. Fernandes et al., 2007), listeners at different stages of their linguistic development consider native speech segmentation cues as more reliable than the statistical segmental information conveyed by an AL stream.
6.7.2. Lexicalization test: inhibitory priming in lexical decision

As in Experiment 8, RTs were measured from stimulus onset to response onset, errors were analyzed separately and the same trimming procedure in RTs was applied (exclusion of less than 3.5% of the data).

Mean RTs for the two lists of base-words (lists A and B) are presented separately for each AL and moment of testing in Table 15.

<table>
<thead>
<tr>
<th>Moment of Testing</th>
<th>Immediately</th>
<th>Post 1 Week</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Primed</td>
<td>Unprimed</td>
</tr>
<tr>
<td>943.4 (26.15)</td>
<td>916.5 (23.92)</td>
<td></td>
</tr>
<tr>
<td>895.9 (20.56)</td>
<td>852.8 (21.63)</td>
<td></td>
</tr>
</tbody>
</table>

Immediately after familiarization, the output of the segmentation procedure adopted by listeners was not lexicalized. Indeed, the mixed 2x2 ANOVA with AL as between-subject factor (AL\textsubscript{A2}; AL\textsubscript{B2}) and base-words list as within-subject factor (list A; list B) showed that neither the interaction between the two factors at study \([F= 1; \, \eta^2_p = .04]\) nor the main effect of base-words list \([F < 1; \, \eta^2_p = .01]\) were significant. We thus found no difference between the processing of primed and unprimed base-
words, namely between those that overlapped with the segmentation outputs (here, the part-words) and those that did not.

Surprisingly, the main effect of AL was significant \[F(1, 22) = 7.7, p = .01, \text{MSe} = 11687; \eta_p^2 = .26\], with listeners exposed to \(\text{AL}_{A2}\) responding faster than listeners exposed to \(\text{AL}_{B2}\). This result was unexpected, since the two ALs were well matched in all relevant parameters (see section 6.4.2. of the present chapter). Furthermore, in the previous experiment the same phonological repertoire was used in ALs and no global differences were found there. It seems to be due to the manipulation of the AL factor as a between-subjects variable.

Much more important is the fact that there was no hint of an effect showing that the output of phonotactic segmentation had acquired some lexical status. Coherently with the non-significant effect of base-words list and its absence of interaction with AL, for both ALs the inhibitory effect (measured as the lexical decision latency difference between primed and unprimed base-words) did not differ from zero \([t_{11} < 1, ps > .10]\).

In the errors analysis, no significant effect was found \(Fs < 2.5, ps > .10\), showing that there was no speed/accuracy trade-off. For listeners familiarized with \(\text{AL}_{A2}\), errors reached on average 11\% (SE = 3.0) for base-words A and 14\% (SE = 4.0) for base-words B; for listeners exposed to \(\text{AL}_{B2}\), they reached 11\% (SE = 3.0) and 12\% (SE = 4.0), respectively.

The same pattern found one week later. Listeners familiarized with \(\text{AL}_{A2}\) were faster than the ones familiarized with \(\text{AL}_{B2}\) \([F(1, 22) = 5.9, p = .024, \text{MSe} = 2598, \eta_p^2 = .21]\), but neither the main effect of base-word list \([F(1, 22) = 1.5, p > .10, \text{MSe} = 2598; \eta_p^2 = .06]\) nor the interaction between the two factors \([F(1, 22) = 2.9, p\)}
Chapter 6. On-line measures and Lexicalization during ALL

=.10, $MSe = 2598; \eta^2_p = .11$] were significant.

The analysis on errors did not reveal any significant effect [AL condition: $F(1, 22) = 1.32, p > .10$; base-words list: $F(1, 22) = 3.3, p = .10$] or interaction between the factors [$F < 1$]. For listeners familiarized with $AL_A$, errors reached on average at 9% (SE = 3.0) for base-words A and 12% (SE = 4.0) for base-words B; and for listeners exposed to $AL_B$, they reached 7% (SE = 3.0) and 9% (SE = 4.0), respectively.

Thus, although one may expected the incubation period, namely the delay in lexical engagement of speech segmentation outputs (cf. Dumay & Gaskell; 2003a; 2003b; 2007) to be particularly crucial in conditions of incongruent segmentation cues, one week after familiarization speech segmentation outputs were still not integrated into the mental lexicon as potential word-like units.

**Sub-analysis**

In order to guarantee that these null effects also hold true for participants whose segmentation outputs consistently corresponded to the part-words, we ran two further ANOVAs (according to moment of testing) with the same factors as in the previous analyses but considering only the participants that had displayed a clearly below-chance performance ($\leq 45\%$) in the forced-choice test. Remarkably, both immediately after familiarization and post-one week, neither the base-word list main effect [immediate: $F < 1$; post-one week: $F(1, 11) = 2.9, p > .10$] nor the interaction of that factor with AL [immediate: $F < 1$; post-one week: $F(1, 11) = 1.8, p > .10$] were significant. The main effect of AL was also nonsignificant [$F \leq 1$ for both immediate and post-one week tests]. Thus, even for listeners that systematically
adopted a segmentation procedure based on the wordlikeness of part-words, no hint of a lexicalization effect of their segmentation outputs was found.

The current results suggest that when different types of segmental information indicate incongruent word boundaries in the AL, listeners are still able to segment the stream, with knowledge of their native language being preponderant over AL statistical information, which is in line with previous findings (e.g., subsegmental information: Fernandes et al., 2007; Johnson & Jusczyk, 2001; prosodic contours: Shukla et al., 2007; Experiment 3 and 6) and reveals the differential weighting attributed to different types of sublexical cues in speech segmentation (Mattys et al., 2005).

However, the AL units segmented on the basis of only wordlikeness and thus here against the statistical information conveyed by the AL stream- did not seem to have acquired any sort of lexical status, either immediately or one week after familiarization to the AL. Indeed, the speed of processing for existing lexical items that strongly overlapped with part-words was no different from the one found for other, unrelated, lexical items.

### 6.8. General Discussion

The ALL paradigm has become a important tool in the study of language acquisition (for a review see Gómez, 2007) as well as in the context of a full mature speech perception system (e.g., Fernandes et al., 2007; Onnis et al., 2005; Saffran et al., 1996b; Vroomen et al., 1998). Indeed, it allows systematic manipulation of different segmentation cues, reducing the set of uncontrolled variables, and hence provides a powerful way to test various theoretical proposals (e.g., Fernandes et al.,
2007; Shukla et al., 2007). However, two aspects at the core of ALL have been left largely unspecified. To shed light on them, in the present study, we adopted ALL settings combined with conventional experimental techniques.

First, in Experiment 7, we combined ALL with a variant of the word-spotting task, providing a direct demonstration of on-line statistical, TP-based, speech segmentation. Second, in Experiments 8 and 9, adopting the approach of Gaskell and Dumay (2003a; 2003b) within an ALL setting, we demonstrated that, at least in some conditions, the byproducts of statistical segmentation can be rapidly integrated into the listeners' mental lexicon, exhibiting a lexical competition signature.

In Experiment 7, we showed that after familiarization to an AL, listeners are able to use statistical information on-line in segmenting speech. In a subsequently presented AL version of the word-spotting task, they were both more accurate to spot and faster to recognize TP-words than part-words embedded in four-syllable nonsense strings. This result cannot be attributed to a strategic detection of AL syllables, since both types of AL stimuli were presented in the stream during the familiarization phase and were constituted by the same syllabic repertoire. Remarkably, even participants' disyllabic fragment responses were consistent with the notion that they did use TPs: for part-words bearing sequences, participant’s disyllabic responses corresponded more frequently to syllables belonging to the same TP-word than to syllables belonging to different TP-words, because of the higher TPs of the former fragments. In addition, disyllabic responses belonging to the same TP-word were as frequent for part-words as for TP-words bearing sequences, probably because these fragments had similar TPs and hence could not be distinguished on statistical grounds.
It is important to note, however, that the present spotting latencies were longer than the ones observed in previous studies (e.g., Cutler & Norris, 1988; McQueen et al., 1994; Norris et al., 1995), which at first sight could raise some doubts regarding the on-line nature of the collected data. Yet, three aspects that distinguish Experiment 7 from conventional word-spotting experiments probably underlie this effect.

First, for reducing the temporal lag between detection and vocal responses, we adopted a procedure in which participants did not give any speeded motor response (as usually done in previous word-spotting studies): latencies were measured from the onset of the string to the onset of vocal responses, which might have increased RTs. Second, the utterances used in both the AL familiarization and word-spotting phases were synthesized, which obviously constitutes an extremely impoverished material in comparison with naturally produced utterances. Although this may have had an adverse impact on word-spotting performances, the use of such a synthesized material was deliberate, since it allowed guaranteeing that only statistical information would be available for AL word-spotting. Finally, and probably most importantly, participants were asked to segment a non-lexical stimulus (the AL stimulus) embedded in a nonsense string, with only AL statistical information being available to assist them. Therefore, listeners could not use their native language knowledge (either sublexical information such as lexical stress or high-level information such as lexical knowledge) to correctly parsed the AL-stimuli from the nonsense carrier sequences. Even in previous word-spotting studies, in which participants spot real words with full lexical configurations (i.e., with associated phonological, orthographic, semantic and syntactic information), error
rates can be very high (reaching levels between 24% and 45%, even when the carrier sequence is not the beginning of any real word; McQueen et al., 1994; Norris et al., 1997; see also Cutler & Norris, 1988) and responses may be rather slow. Thus, it is not surprising that this task becomes quite hard when no high-level information could assist segmentation since the only available segmentation cue was the TPs conveyed by the AL stream presented to the listeners during a 21-minutes period.

In any case, the present results confirm the validity of ALL effects based on forced-choice testing (e.g., Saffran et al., 1996b) and on RT tasks with isolated stimuli (McLennan & Luce, 2004; McLennan et al., 2004; Onishi et al., 2002; Warker & Dell, 2006). In addition, our results draw an additional parallel (Perruchet & Pacton, 2006) between ALL based on TPs and other sequence learning effects studied within the framework of implicit learning like those observed in the serial reaction time task (e.g., Cleeremans & McClelland, 1991), in which on-line sensitivity to statistical information like TPs has already been reported (Hunt & Aslin, 2001; Remillard & Clarck, 2001).

The results of Experiment 8 showed in addition that statistical segmentation outputs can also be rapidly available for integration in adult listeners’ mental lexicon. TP-words that also strongly overlapped from onset with real words in the listeners’ native language were not only correctly parsed from the stream, as revealed by very good performance in the forced-choice test (used as an ALL index), but also gained a lexical status, exhibiting a lexical competition signature. Indeed, both immediately after the AL familiarization period and one week later, the speed of processing of real words members of the same cohort than the TP-words parsed from the AL stream was penalized in comparison to the processing of unrelated real words.
These results add to previous findings of word learning in the absence of semantic knowledge (e.g., Bowers et al., 2005; Gaskell & Dumay, 2003a; 2003): they demonstrate that novel lexical items can be acquired in the context of a continuous speech stream, which probably closely resembles natural conditions of word learning. This lexicalization effect found with adult listeners is also in line with infants data (Saffran, 2001; Swingley, 2007).

In Experiment 9 we evaluated whether wordlikeness of AL stimuli could by itself drive segmentation, against the statistical information available in the AL stream, and whether the byproducts of such a segmentation process could also be lexicalized. Wordlikeness of AL stimuli was able to drive the segmentation process, since listeners considered the part-words (which had lower TPs in the AL but strongly overlapped with real words) as the “units” of the novel language, discarding the TP-words. This is in line with previous findings. Listeners in different stages of linguistic development consider their language specific information (e.g., lexical stress; coarticulation) as a more reliable cue to speech segmentation than general domain TPs (e.g., Fernandes et al., 2007; Jonhson & Jusczyk, 2001). Furthermore, phonotactics has an important and unique role in wordlikeness classification of polysyllabic nonwords (Frisch et al., 2000), and listeners are able to use this segmental information of their language since an early phase in linguistic development (e.g., Mattys & Jusczyk, 2001) until adulthood (e.g., McQueen et al., 1998; Vitevitch & Luce, 1998; 1999; 2005). With intact speech, this segmental information is also able to drive speech segmentation, even when other sublexical cue, such as lexical stress, is available in the stream and suggests incongruent segmentation points (Mattys et al., 2005; Experiment 2). In ALL settings, Onnis et
al. (2005; see also Perruchet et al., 2004) already demonstrated that, when the statistical information available corresponds to TPs between nonadjacent syllables, phonotactics can also underlie listeners ALL.

However, in contrast to what was found in Experiment 8, the byproducts of the segmentation procedure adopted by the listeners with the incongruent cues used in Experiment 9 were not lexicalized. The difference found in Experiment 9 between the results of the forced-choice test and of the lexicalization test is in accordance with Leach and Samuel’s (in press) distinction between items configuration (in the present study, the phonological representations parsed from the AL stream) and their lexical engagement, namely the items’ involvement in lexical dynamics. Indeed, whereas the ALL effect observed in the forced-choice test is the outcome of listeners’ ability to segment the speech stream into units, which corresponds to their knowledge of AL “words” configuration; this information is not able to predict the potential lexical status of the AL units.

An alternative explanation of the effects found in lexical decision may however be that they reflect a transient activation of the base-words during the AL familiarization period, due to the fact that the beginning of either TP-words (in Experiment 8) or part-words (in Experiment 9) formed possible words up until the nonword point (when they deviated from the base-words). However, this is extremely unlikely. In phoneme monitoring, Pitt and Samuel (1995) did not find any hint of lexical effects for words and nonwords derived from them in early-unique conditions (as it was the case in present Experiments 8 and 9). While sublexical information (as in the experiments presented in this chapter) can rapidly modulate spoken word processing (e.g., Pitt & McQueen, 1998; Pitt & Samuel, 1995),
lexically-driven, top-down effects take time to build, probably not emerging until well into the word. Therefore, since the AL material was not constituted by real words, and the critical AL stimuli (i.e., the TP-words in Experiment 8, and the part-words in Experiment 9) diverged early on from real words, top-down effects were unlikely to have occurred. Even if this could have been possible, this activation would rapidly vanish, since in Experiments 8 and 9, the outputs of speech segmentation differed from real words (i.e., from the base-words) on the last syllable. Such an effect would result in transient facilitation of base-words processing, thus in a facilitatory priming effect which would be observed immediately after familiarization, in both experiments, but not after one week. On the opposite, whereas in Experiment 9 no hint of lexicalization effects was found, in Experiment 8 the effect was inhibitory, long-lasting and even strengthened between the two moments of testing.

This strengthening is consistent with general evidences of sleep being implicated in learning and memory consolidation (for a review, Walker & Stickgold, 2004) and, in particular, in word learning (Dumay & Gaskell, 2007; Fenn, Nusbaum, & Margoliash, 2003) and vocabulary growth (Gais, Lucas, & Born, 2006). This increase of the priming effect over time is also coherent with the idea that long-term priming effects reflect word learning rather than the temporary activation of words tapped by short-term priming (Bowers, 1999, 2000; Sumner & Samuel, 2007). Under this view, long-term priming involves structural changes, and these structural changes affect the later processing of repeated items.

One may however wonder whether the inhibitory effect observed in Experiment 8 may reflect strategic processing instead. Indeed, the origin of the
inhibitory priming effect linked to onset overlap is highly debated, at least in immediate priming studies. For example, Goldinger (1999) argued that the inhibitory priming effect does not provide an accurate picture of lexical competition, because it co-occurs with evidence for response biases. More recently, Pitt and Shoaf (2002) argued that participants’ surprise when they encounter the first related trials might explain the inhibitory priming effect, which they showed to dramatically decrease between the beginning and the end of the experiment. Other authors have argued that long-term priming effects may also be contaminated by strategic processing. For example, Luce et al. (2000) consider as strategically-based the long-term priming effects observed by Monsell and Hirsh (1998) with lags of 1 to 5 min between primes and targets overlapping at their onset, among others because they used a high proportion of overlapping items in relatively small stimulus sets (Goldinger, 1998a; Goldinger et al., 1992), leading participants to notice the occurrence of overlapping items, as acknowledged by the authors (Monsell & Hirsh, 1998, p.1511).

However, numerous studies support the notion that the inhibitory priming effect linked to onset overlap is largely lexical in nature and does not reflect purely strategic processes. Indeed, contrary to facilitatory priming, inhibition due to multiple-phoneme onset overlap is, in fact, stronger in conditions in which strategic confound is minimized, for example when the proportion of related prime-target pairs is low (Hamburger & Slowiaczek, 1996). Even surprise cannot fully explain the inhibitory priming effect, since such an effect is observed even when participants were presented with related pairs in the training session (Dufour, Frauenfelder, & Peereman, 2007). In addition, it is worth noting that the study of Pitt and Shoaf (2002) showed that the size of the inhibitory priming effect decreases as strategic
processes build-up over the course of the experiment (see also Hamburger & Slowiaczek, 1996). Thus, at least in immediate priming experiments, response biases are more likely to mask than to cause the inhibitory priming effect (Dufour et al., 2007). Second, the size of the inhibitory priming effect is influenced by lexical factors such as the lexicality of the primes (Slowiaczek & Hamburger, 1992), the relative frequency of the primes and targets (Radeau, Morais, & Segui, 1995), and the neighborhood density of the target words (e.g., Dufour & Peereman, 2003; Dufour et al., 2007; Luce & Large, 2001; Vitevitch & Luce, 1998; 1999; 2005). Third, the inhibitory effect observed in long-term priming studies is limited to word primes, supporting a lexical origin of the effect (Monsell & Hirsh, 1998; Sumner & Samuel, 2007).

In Experiments 8 and 9, we tried to avoid response biases by never presenting the base-words to listeners before the lexical decision task, never giving them any information about their possible relation with the AL stimuli, and using a low proportion of trials (less than 17%) in which real words had a phonological relation with AL stimuli, as well as a long lag between presentation of AL isolated stimuli (in the forced-choice task) and presentation of these base-words (for lexical decision). However, one potential problem lies in the fact that we evaluated lexicalization only through lexical decision, a task that, although widely used, involves a decisional component and thus is also often considered as involving substantial post-lexical processing (e.g., Abrams & Balota, 1991; Goldinger, 1996a; Slowiaczek & Hamburger, 1992) and hence to be highly sensitive to strategic biases in priming studies (e.g., Holender, 1992; Norris, McQueen, & Cutler, 2002; Radeau, Morais, & Dewier, 1989; Slowiaczek, Soltano, Wieting, & Bishop, 2003). There is however
some evidence that lexicalization effects are not dependent on the specific lexicalization test chosen. As a matter of fact, Gaskell and Dumay (2003a) reported consistent lexicalization effects in both lexical decision and pause detection tasks. In addition, the latter task seems to be less prone to strategic biases because it offers a measure of the overall level of lexical activity in the absence of any explicit linguistic judgment, since listeners are merely asked to detect periods of silence in speech (Mattys & Clark, 2002). While future work should aim at checking whether the same holds true as regards lexicalization in ALL, it is worth noting that task idiosyncrasy cannot explain why an inhibitory priming effect was observed only in the context of the congruent cues used in Experiment 8, and not with the incongruent cues used in Experiment 9.

Two factors can actually explain the inconsistent pattern of results of Experiments 8 and 9 as regards the lexicalization effect: the frequency of occurrence of the AL stimuli, and the (in)congruence of the available segmentation cues.

Possibly, novel phonological sequences need to be sufficiently familiar before they are lexicalized. Indeed, it would be counter-productive if all phonological forms to which listeners are exposed would be stored as different categories in mental lexicon. In addition, it has been shown that novel item’s frequency has a role in lexicalization (Gaskell & Dumay, 2003b). In Experiments 8 and 9, due to material constraints, part-words occurred less often than TP-words during familiarization, as in most ALL studies, e.g., Saffran et al., 1996a). For sure, raw frequency of the AL stimuli is not able to explain all ALL patterns, since ALL has been reported even when the raw frequency of part-words and TP-words had been equated. (e.g., Aslin, Saffran & Newport, 1998; Saffran et al., 1996a). But this factor is well known to
affect language processing (e.g., Brent & Cartwright, 1996; Magnuson et al., 2003) and may account for at least some ALL patterns (e.g., Perruchet & Vinter, 1998). Thus, the absence of a significant lexicalization effect in Experiment 9 might be related to the lower frequency of part-words in comparison to TP-words. Nevertheless, previous data suggest that the raw frequency of part-words was sufficiently high in Experiment 9 to expect lexicalization to occur. Indeed, although using a different paradigm and presenting participants with already segmented stimuli, Gaskell and Dumay (2003a) reported a lexicalization effect in lexical decision with about 30 phoneme monitoring trials, and Gaskell and Dumay (2003b) even found an inhibitory effect post-one week for “low-frequency” novel items that had occurred only 12 times each during phoneme monitoring. In Experiment 9, part-words occurred in the AL stream 63 times each (i.e., twice more than in Gaskell and Dumay, 2003b, and five times more than Gaskell & Dumay’ 2003b low-frequency novel items), which was enough for listeners to consider them as the “words” of the AL (in the forced-choice test) but not to lexicalize them. It is of course possible that exposure to a continuous stream acts as some kind of noise to the lexicalization process, hence impeding massive exposure to AL stimuli to endorse an immediate lexicalization effect. However, the fact that we observed no lexicalization gain one week after familiarization suggests that frequency of occurrence is not the only factor at play.

The pattern of results found in Experiments 8 and 9 might alternatively reflect the fact that the consistency of the segmentation hypotheses afforded by different sources of information is important not only for speech segmentation (e.g., Fernandes et al., 2007) but also for word learning, modulating the potential
lexicalization of the segmentation outputs. Thus, only phonological representations parsed from the continuous speech stream on the basis of the strongest evidence, i.e., with different segmentation cues suggesting the same parsing, would be lexicalized. Such a view would be in line with the suggestion that the integration of congruent segmentation cues provides evidence about aspects of linguistic structure that are not available from any single source of information, reducing the potential for making false generalizations (Christiansen, Allen, & Seidenberg, 1998; Christiansen & Curtin, 2005).

More generally, further work should also aim at examining how close the present results may be related to the concept of ecological validity. In natural settings, listeners are exposed to a rich combination of different signal-derived and lexically driven segmentation cues. But most of the time these cues converge rather than diverge. Thus, in less ecological situations, e.g. when cues collide, as was the case in Experiment 9, listeners might be reluctant to accept the by-product of segmentation as “true” words. Whatever the outcome of this future research, our results clearly show that the lexicalization effect found in Experiment 8 is not only a matter of lexical similarity.

Future work should also be aimed at examining which computational models may most adequately account for such word learning effects occurring in the context of speech segmentation and, in particular, for such rapid acquisition and integration of statistical information. While any discussion of this point remains highly speculative in the absence of precise simulations, it seems to us that the Adaptive Resonance Theory or ART (Carpenter & Grossberg, 2003; Grossberg, Boardman, & Cohen, 1997; Grossberg & Myers, 2000; see also Sumner & Samuel, 2007, and
Vitevitch & Luce, 1999) might be a good candidate to apprehend word recognition and word learning within a continuous flow, as well as to account for the present findings.

In ART, the resonance created by the match between bottom-up information and top-down expectations (stored in Long Term Memory – *LTM*) is responsible for the creation of the conscious spoken percept (Grossberg et al., 1997; Grossberg, 2003) and also drives word learning (Carpenter & Grossberg, 2003). While the spoken input unfolds, it activates speech segments that in turn activate “items” represented at Working Memory (WM), where they are also unitized, into “list chunks” of variable length (e.g., phonemes, syllables or words) within the same masking (inhibitory) field.

When novel words are presented in the speech stream, as happens during familiarization to an AL, novel information cannot form a good enough match with the expectations that are read-out by previously learned recognition categories. This will trigger a memory search, or hypothesis testing, that leads to selection and learning of a new recognition category that can better match the input. Since in the masking field, longer list chunks will promote stronger inhibition, a chunk that codes a longer novel list like /suset3/ (one of our TP-words in Experiment 8) will have a-priori advantage over chunks that code shorter lists like the familiar nonsense syllables that constitute it. It is precisely the masking advantage of longer list chunks that enables novel words to successfully compete with “amplifier” familiar chunks of shorter lists. When top-down expectation achieves a good enough match with bottom-up data, this match process focus attention upon those feature clusters in the bottom-up input that are expected. If the expectation is close enough to the input
pattern then a state of resonance develops as the attentional focus takes hold, and the novel word is learned. The resonant process will be interrupted or terminated by mismatch reset and hence a new speech segmentation point will be defined. The repeated exposure to specific spatial patterns (in the present case, the repeated occurrence of TP-words in the speech stream) permits learning by the LTM traces in the adaptive pathways between item nodes and the list nodes, reflecting the ALL and lexicalization effects found in Experiment 8.

ART is also able to explain the pattern found in Experiment 9. When different informations are available in the stream, suggesting different competing chunks of the same length (e.g., TP-words vs. more real word-like part-words), resonance is actively reset by input mismatching even before reaching a stable state. The simultaneously activated competing list chunks promote only partial matching resonant loops and hence the new information cannot be stably stored in LTM. Consider for example a masking field that is tuned to expect two chunks like /kufune/ (a TP-words of the AL that based on bottom-up, on-line statistical information is an unifying category) and /funetu/ (a part-word of the same AL, but that according to its high wordlikeness and hence phonotactic information could easily constitute an unifying category). These chunks will strongly inhibit each other based on the shared items /fune/, and although both promote resonant loops, an equilibrated resonant state will not be achieved, and hence novel words will not be stored in LTM.
PART III
GENERAL DISCUSSION & CONCLUSIONS
Chapter 7. General Discussion

The work presented in this thesis was aimed at evaluating the role of different sublexical sources of information in speech segmentation, in the context of a full mature speech perception system. To this general aim, we have considered sublexical segmentation cues of the three types mentioned in the General Introduction of this thesis (see section 1.1.2. of chapter one). In particular, we have considered: coarticulation, a subsegmental cue; TPs between adjacent syllables, a statistical segmental cue; and suprasegmental information such as word primary stress (a language-specific cue) and IP right-edges (a universal prosodic cue; e.g., Grice, 2004; Shukla et al., 2007).

In order to evaluate the role of these sublexical cues in speech segmentation per se, we adopted the ALL paradigm implemented by Saffran and colleagues (Saffran et al., 1996a; 1996b). In fact, this paradigm has revealed itself perfectly suited for the study of sublexical sources of information in a highly controlled condition, in which high-level information is not available (or at least is largely reduced). It also enables the simultaneous assessment of different available cues, both in conjunction (i.e., suggesting the same AL parsing) and in confrontation (i.e., when those cues suggest different, conflicting segmentation hypotheses).
On the grounds of Mattys and colleagues’ (Mattys et al., 2005; Mattys, 2004) theoretical proposal, we have evaluated the weighting (or in Mattys et al.’s terms: the hierarchical organization) of different sublexical cues represented, according to these authors, at different tiers (tiers II and III; see Figure 1 in chapter one of the present thesis) and at the same tier (i.e., segmental and subsegmental information at tier II; lexical stress and universal prosody at tier III) in different (qualitative and quantitative) listening conditions.

7.1. Main findings of the present study

7.1.1. Speech segmentation with physically degraded signal

Experiment 1

In Experiment 1 (see chapter two), within an ALL setting, we investigated, through three signal’s physical quality conditions, the relative impact of two types of sublexical information in speech segmentation. These two sublexical cues are represented at the same middle tier in Mattys et al.’s proposal: statistical information – TPs between adjacent syllables, and subsegmental information – coarticulation.

Mattys (2004; Mattys et al., 2005) had already demonstrated coarticulation reliability in adults, when this cue was in conflict with lexical stress. Our results (see chapter two) thus add to this evidence. In good listening conditions, coarticulation is given priority to either lexical stress (Mattys, 2004; Mattys et al., 2005) or to a general segmental statistical cue (TPs, in Experiment 1). Most importantly, coarticulation overruled statistical information when both cues were put in conflict in
the speech stream, which is in line with Johnson and Jusczyk’s (2001) data on infant listeners. The incongruent cues condition showed that coarticulation overrides TPs only in intact speech, while TPs override coarticulation in strongly degraded speech. As a matter of fact, both Experiment 1 of the present study and Mattys’ data (Mattys, 2004; Mattys et al., 2005) show that coarticulation is much more affected by the presence of physical noise than other cue types. Therefore, the segmentation process adopted by listeners, when both statistics and coarticulation are available in the stream, varies as a function of signal quality (cf. Mattys et al., 2005). Moreover, the weighting change of the available segmentation cues largely depended on both the unavailability of coarticulation and the high resilience of statistical information to physical noise superimposition. The pattern observed with congruent cues suggests an additive effect of converging cues, at least in good listening conditions, allowing the optimization of the speech segmentation process. However, as signal degradation increased, the integration of the available cues was affected and hence the redundancy gain was largely reduced.

In Experiments 2 and 3 (see chapter three), within an ALL setting, we investigated, through different physical quality conditions, the relative impact of two types of sublexical information in speech segmentation, represented at different tiers in Mattys et al.’s proposal: statistical information – TPs between adjacent syllables, and suprasegmental information. Contrary to what is proposed for subsegmental information (e.g., coarticulation), prosody is represented at the lowest tier in Mattys and colleagues’ theoretical framework, and is considered a “last resource segmentation heuristic”, particularly important in physical impoverished conditions. However, the impact of physical noise on suprasegmental cues was, until the present,
only evaluated regarding word primary stress, and only with English listeners (Mattys et al., 2005). On the one hand, Mattys et al. (2005) had already suggested that the weighting of language-specific cues, such as lexical stress, probably depends of their word-boundary predictability in the specific language (cf. Mattys et al., 2005; see section 1.4.2. of the present chapter). Therefore, cross-linguistic differences would be expected on these cues weighting. On the other hand, Shukla et al. (2007) demonstrated that IPs (i.e., one of the highest prosodic structures, that seems to be an universal physiologically-based cue; Grice, 2004) is a preponderant segmentation cue, even when another source of information represented at an above tier in Mattys et al.’s (2005) proposal (i.e., segmental information: TPs between adjacent syllables) is available in the stream and the signal is phonetically intact.

Experiment 2

In Experiment 2 (see chapter three), we investigated the weighting of statistical segmental information (i.e., TPs computation) and universal prosodic cue (by marking prosodic right-edges through syllable lengthening and falling pitch) on an AL speech segmentation.

Universal prosody was not only as resilient to physical degradation of the signal as TPs computation, but it was also a powerful segmentation cue. Indeed, the former cue was able to drive the segmentation processes, even when TPs were incongruently available in the stream, in any physical quality condition. When TPs and prosodic right-edges suggested the same segmentation hypotheses a performance gain was also observed.

Remarkably, in line with Shukla et al.’s (2007; Experiment 4) findings,
listeners’ performance in the ALL test phase was modulated by the type of stimuli with which statistical outputs (i.e., TP-words) were confronted with (i.e., part-words). Recall that, in the forced-choice test, listeners are presented in each trial with a TP-word (i.e., a statistical segmentation output) and a part-word (i.e., AL-stimuli not supported by the statistical segmentation procedure). Therefore, listeners familiarized with the AL in incongruent cues conditions (i.e., in which statistical and prosodic informations suggested incongruent segmentation hypotheses), were presented, in the forced-choice test, with two types of part-words according to their potential support by the prosodic cue. One type of part-words was not supported by any type of information - neither by statistical information (i.e., these part-words spanned a TP-boundary) nor by the prosodic cue (i.e., their last syllable was not prosodically marked as a right-edge and hence they also spanned a prosodic-boundary). The other type of part-words, although not supported by the statistical information, were prosodic cohesive units, since their last syllable was acoustically marked by a right-edge (i.e., the third syllable of these part-words was lengthened and had a falling pitch). In Experiment 2 we have observed the predominance of the universal prosodic cue over the statistical information in both intact and physically degraded conditions.

If listeners adopted an all-or-none segmentation procedure, in the incongruent cues condition we would not expect that listeners’ TP-word choices would be above the chance level. In particular, in trials in which neither of the two AL-stimuli (i.e., the TP-word vs. part-word) corresponded to the byproducts of a prosodically-driven segmentation procedure, listeners’ performance should have been at chance, since in that case neither the TP-word nor the part-word were compatible with prosodically-
driven speech segmentation. In trials in which TP-words were confronted with part-words that were prosodic cohesive units, we would expect to observe a performance below the chance level. Instead, we have found a different pattern of results. In the incongruent cues condition, in the forced-choice test’s trials in which TP-words were confronted with part-words not supported by any available segmentation cue, TP-words were more often chosen as the plausible units of the new language than when TP-words were confronted with prosodically-based part-words (*see* section 3.3. of chapter three). This is an important result because it demonstrates that during speech segmentation processing (even when high-level information is not available, as in ALL settings), when **incongruent cues are available, listeners are also sensitive to the multiple possible byproducts of those** (incongruent) **cues**. This fact might explain the preference for TP-words over part-words that were not supported by any segmentation cue (at least in the incongruent cues – 1st syl condition). Besides, listeners were also able to consider one cue (i.e., universal prosodic information) more reliable than the others (*see also* Shukla et al., 2007; Experiment 4), and in the congruent cues condition a performance gain was observed in comparison to the single cue condition (*see* section 3.3. in chapter three). Therefore, listeners’ performance is **influenced by the reliability of the segmentation cues that underlie the presented potential “words”**. This is in line with Mattys et al. (2005) proposal suggesting that the impact of any speech segmentation cue is a graded rather than an all-or-none phenomenon.

**Experiments 3A & 3B**

In Experiment 3A (*see* chapter three), using an off-line forced-choice task, we
ensured that syllable duration is, in EP, a correlate of word primary stress (see also Delgado-Martins, 2002).

In Experiment 3B the weighting of a statistical segmental cue (i.e., TPs between adjacent syllables) and word primary stress (i.e., language-specific prosodic information) was evaluated, within three input-intelligibility conditions.

The role of language-specific prosodic cue evaluated in Experiment 3B was rather different than the one found in Experiment 2 with universal prosodic information. Indeed, in opposition to what was found in Experiment 2, listeners’ performance was not globally affected by the presence of lexical stress patterns. Yet, lexical stress has a role in speech segmentation when the signal is degraded, possibly constraining the extraction of statistical segmentation outputs to the ones with higher support. With phonetically intact speech, both high- and low-TP-words were correctly extracted from the stream, both independently of the number of cues available in the stream and of the stress location. Notably, with mildly degraded signal, a lexical stress effect emerged, narrowing the extraction of AL-units: low-TP-words, which with intact speech were correctly parsed from the stream, were no longer selected as possible words of the new language, in the mildly degraded condition. Additionally, a strongly degraded signal (i.e., 10dB SNR) maximized the strength of primary stress effects (cf. Mattys et al., 2005). Consequently, in this condition, only the statistical outputs with the highest TPs that obeyed to the default stress pattern in EP were selected as plausible words. In stressed conditions diverging from the default one statistical learning was inhibited.

The availability of a congruent lexical stress pattern did not promote any quantitative benefit in ALL. Instead, it narrowed the selection of statistical
segmentation outputs to the ones which had the strongest support from the conjunction of the two segmentation cues available in the stream, enabling the exclusive extraction of these units (i.e., high-TP-words obeying to the default stress pattern in EP). Remarkably this extraction was as efficient when the signal was quiet distorted as when it was intact.

7.1.2. Speech segmentation in cognitive noise

The importance of the listening conditions in the weighting of various speech segmentation cues is an innovative and fundamental aspect of Mattys et al.’s (2005; Mattys, 2004) proposal. Until now this impact has been only evaluated through physical degradation of the signal (i.e., physical noise). Note, however, that Mattys et al. (2005) already suggested that the listening conditions go beyond this physical quality. Under this view, deteriorating the listening conditions while maintaining the signal physically intact would also affect the kind of cues predominantly used in speech segmentation.

This proposal is based on naturalistic observations. Indeed, in natural settings, speech perception, and hence speech segmentation, often occur not only in conditions of physical noise, but also in conditions of attentional load, far from the optimal conditions of a laboratorial sound-proof room.

The impact of cognitive noise (or attention-load) in speech segmentation was evaluated in Experiments 4, 5, and 6 of the present study.

**Experiment 4**

Experiment 4 (see chapter four) explored if a hierarchical organization of
coarticulation and TPs, similar to the one found with physical noise (i.e., in Experiment 1 of the present work), would also be observed in cognitive noise conditions.

The ALL patterns observed in Experiment 4 have shown that cognitive noise differently affects the use of these two cues. Notably, we have found a weighting change of these sublexical segmentation cues modulated by cognitive noise that tended to mirror the one observed with physical noise: The gradual effect of cognitive noise had a clear impact on statistical information but not on coarticulation. When strong cognitive noise prevented participants to optimally extract TP-units from the stream, listeners did rely on coarticulation when TPs were incongruent with it. This pattern of results is in sharp contrast to the one found with strong physical noise in Experiment 1.

When coarticulation and TPs were congruently available in the AL stream, their integration was also achieved in strongly degraded attentional conditions. Indeed, with strong cognitive noise, a redundancy gain (better performance when both statistics and coarticulation suggested the same word-boundaries than when only TPs were available) was exclusively found for the AL words with low TPs.

A soft reduction of the attentional resources available to speech segmentation through passive exposition to visual stimuli (i.e., weak cognitive noise condition, in Experiment 4) seemed to have the same (at least quantitative) impact as a mild physical degradation (i.e., mildly degraded condition in Experiment 1) of the auditory AL stream. In both cases, the product of the available incongruent cues was disrupted.

Notably, the impoverishment of the statistically-driven segmentation
procedure by *strong cognitive noise did not completely prevent statistical learning*. Although listeners exposed to the AL with only statistical information available in the stream presented an impaired performance in the ALL-test (i.e., they were not able to select low-TP-words as the units of the new language), they still presented an above chance performance for AL units with the highest TPs. Thus, although *statistically-driven speech segmentation* is *affected by cognitive noise* (i.e., a severe reduction of attentional resources available) *it still operates* in these attentional adverse contexts.

In order to *specifically assess* the impact of attention in *statistically-driven speech segmentation*, in Experiments 5 and 6 (*see* chapter five), we *evaluated directly* the role of *amount of familiarization*, and the role of the *two facets of attention* (i.e., as a selective mechanism, and as limited-capacity resources) in TP-based speech segmentation.

### 7.1.3. Cognitive noise and statistically-driven speech segmentation

#### Experiment 5

In Experiment 5, we assessed the extent to which the *amount of AL-familiarization* (i.e., between-participants familiarization phases of 7-min, 14-min, 21-min) would *influence statistically-driven* (i.e., *TPs-based*) *speech segmentation* in *conditions of low-* (i.e., weak cognitive noise) and *high-attention load* (i.e., *strong cognitive noise*).

In the *high-attention load condition*, independently of the amount of
exposition to the AL, participants *always succeeded in extracting the high-TP-words* from the stream. However, it was only with 7-min of familiarization time that they were also able to extract the low-TP-words. *With familiarization time, performance for high-TP-words* actually *increased*, while *performance for low-TP-words decreased*, as though the cognitive system adapts itself to the reduction of attentional resources by focusing on the most salient “lexical” units.

In contrast, in the *low-attention load condition*, the system seems to *privilege the extraction of high-TP-words* only when the familiarization time is drastically shortened (namely, *with 7-min*). Even in *this shortest familiarization condition*, *performance was above chance* level for both *low- and high-TP-words*, as it was in the other familiarization time conditions. In addition, with *increasing familiarization times*, the system is able to *extract both high- and low-TP-words* at similar performance levels, although this procedure seems to be *done at the expense of performance on high-TP-words*.

In Experiment 5 we were able to demonstrate clearly that *attention-load affects statistical computation in a graded* rather than in an all-or-none *manner*. This was already suggested in the strong cognitive noise condition when only statistical information was available (Experiment 4 presented in chapter four). Indeed, *cognitive noise* (or in other words, high-attention load) *has a different impact on the extraction of high- and low-TP-words*. Reducing attentional resources and/or diverting attention from the AL do not completely prevent statistical learning. Instead, the extraction of the most salient units is slowed down, while the extraction of the less salient units is progressively inhibited.

Indeed, it was *in the longer familiarization time* at study in Experiment 5
(i.e., 21-min) that the impact of the reduction of the available resources was more severe. In this condition only high-TP-words were correctly extracted from the stream, leading to the observation of a TP gradient effect. Furthermore, this result also suggests that the TP gradient effect observed in Experiment 4 in the strong cognitive noise condition can not be due to a weak reduction of attentional load.

**Experiment 6**

This experiment was aimed at disentangling the role of attention as a selective mechanism and as a central resource in TPs computation. In Experiment 6 we demonstrated that the efficient operation of the statistical learning mechanism is not a matter of selective attention, but one of attentional resources.

In a condition of severe reduction of the available attentional resources through the requirement to perform another resource-consuming task, as in the high-attention load condition (i.e., strong cognitive noise), statistically-driven speech segmentation operates in an impaired fashion. Additionally, even when participants were focusing their attention on the AL stream (i.e., intentional learning condition), statistical computations still suffered a dramatic impairment as long as the attentional resources were scarce.

The TP gradient effect observed in high-attention load in the TP – single cue conditions (in Experiment 4, 5, and 6) is a deleterious effect of reduced available resources. In order to operate efficiently, the statistical learning mechanism requires at least some attentional resources. When this requisite is not fulfilled, TPs are still computed but in an impaired manner, leading to the inability of correctly extracting from the stream the AL-units with the lower TPs.
Statistical learning occurs even in conditions in which processing the auditory stream is not required and is not beneficial (being presented as potentially harmful) to the performance of an independent task. In that sense, although statistically-driven speech segmentation is dependent of the attentional-resources available, and thus not a purely “stimulus-driven” mechanism (cf. Moors & De Houwer, 2006), it may nevertheless be conceived as automatic. This is in line with Tzelgov’s (1997) and Bargh’s (1992) proposals that an automatic process is one that is autonomous, running without monitoring (it can be incidental), and even when it is not part of participants’ task requirement.

In fact, in the present study we have shown that cognitive noise has a differential impact on the available sublexical cues. It is important to note that the results observed in cognitive noise (high attention-load) conditions cannot be due to a general increase in task difficulty with attentional load. Indeed, in Experiments 1, 2 and 3B, we have demonstrated that sensory degradation (i.e., reducing the physical quality of the signal by white-noise superimposition), which also increased task difficulty, had a different impact on the relative weighting of the available speech segmentation cues. In other words, the observation of different patterns of ALL as a function of the nature of the degradation instigated (i.e., physical or cognitive) rules out general task difficulty as an alternative account for the effects observed in the present study. Both physical and cognitive noise increase general task difficulty but they clearly have independent and different impacts on statistically driven speech segmentation processes and in the weighing of different segmentation cues.

The experiments presented in this thesis and reviewed thus far suggest that the ALL paradigm can be a useful tool. However, this paradigm suffers from one
**particular limitation** regarding the task by which the speech segmentation procedures adopted by listeners during the AL-familiarization phase are accessed. Indeed, the **two-alternative forced-choice test** provides only **indirect evidence of speech segmentation processes**. Furthermore, the status of statistical segmentation outputs in what regards listeners’ linguistic knowledge, in particular in the context of a full mature speech perception system, is far from clear.

7.1.4. The ALL paradigm revised

In *Experiments 7, 8, and 9* (see chapter six) of the present study, we adopted an **ALL approach combined with conventional experimental techniques** in order to **obtain evidence of on-line statistical speech segmentation** and to evaluate whether **statistical segmentation outputs could acquire some lexical status** in the adult speech system.

**Experiment 7**

In Experiment 7, adopting a variant of the word-spotting task, we **directly demonstrated** that indeed listeners are able to **use on-line statistical information on speech segmentation**. In “**AL-word-spotting**”, listeners were both more accurate to spot and faster to recognize TP-words than part-words embedded in four-syllable nonsense strings. This is the first direct evidence that after an AL-familiarization period, in which listeners were never presented with previously segmented AL-stimuli, they were able to correctly extract from a four-syllabic sequence the trisyllabic stimuli that corresponded to AL words.
Experiment 8

In this experiment, TP-words based on the statistical information available in the AL stream also strongly overlapped from onset with real words of listeners’ native language. In this condition, TP-words were not only correctly parsed from the stream, as revealed by the highly accurate performance in the forced-choice test (used as an ALL-index), but these statistical segmentation outputs had also a lexical status, exhibiting a lexical competition signature. Indeed, both immediately and one week after the AL familiarization period, the speed of processing of real (already existing) words, members of the same cohort than the TP-words parsed from the AL stream, was penalized in comparison to the processing of unrelated real words.

These results add to previous findings of word learning in the absence of semantic knowledge (e.g., Bowers et al., 2005; Gaskell & Dumay, 2003a; 2003) and demonstrate that novel lexical items can be acquired in the context of a continuous speech stream, which probably closely resembles natural conditions of word learning. This lexicalization effect found with adult listeners is also in line with infants’ data (Saffran, 2001; Swingley, 2007).

Experiment 9

In the last experiment of this work, we evaluated whether wordlikeness of AL-stimuli by itself, against the statistical information available in the AL-stream, could drive the segmentation processing and whether its byproducts could also be lexicalized.
In this condition, listeners adopted a segmentation procedure compatible with their native language’s specific information. This is in line with previous findings suggesting the preponderance of listeners’ native language information (e.g., coarticulation; prosody; phonotactics; vowel harmony) in AL speech segmentation (e.g., Fernandes et al., 2007; Onnis et al., 2005; Shukla et al., 2007; Vroomen et al., 1998). Nevertheless, the byproducts of this segmentation procedure adopted by listeners in the incongruent cues condition (Experiment 9) were not lexicalized. This is in contrast with what was found in the congruent cues condition (Experiment 8). These results were at first sight quite surprising. However, the difference found in Experiment 9 between the pattern of results on ALL-test and the lexicalization-test is in accordance with the proposal of Leach and Samuel (in press) on the distinction between “items configuration” (in the present study, the phonological representations parsed from the AL stream) and “lexical engagement” (their involvement in lexical dynamics). Indeed, the ALL-effect is the outcome of listeners’ ability to use the available segmentation cues on-line, as it was demonstrated in Experiment 7 of the present study, which corresponds to listeners knowledge of those AL-words’ configuration. However this information cannot predict either the potential lexical status or the lexical engagement of the segmentation outputs.

In natural settings listeners are exposed to a rich combination of different signal-derived (and also lexically-driven) segmentation cues, and hence, it would be counter-productive if all phonological forms to which listeners are exposed would be stored as different categories in mental lexicon (cf. Gaskell & Dumay, 2003b). The pattern of results found in Experiments 8 and 9 is in line with the suggestion that the consistency of segmentation hypothesis proposed by different available sources of
information can have an important role, possibly modulating the potential lexicalization of the segmentation outputs.

7.2. Theoretical framing of the present results

The pattern of results reported in this thesis is generally in agreement with Mattys and colleagues’ (2004; Mattys et al., 2005) proposal. Indeed, we have observed through these experiments that the segmentation process adopted by listeners varies as a function of both the types of cues available in the speech flow and of the listening conditions. Additionally, it was shown that the influence of noise (whether physical, as in Experiments 1, 2 and 3B, or cognitive, as in Experiments 4, 5, and 6) also varies in a graded, rather than in an all-or-none, manner.

Note, however, that in conditions in which segmental (phonotactic, cf. McQueen, 1998) and subsegmental (coarticulation, cf. Mattys, 2004) information were independently put against metrical prosody, both overruled the last cue. This seems to support a broad distinction between sub-lexical (either segmental or sub-segmental) cues and prosodic information. Considering suprasegmental information, in the present study, different weightings were observed depending on the particular type of prosody cue (i.e., word primary stress and universal prosody) available in the stream. This could, at first sight, suggest that within the tier at which prosodic information is represented, a sub-division would be required. However, we do not advocate the exhaustive “listing” of sublexical cues and theoretical modifications based on each new segmentation cue studied. Furthermore, at an acoustic level, the link between acoustic correlates and the sublexical information is not always
straightforward. For example, some acoustic factors frequently manipulated as correlates of suprasegmental information (e.g., syllables duration) are also related with other types of sublexical cues, such as subsegmental information. Indeed, while findings suggest that differences in the duration of segments and syllables are used by listeners in speech segmentation’s assistance, authors have considered them related with subsegmental (e.g., Davis et al., 2002) and suprasegmental prosodic information (e.g., Salverda et al., 2007). Since Mattys et al. (2005) already demonstrated that subsegmental and suprasegmental information have independent and different impacts modulated by listening conditions, caution is needed in the consideration of acoustic correlates and their relation with different sources of information.

Does the specific pattern of results imply that the organization of segmental, subsegmental and suprasegmental information posited by Mattys et al.’s (2005) should be rearranged? Providing a definite answer to this question is difficult on the basis of the available evidence. However, we do not believe that the present results are compatible with that sort of theoretical restructuring. Instead, we propose that according to the data reported in the present thesis, the differential weighting of cues and their integration (whose outcome corresponds to speech segmentation) may depend not only on the structural grain of the cues (cf. Mattys et al., 2005), but also on two basic factors and their relationship with the modular speech perception system (cf. Fodor, 1983; 2000). These factors, i.e., the domain-generality of the cues and their role in specific languages – universal vs. language-specific cues – allied with the hierarchical organization proposed by Mattys and colleagues could accommodate not only the results presented in this thesis but also other recent
evidences (e.g., Shukla et al., 2007).

7.2.1. A theoretical proposal – the nature of the available sources of information

**Domain-Generality**

The first factor that probably underlies the weighting of the available cues in different listening conditions regards their *domain generality*. Indeed, the ability to track TPs involves a general-domain learning mechanism, as demonstrated by the fact that TPs are also extracted in tone (Saffran et al., 1999) and visual (Fiser & Aslin, 2001) sequences. On the contrary, coarticulatory information and metrical prosody are speech-specific.

This factor possibly underlies the weighting of cues in degraded (both in physical and cognitive noise) conditions. Speech-specific cues are probably processed in an encapsulated manner as part of the modularity speech perception system (for a review on evidences of modularity of the speech perception system see Trout, 2001) and hence highly independent of “horizontal faculties” (Fodor, 1983). This would confer speech-specific cues a higher weighting in speech segmentation than general domain ones have, as long as the reliability of the former is not weakened by the listening conditions (see Figure 13 for an illustration of this proposal).

**Speech-specific (Universal vs. Language-specific) cues**

Within the speech domain, it might be useful to consider the further
distinction between *universal cues*, like intonational phrases, which partly correspond to physiological mechanisms like breath groups (Grice, 2004; Shukla et al., 2007), and *language-specific cues*, namely properties that depend on the particular language (see Figure 13). The latter properties obviously need to be learned. Both domain-general statistical mechanisms and speech-specific universal cues are available since a very early phase in language acquisition (e.g., Dehaene-Lambertz & Houston, 1998; Thiessen & Saffran, 2003) and are also used by adults even with unknown or foreign languages (Shukla et al., 2007). While universal cues could maintain an important role in sublexically driven speech segmentation processing through all linguistic development, the weighting of language-specific cues probably depends on their word-boundary predictability in a specific language (cf. Mattys et al., 2005). For example, since lexical (or word primary) stress is an abstract property not always acoustically realized, the possible effectiveness of a primary stress segmentation procedure in the context of the adult speech perception system is obviously reduced. This holds true, particularly in languages (like EP) in which stress location seldom indicates a word-boundary. These cues would occupy a particular role during a transitory phase in linguistic development, being lower weighted in adulthood.

**The impact of noise**

Both general-domain statistical mechanisms and speech-specific universal cues might have a central role in development as guides to other segmentation cues. To be reliable guides, such cues should be relatively immune to signal degradation. In other words, as suggested by Mattys et al.’s (2005), the more resilient (but lower-
weighted in normal listening conditions) cues in adult speech segmentation could correspond to the earliest acquired, and hence, to the most critical cues at the onset of language development. Therefore, both general-domain and universal speech-specific cues would be highly resilient to physical degradation of the signal, with the latter occupying a preponderant role in segmentation over the former, probably due to its encapsulated processing. The role of sublexical language-specific cues would be, as already proposed by Mattys and colleagues, highly dependent of the listening conditions (Fernandes et al., 2007; Mattys et al., 2005).

Figure 13: The theoretical proposal outlined in the present thesis allied with Mattys and colleagues proposal for language-specific cues (see also Figure 1 in section 1.4. of chapter one).

The pattern of results found in the present study is in line with this theoretical proposal. In Experiment 2, it was observed that both TPs computation (a general-domain cue) and universal prosody (a speech-specific cue) are very resilient to the physical degradation of the signal, being able to drive speech segmentation at similar levels with intact and strongly impoverished signal. Furthermore, the predominance of coarticulation over statistical information in cognitive noise, observed in
Experiment 4, is also consistent with these cues’ distinction on modularity grounds. On the one hand, with physical noise, the fundamental statistical cue overruled coarticulation. This was probably due to listeners’ inability of processing low-level, fine-grained acoustic aspects required for speech segmentation driven by coarticulation, in physically impoverished conditions. On the other hand, in cognitive noise, coarticulatory processing could not be “switched-off” simply by drastically reducing the available attentional resources, and thus it maintained its ability to drive speech segmentation.

The integration of the available segmentation cues

In the present study we have suggested that the two factors mentioned above, incorporated in the theoretical framework of Mattys and colleagues (see Figure 13), could apprehend, not only the role of different cues, but also their weighting in different listening conditions in speech segmentation’s assistance.

Shukla et al. (2007) have recently proposed a different approach to the integration of the available segmentation cues. These authors suggest that universal prosody could act as a filter for the output of statistical computations, suppressing the TP-words that are incompatible with prosody segmentation hypotheses. Thus, although all TP-words would be extracted from the stream, listeners would only consider the statistical outputs compatible with the prosodic filter as the ones corresponding to a correct parsing.

This proposal could be, of course, extended to any speech-specific cue, such as coarticulation, lexical stress or any other cue. However, Shukla et al.’s proposal does not seem to be able to accommodate the results presented in this work. If
prosody was to be used only as a “higher” filter of statistical outputs, allowing or suppressing the units extracted on a TP-basis, then when TPs and other sublexical cues suggested the same parsing (i.e., congruent cues condition; e.g., TPs and coarticulation in Experiments 1 and 4; TPs and universal prosodic cues in Experiment 2; TPs and wordlikeness in Experiment 8) no redundancy gain would be observed. In fact, in Experiment 2, a redundancy gain was observed both with intact and physically degraded signal, that was promoted by the congruent availability of TPs and a universal prosodic cue. In Experiment 1, a redundancy gain derived from the congruent availability of TPs and coarticulation was also observed in intact conditions.

Additionally, universal prosody was also able to drive the segmentation process even when no statistical outputs were congruent with it and hence no TP-words were compatible with the universal prosodic segmentation procedure. The same pattern of results was observed in intact and strong cognitive noise conditions when, instead of universal prosody, coarticulation was the cue available in the AL stream. Moreover, three findings of Experiment 3B suggest that language-specific prosody does not act as a filter of statistical segmentation outputs. First, stress pattern had no impact on listener performance with intact speech. As a matter of fact, all listeners independently of the presence (/absence) and location of the primary stress’ (either in the first, second, or third syllable of TP-words) were able to correctly extract from the stream the statistical segmentation outputs, at the same level. Second, stress pattern effects only emerged in degraded signal, in line with Mattys and colleagues proposal (Mattys, 2004; Mattys et al., 2005) and were maximized in the strongly degraded (i.e., 10dB SNR)
condition. Third, when both statistical and stress cues were available in the stream and the strong physical impoverishment (i.e., 10dB SNR) enabled stress cues to operate as efficiently as the statistical learning mechanism, only the units parsed from the stream that were highly compatible with the conjunction (i.e., intersection) of the available cues were considered by listeners reliable outcomes of the segmentation process.

Thus, speech segmentation does not seem to depend on one predominant cue acting as a “higher” filter of the outputs of another lower weighted cue. Instead, a more parsimonious account would suggest that the outcome of speech segmentation is largely the result of the conjunction of the available cues, both in accordance with their nature and their weighted reliability.

Therefore, when lower weighted cues are available in the stream, only units strongly supported by the conjunction of those cues will be considered. In particular, as it was observed in the strong degraded condition in Experiment 3B, when the available cues were lower weighted and thus weakly reliable in speech segmentation, as it is the case of lexical stress – the “last segmentation resource heuristic” (cf. Mattys et al., 2005; see also Valiant & Levitt, 1996) and general domain TPs computation, only the units extracted from the stream that are strongly supported by both cues in conjunction were considered likely correct. In that case, no redundancy gain should be observed, as it was found in Experiment 3B.

As long as a reliable cue is available in the speech stream (e.g., in intact speech, coarticulation: Experiment 1; in any physical condition, universal prosodic cues: Experiment 2), its byproducts are considered highly reliable. If these units are also supported by a lower weighted cue, such as the general domain TPs, a
redundancy gain will probably be observed. Indeed, in Experiment 1 of the present study when both coarticulation and TPs were congruently available in the stream, with intact speech, an additive effect (for both high- and low-TP-words) allowed the optimization of speech segmentation processing. Accordingly, in Experiment 2, when TP-words were acoustically marked by a universal prosodic edge, their correct parsing was also optimized. In Experiment 8, when the statistical segmentation outputs also strongly overlapped with existing words, listeners’ ALL performance reached a high level: note that 17 out of 32 participants chose TP-words as the “lexical units” of the new language in at least 90% of the forced-choice test trials.

As already suggested by Christiansen and colleagues’ computational work (Christiansen et al., 1998; Christiansen & Curtin, 2005), a cue that insures a deeper encoding of structural regularities of the input also enables the reliance on more subtle aspects of the input for making correct predictions. Thus the integration of different cues does not necessarily promote a quantity gain (i.e., more units parsed from the stream) but it can promote a qualitative one (i.e., the correct parsing and deeper encoding of highly supported units). Consequently, the correct extraction of the byproducts of the conjunction (i.e., intersection) of the available sources of information can minimize errors and unwanted over-generalizations. This holds true, particularly, in conditions in which the available cues are lower-weighted.

The fact that speech segmentation is a robust phenomenon emerges from the pattern of results presented in this study. In other words, as already suggested on Mattys and colleagues’ proposal (Mattys, 2004; Mattys et al., 2005) and supported by studies of anatomical and functional neural organization (for a review see Scott &
Johnsrude, 2003), speech perception (and hence speech segmentation) is very resilient to different listening conditions. The availability of multiple cues (both signal-derived and lexically-driven), differently weighted according to their nature (whether the tier at which they are represented, Mattys et al., 2005, or their distinction on domain-generality and their role in particular languages, as suggested in the present chapter) enables the listener to rapidly fulfill speech segmentation task in any listening conditions. In other words, the robustness of speech segmentation processing seems to be due to the fact that congruently available cues can act as complementary sources of information, being differential weighted according to the type of impoverishment (cognitive or physical) instigated on listening conditions. Therefore, an apparent paradox is observed in degraded conditions when different congruent sources of information (in the present study, sublexical cues) are available in the speech stream. Noise modulates the weighting of these segmentation cues, but at the same time, it does not obligatorily impair speech segmentation in a drastic way (note for example that in Experiments 1, 2, and 4 of the present study, listeners exposed to the AL with congruent cues available, had an ALL performance above chance in any one of the degraded listening conditions). This is in line with Mattys’ (2004, Mattys et al., 2005) proposal that speech segmentation is largely the product of listening conditions and the available segmentation cues. Therefore, at least at the levels of noise studied in the present experiments, speech segmentation was achieved even in degraded conditions due to the fact that lower weighted cues were called upon to drive speech segmentation in impoverished conditions and thus disabled any observation of a main impact of noise.

The present study also demonstrated that TPs computation still play an on-
line role (see Experiment 7 presented in chapter six) in language processing in adults. For sure, TPs computation plays a less fundamental role in adulthood than in infancy. Nevertheless, adults are still sensitive to statistical information such as conditional probabilities, not only in speech processing (e.g., Saffran et al., 1996a; Vroomen et al., 1998), but also in other sequence learning tasks (e.g., Boyer et al. 2005; Conway & Christiansen, 2005; Fiser & Aslin, 2001; Perruchet & Pacton, 2006). Indeed, since the intensely dynamic environment that surrounds us mainly involves stimuli occurring in temporal and/or spatial sequences, statistical learning would enable adults to rapidly structure novel sequences of events into emergent units (i.e., unitized into sets of temporally or spatially contiguous events). More generally, the process of reduction of uncertainty (Gibson, 1991) drives learners to seek invariant structure in the stimuli array (Gómez, 2002).

However, the nature of the phonological representations corresponding to statistical segmentation outputs was until the present largely obscured. In other words, at least for adults, whether the statistical segmentation outputs could actually be lexicalized was largely unknown. In order to shed light on this aspect in the context of a full mature speech perception system we designed the two last experiments of the study reported in this thesis.

7.2.2. The nature of statistical segmentation output

In Experiments 8 and 9, adopting the approach of Gaskell and Dumay (2003a; 2003b) within an ALL setting, we demonstrated that, at least in some conditions, the byproducts of statistical segmentation can be rapidly integrated in listeners’ mental lexicon, exhibiting lexical competition signatures.
The results found in Experiment 8 provide an important contribution to the understanding of the nature of statistical segmentation output, adding to Saffran’s (2001) and Estes et al.’s (2007) findings with infant listeners. The output of the statistical learning mechanism can have a lexical status, being rapidly integrated in mental lexicon, even in the context of a fully mature perception system.

Notably, the lexical engagement of the output of statistical segmentation observed in Experiment 8 is long-lasting and it continued to be strengthening between the first (immediately) and second (post-one week) moments of testing. This pattern of results is in line with Dumay and Gaskell’s (2007) proposal. It is also in accordance with general evidences of sleep being implicated in memory consolidation (for a review see Walker & Stickgold, 2004) and, in particular, in word learning (Dumay & Gaskell, 2007; Fenn, Nusbaum, & Margoliash, 2003; see also Clay, Bowers, Davis & Hanley, 2007) and vocabulary growth (Gais, Lucas, & Born, 2006).

Models of spoken word recognition must be able not only to apprehend how different sublexical sources of information are weighted (or in Mattys et al., 2005 theoretical terms, are “hierarchically organized”) in speech segmentation, but also what their role (and the one of their congruency) in word learning. Regarding the role of segmentation cues in word learning, it is important to consider their involvement in structural changes with consequences on how listeners represent speech at different (e.g., sublexical, lexical) levels (see e.g., Sumner & Samuel, 2007), as it was the case in Experiment 8 of the present work.

Future work should be aimed at examining which computational models may most adequately account for such word learning effects occurring in the context of
speech segmentation and, in particular, for such rapid acquisition and integration of statistical information. While any discussion of this point remains highly speculative in the absence of precise simulations, it seems to us that the Adaptive Resonance Theory or ART (Carpenter & Grossberg, 2003; Grossberg, Boardman, & Cohen, 1997; Grossberg & Myers, 2000; see also Sumner & Samuel, 2007, and Vitevitch & Luce, 1999) might be a good candidate to apprehend word recognition and word learning within a continuous flow, as well as to account for the present findings.

In particular the ARTWORD model (Grossberg & Myers, 2000) which extends the earlier ARTPHONE model (Grossberg et al., 1997), could integrate word recognition and word learning within a continuous speech flow (Carpenter & Grossberg, 2003). ART is also able to explain the pattern of results found in both the ALL-test and the lexicalization of the Experiments 8 and 9 of the present study.

**The ART**

For the ART model, the resonance created by the match between bottom-up information and top-down expectations (stored in LTM) is responsible for the creation of the conscious spoken percept (Grossberg et al., 1997; Grossberg, 2003). While the spoken input unfolds, it activates speech segments that in turn activate “item” nodes. “Items” processed through time, generate an evolving spatial pattern of activation across WM that represents item information (which items are stored) and temporal order (the sequence in which they are stored). This enables items to be grouped, or unitized, into categories or “list chunks”. These “list chunks” have their maximal length restrained by WM, but can represent items or larger groupings of variable length (e.g., phonemes, syllables or words), with all of them represented
within the same masking field.

In ART, masking denotes inhibition between “list chunks” (of different lengths): the longer the list chunks, the stronger will be the inhibition (or masking) promoted. However, list chunks will only fire when enough bottom-up evidence is received, which occurs after all the items in the list had been activated. Thus list chunks begin sending top-down feedback (representing a learned expectation stored in LTM) to associated items that are currently stored in WM. Chunks whose top-down signals are best matched to the sequence of incoming data reinforce the WM items and, in turn, receive greater bottom-up signals in return, winning masking field competition and thereby selectively amplifying and focusing attention upon consistent WM items, while suppressing inconsistent ones. This resonance loop will result in an equilibrated resonant state, thereby creating an emergent conscious percept.

In order to parse a continuous stream into multiple discrete words, the positive feedback loop of any resonance cannot continue indefinitely. ART explains it as an example of “resonant reset” (Grossberg, 2003). “Mismatch reset” occurs when new phonemic information arrives and is sufficiently different from the current activated WM pattern to warrant an arousal burst that rapidly resets activity in the masking field. The network is “reset” into a nonresonant state, so that the next resonance can be initiated.

In ART, the resonant state also drives the learning process (Carpenter & Grossberg, 2003). When novel words are presented in the speech stream, as it happens during an AL-familiarization period, novel information cannot form a good enough match with the expectations that are read-out by previously learned
recognition categories. This will trigger a memory search, or hypothesis testing, that leads to selection and learning of a new recognition category that really matches the input. Since in the masking field, longer list chunks will promote stronger inhibition, a chunk that codes a longer novel list like any one of the trisyllabic TP-words of the ALs used in this study, will have an a priori advantage over chunks that code shorter lists, like the familiar nonsense syllables that constitute it. It is precisely the masking advantage of longer list chunks that enables novel words to successfully compete with “amplifier” familiar chunks of shorter lists. When a top-down expectation achieves a good enough match with bottom-up data, this matching process focuses attention upon those feature clusters in the bottom-up input that are expected. If the expectation is close enough to the input pattern then a state of resonance develops as the attentional focus takes hold, and the novel word is learned. The resonant process will be interrupted or terminated by mismatch reset and hence a new speech segmentation point will be defined. The repeated exposure to specific spatial patterns (e.g., the repeated occurrence of TP-words in the speech stream) permits learning by the LTM traces in the adaptive pathways between item nodes and the list nodes, reflecting the ALL and lexicalization effects found in Experiment 8 of the present study. ART is also able to explain the unobservation of lexicalization effects when incongruent cues were available in the AL stream (see Experiment 9 presented at chapter 6 of this thesis). When different informations are available in the stream, suggesting different competing chunks of the same length (e.g., TP-words and Part-words, in conditions in which both are supported by the available incongruent cues), resonance is actively reset by input mismatching even before reaching a stable state. The simultaneously activated competing list chunks promote only partial matching
resonant loops and hence the new information cannot be stably stored in LTM. The competitive chunks will strongly inhibit each other based on shared items and although both promote resonant loops, an equilibrated resonant state will not be achieved, and consequently novel words will not be stored in LTM.

ART model resembles the PARSER model (Perruchet & Vinter, 1998) in its focus on *primitive chunks* and on the importance of selective attention in chunks formation, although PARSER does not propose any mechanism of word learning.

Perruchet and colleagues (Perruchet, 2005; Perruchet & Pacton, 2006; Perruchet & Vinter, 1998) proposed that simply relying on chunk formation through the repeated occurrence of stimuli is sufficient to explain statistical learning effects. This obviates the need to postulate that listeners are able to perform statistical computations such as the ones involved in TPs. Based on this fairly simple algorithm PARSER could represent an *elegant* proposal, challenging the statistical learning approach (e.g., Cleeremans et al., 1998; Saffran et al., 1996a; 1996b). However, the pattern of results presented in this thesis does not seem to be compatible with Perruchet and colleagues’ (Perruchet, 2005; Perruchet & Pacton, 2006; Perruchet & Vinter, 1998) proposal.

**The PARSER**

PARSER (Perruchet & Vinter, 1998) is a fragment-based model that proposes that the knowledge acquired in statistical learning tasks might result from simple learning algorithms, giving rise to little more than explicitly memorized short fragments or “chunks”. Thus sensitivity to statistical structure is conceived as a simple by-product of chunk formation (*see for this discussion*: Bonatti et al., 2006;
and Perruchet et al., 2006). In PARSER (Perruchet & Vinter, 1998) chunks are formed from the outset on a random basis, as a natural consequence of the capacity-limited attentional processing of the incoming information, with each chunk corresponding to one attentional focus. As a consequence of their temporal proximity (in language or in any temporal sequential learning; Perruchet & Pacton, 2006), chunks become the constituents a new representational unit and their future depends on the laws governing associative learning and memory. If the association between primitives that form a percept is not repeated, the internal representation created by this percept rapidly vanishes, as a consequence of both natural decay and interference with similar material processing. If the same percept reoccurs, the internal representation is progressively strengthened. The probability of repeatedly selecting the same group of syllables is higher if these syllables form intra-word rather than between-words components.

This fairly simple learning algorithm provides an alternative mechanism underlying the observation of listeners’ statistical sensitivity in speech segmentation, and can explain (at least) some statistical learning results (e.g., Perruchet et al., 2004; see also Dahan & Brent, 1999; Valian & Levitt, 1996). Remarkably, PARSER can also account for the moderate level of performance reported in incidental learning conditions (i.e., Saffran et al., 1997). In fact, a reduced amount of attention (i.e., processing only a fragmentary part of about 3-5% of the AL presented) was sufficient for PARSER to mimic Saffran et al.’s results. Thus, a framework for the word segmentation tasks grounded on the properties of attentional mechanisms, such as PARSER, would naturally predict a general degradation of performance under attentional-load conditions.
Due to PARSER emphasis on the role of attention in statistical learning, with this kind of learning being ruled by general laws of learning and forgetting, this fragment-based model is able to correctly predict a global degradation of performance under attentional-load conditions. Additionally, it is also compatible with the absence of an overall performance gain with the amount of familiarization, as observed in the present study (see Experiment 5 presented in chapter five).

However, this fragment-based proposal has two major problems denoted in the pattern of results found in the study reported in this thesis. In particular, the impact of attention-load; and the observation of TP-gradient effects (i.e., better performance for high- than for low-TP-words) that do not dependent of AL-stimuli frequency of occurrence.

First, PARSER would predict that the effect of amount of familiarization in ALL would be a general one: both high- and low-TP-words, which had the same raw frequency, would have the same cost/benefit allied with an increase of AL familiarization time durations. This was clearly not the case in Experiment 5. Note that the differential impact of familiarization time on high- and low-TP-words suggests that statistically-driven speech segmentation cannot be confined to chunk formation.

Second, if statistically-driven speech segmentation is instead based on the statistical properties of the stimuli and not in “chunk formation”, participants’ performance should reflect those statistical properties. In particular, when TPs are the only information available in the stream, with raw frequency of the AL units controlled (Aslin et al., 1998), statistical learning is still observed. Additionally, Saffran et al. (1996b) already demonstrated that when only statistical information is
available in the stream, listeners’ ALL performance is modulated by the TP-level of AL-words. In Saffran et al. (1996b) listeners showed better performances on AL-words with the higher-TP (from 1.0 to 0.75) that on the lower-TP AL-words (from 0.50 to 0.37). One should remark that PARSER failed to reproduce this TP-gradient effect of the empirical data (Perruchet & Vinter, 1998). However, Perruchet & Pacton (2006) replied that taking into account the effect of interference promoted by forgetting, since the sequential material used in statistical learning is typically generated by recombining a reduced number of primitives (e.g., syllables), is sufficient to make chunk strength sensitive to TPs, thus, revealing the TP-gradient effect observed by Saffran et al. (1996b). Yet, the sensitivity to TPs observed in Perruchet and colleagues model is based on the different raw frequencies of the AL material (i.e., half of the AL-words had higher frequency of occurrence than the other half, as in Aslin et al., 1997). Thus, the observation of TP-gradient effects, even when high- and low-TP AL-words have the same raw frequency (e.g., Saffran et al., 1996b), pose some problems to Perruchet and colleagues’ interpretation. Interestingly, whereas in Experiment 1 of the present study we did not find any hint of a TP-gradient effect on listeners’ ALL performance, when listening conditions were degraded by attentional load (in Experiments 4, 5, and 6 of the present study), or when a language-specific prosodic cue (i.e., word primary stress) was made reliable by physically impoverishment of the signal (in Experiment 3B), we also observed a TP-gradient effect on listeners’ ALL performance that can not be subsumed under the frequency of AL-stimuli. Note that in the present study both high- and low-TP-words had the same frequency of occurrence in the AL stream.

Additionally, human listeners (adults and 8-month-olds) are sensitive to
nonadjacent dependencies that can not be predicted based on adjacent ones (Gómez, 2002). Thus, learning effects based on nonadjacent TPs (e.g., Kuhn & Dienes, 2005) also seem to challenge the chunking explanation, since those computations are not based in contiguous primitives.

The pattern of results found in the work reported in this thesis cannot be explained by raw frequency differences of TP-words and part-words. Frequency effects are well known in language processing (e.g., Brent & Cartwright, 1996; Magnuson et al., 2003), and are also able to explain some ALL patterns (e.g., Dahan & Brent, 1999; Perruchet & Vinter, 1998; Vallian & Levitt, 1996). Yet, if the participants had only used the frequency of the trisyllabic AL items to analyze the speech stream, the performance should have been similar in all between-participants conditions at study in this thesis. Thus, all groups should have equally learned the AL, with performance probably depending exclusively on task difficulty (i.e., signal-quality and attentional load). On the opposite, the ALL patterns found in the present study differed both within and between experiments, depending on the segmentation cues available in the speech stream and on the listening conditions’ contingency.

Therefore, considering the similar emphasis on chunk formation made by PARSER (Perruchet & Vinter, 1998; i.e., primitive chunks) and ART (Grossberg et al., 1997; i.e., list chunks), it is difficult to understand how these models could explain both TP-gradient effects (better performances for AL-words with high- vs. low-TPs, while their raw frequency is held constant; e.g., Saffran et al., 1996b; and the present study) and statistical learning based on nonadjacent TPs (e.g., Kuhn & Dienes, 2005).

The present results may also help to shed some light on the nature of the
processes that are responsible for statistical learning. Indeed, the observation of TP gradient effects and the differential effect of attentional load as a function of TP level is (as already commented on) difficult to reconcile with fragment-based models of implicit learning, such as PARSER (Perruchet & Vinter, 1998). The present results suggest that participants’ sensitivity to statistical information is probably based on distributed knowledge that is represented as subsymbolic information (Cleeremans et al., 1998; Cleeremans & Jiménez, 2002). Accordingly, neural network models (e.g., Cleeremans & McClelland, 1991) assume that this incidental sequence knowledge, such as the one emerging by statistically-driven speech segmentation, is in essence a side-effect of information processing over the course of time, and could arise instantaneously from any interaction with structured environments (Cleeremans et al., 1998; Cleeremans & Jiménez, 2002).

7.3. Some unanswered questions

The issues of compatibility of sublexical cues with prelexical representations, and the phylogenetic root of statistical computations were not directly approached in the study reported in this thesis. However, due to their connection with the sublexical segmentation cues studied in the work here presented, those matters will be considered in this section.

7.3.1. Compatibility between sublexical segmentation cues, prelexical units and lexical mechanisms

Sublexical segmentation procedures are attractive because they provide a
unified solution to the segmentation problem of both infant and adult listeners. Furthermore, these procedures are neither incompatible with lexical competition (e.g., McQueen et al., 1994; Norris et al., 1995), nor with prelexical representations (e.g., Kolinsky, Goetry, Radeau, & Morais, 2000; Mattys & Melhorn, 2005).

On the one hand, a compromise between signal-derived and lexically-driven mechanisms seems to be established through all linguistic development (i.e., from infancy: Bortfeld et al., 2005; Swingley & Aslin, 2007, until adulthood: Davis et al., 2002; Gow & Gordon, 1995; Salverda et al., 2007). Recent evidences suggest that early speech segmentation abilities in infancy are directly related with vocabulary growth (Newman et al., 2006), and hence with lexical acquisition. In adulthood, sublexical segmentation procedures seem to constrain the alignment process in speech segmentation (e.g., McQueen et al., 1994; Norris et al., 1995) and also to modulate lexical activation (e.g., Davis et al., 2002; Gow & Gordon, 2003; Salverda et al., 2003; 2007).

On the other hand, sublexical segmentation cues are also compatible with a prelexical stage of processing in which the output of the auditory system would be classified into a sequence of prelexical units (e.g., syllables).

A clear distinction needs to be made in what regards the classificatory “units of perception” (i.e., the building blocks of prelexical processing, Kolinsky, 1998) and sublexical segmentation procedures (see Mattys & Melhorn, 2005). While the output of perceptual stages corresponds to the building blocks, segmentation cues (e.g., stress: Mattys, 2000; acoustic-phonetic cues: Davis et al., 2002; Mattys et al., 2005) may also involve units (e.g., syllables) as well, but these are often considered regarding their word-boundary predictability: i.e., where in the speech stream the
process of lexical access could be initiated. Therefore, segmentation cues often regard the discovery of where word boundaries are located, while perceptual units correspond to what intermediate representations possibly mediate lexical access. Therefore, the process of classification is obviously distinct from the process of segmentation, and segmentation units can differ from classificatory ones (Kolinsky, Morais & Cluytens, 1995).

**Segmentation (where) and classification (what) units**

Whereas classification inevitably entails segmentation, the reverse is not necessarily true. For example, a prosodic segmentation procedure such as the MSS (Cutler & Norris, 1988) is a segmentation device based on the rhythmic properties of listeners’ native language (e.g., in English and in Dutch, at the onset of any strong syllable; Cutler & Norris, 1988; Vroomen, van Zon, & de Gelder, 1996). Its role is to indicate where in the speech stream the process of lexical access could be initiated. Yet, this segmentation device does not provide any information about the prelexical representations (e.g., feet, syllables) of such classification. In this sense, segmentation procedures are compatible both with models of speech perception involving prelexical stages as well as with models involving no prelexical unit.

Note however that word recognition would benefit if the speech code could be “broken” prelexically. The mediation of a prelexical level of representation (perhaps separated at sub-stages, McQueen, 2005) between the output of the auditory system and the lexicon could remove considerably redundancy that otherwise would have to exist at lexical level (but see Goldinger 1996; 1998). There is no general consensus, however, in what regards the size of these prelexical units.
The size of prelexical units

Among proposals of prelexical units, the most influential is probably the one of Mehler and colleagues (Mehler, Dommergues, Frauenfelder, & Segui, 1981). In a version of the monitoring paradigm (i.e., in the fragment detection task), response times of French listeners were longer when the target (e.g., pa) and the carrier word (e.g., pal#mier\(^{25}\)) did not match on syllabic structure than when they did (e.g., pa in pa#lace; pal in pal#mier). This “syllabic effect” was taken as evidence for a syllabification procedure and hence for the existence of classification prelexical units (i.e., the syllable) that would operate during the process of word recognition:

“(…) the syllable is probably the output of the segmentation device operating upon the acoustic signal. The syllable is then used to access the lexicon…” (p. 342, Mehler et al., 1981).

However, Cutler, Mehler, Norris, & Segui (1983) demonstrated in a cross-linguistic study with French and English listeners that the syllabic effect was not (or at least it did not seem to be) universal. Indeed, English listeners did not present the syllabic effect observed in French listeners. This cross-linguistic difference was interpreted as a consequence of the specific structure of the two languages, probably based in their rhythmic properties. In French (a syllable-timed language), listeners segment the speech stream into syllables, since their language displays a clear and reduced variety of syllabic structures. Yet, in English (a stress-timed language), listeners would not use this syllabic strategy because it is not suited to a language presenting widespread ambisyllabicity and a large variety of syllabic structures. The

\(^{25}\) The “#” marks a syllabic boundary.
study of Cutler et al. (1983) introduced the shift on the focus of research from the universal perceptual building block to language-specific segmentation strategies (see also Kolinsky et al., 2000).

The syllabic effect: compatible or incompatible with sublexical cues?

Syllables have often been found to play a significant functional role in spoken word processing, not only in Romance Languages (e.g., in French: Dumay, Benraïss, Barriol, Colin, Radeau, & Besson, 2001; Dumay, Banel, Frauenfelder & Content, 1998; Kolinsky et al., 1995; Mehler et al., 1981; Pallier, Sébastian Gallés, Felguera, Christophe & Mehler, 1993), but also in other, non-Romance ones, like English (e.g., Finney, Protopapas & Eimas, 1996; Mattys & Melhorn, 2005) and Dutch (e.g., Zwitserlood, Schriefers, Lahiri, & van Donselaar, 1993). However, even in Romance Languages, the syllabic effect is not systematically observed in all experimental conditions nor with different materials (e.g., Content, Meunier, Kearns, & Frauenfelder, 2001; Sebastian-Gallés, Dupoux, Segui, & Mehler, 1992).

Morais, Content, Cary, Mehler, & Segui (1989), demonstrated the important role of syllables in Portuguese speech processing, using a variant of the sequence monitoring paradigm. However, using the migration paradigm (for an overview see Kolinsky & Morais, 1996), Kolinsky et al. (1995; Kolinsky, 1998) reported that in Portuguese (for both EP and BP) syllables seems to have a less important role than other sub-syllabic units. Indeed, in Portuguese the initial consonant is the property that blends the most.

Is it possible that listeners (at least Portuguese listeners) could be sensitive to different prelexical units? Kolinsky (1998) already suggested that all listeners are
able to use different types of routines (e.g., syllabic and subsyllabic) according to task requirements (for a striking example see Goldinger & Azuma, 2003) and language-specific strategies. However, in what regards the sequence monitoring paradigm, its use to study pre-lexical units can be seriously questioned, since it probably involves metaphonological processing (Kolinsky, 1998). Therefore, while results with the fragment monitoring task could reflect listeners’ strategies rather than underlying speech representations, the results found with tasks that do not require intentional retrieval of sublexical units (e.g., the migration paradigm) reduce the probability of the use of strategies and late (e.g., metaphonological) representations (see e.g., Kolinsky, 1998; Kolinsky et al., 2000). Interestingly, using the migration paradigm the role of the initial consonant was also found in EP pre-literate children (Castro, Vicente, Morais, Kolinsky, & Cluytens, 1995) and illiterate adults (see Kolinsky, 1998; Kolinsky et al., 1995), suggesting the prelexical locus of this effect.

Using the fragment-priming paradigm, Tabossi, Collina, Mazzetti, and Zopello (2000) showed that in Italian, words matching the syllable structure of the fragments (e.g., si#lenzio – silence - and si#l) were more strongly activated than words that mismatch the fragment (e.g., sil#vestre – sylvan – and sil#l), which is compatible with the “syllabic effect”. Nevertheless, small durational differences between the vowels in these fragments appear to have signaled the difference in syllabic structure. This could suggest that the “syllabic effect” is instead a product of subsegmental differences in the input (see also Content et al., 2001). In other words, Tabossi et al.’s (2000) syllabic effect could be due to sublexical segmentation procedures and not necessarily related with prelexical classification units. Indeed, the
importance of syllables in speech perception has often been related to the fact that, in
natural speech, coarticulation is lower between than within syllables (Liberman,
Cooper, Shankweiler, & Studdert-Kennedy, 1967). Furthermore, as demonstrated in
the work presented in this thesis, coarticulation is a reliable cue in speech
segmentation in adulthood (Experiment 1 and 4 of the present study).

Therefore, sublexical segmentation cues such as the degree of coarticulation
between segments could, by themselves, account for the observed syllabic effect.
Indeed, using the migration paradigm, Kolinsky et al. (1995) also demonstrated the
syllabic effect in French. At first sight these results could suggest that the
segmentation devices were the only procedures required for the observation of
syllabic effects without calling upon any prelexical classificatory unit. However, if
that was the case, we would always observe a syllabic effect mediated by sublexical
segmentation devices, in every language. Note that the results observed in
Portuguese with the migration paradigm (Kolinsky, 1998; Kolinsky et al., 1995)
refute this argument. Based on the results found with coarticulatory cues in the study
presented in this thesis (see Experiments 1 and 4, in chapters two and four,
respectively), we already know that in EP coarticulation is a reliable segmentation
cue. This means that the migration of a subsyllabic unit in EP (Castro et al., 1995;
Kolinsky et al., 1995; Kolinsky, 1998) cannot be simply due to weaker coarticulatory
influences in EP than in languages in which syllables are the unit that migrates the
most (e.g., in French, Kolinsky et al., 1995; in English, Mattys & Melhorn, 2005).
Thus, a variety of (both segmentation and classificatory) units could be involved in
prelexical level, with their weight depending on their relevance in any particular
language. An approach to the speech segmentation problem relying on several
sources of information extracted and integrated from infancy on, through all linguistic development, may provide a more powerful and realistic account of listeners’ accurateness in recognizing continuous speech (Christiansen & Curtin, 2005; Kolinsky et al., 2000; Mattys et al., 2005). Speech perception system could exploit different types of segmentation cues in order to constrain highly structured sublexical representations (Kolinsky et al., 1995).

7.3.2. Cross-species statistical learning

The acquisition and processing of language is, beyond doubt, governed by universal constraints, many of which are derived from innate properties of the human brain (Christiansen & Ellefson, 2002). In Chomsky’s (1980; 1986) approach, the constraints on the acquisition and processing of language are represented in the form of a Universal Grammar (UG), i.e., a large biological endowment of linguistic knowledge, highly abstract, comprising a complex set of linguistic rules and principles that could not simply be acquired from exposure to language during development. Such characteristics are thus considered innate rather than learned and species-specific.

Alternatively, the approach of Christiansen and Ellefson (2002) stresses the adaptation of linguistic structures to the biological substrate of the human brain rather than concentrating on biological changes to accommodate language. Therefore, many constraints on linguistic adaptation derive from non-linguistic limitations on learning and processing of hierarchically organized sequential structure. The underlying mechanisms existed prior to the appearance of language...
but, presumably, also underwent changes after the emergence of language. Consequently, many language universals may reflect non-linguistic, cognitive constraints on learning and processing of sequential structure, rather than an innate UG. These aspects would correspond in Hauser, Chomsky, and Fitch’s (2002) terms to the *Faculty of Language – broad sense (FLB)*. Although many aspects of FLB are shared with other species, the core aspects of *Faculty of Language – narrow sense (FLN)*, such as the recursive aspect (and the notion of discrete infinity; Fitch, Hauser, & Chomsky, 2005), appear to lack any analog in non-human species and are thus unique in humans (*for a discussion see* Fitch et al., 2005; Jackendoff & Pinker, 2005; Pinker & Jackendoff, 2005). In this regard, Hauser et al. (2002) argue for a qualitative difference on the types of mechanism underlying language processing.

Statistical learning based on TPs computation would be one example of a *phylogenetic*-based mechanism, or as in Hauser et al. (2002) terms, a FLB one. Indeed, Yang (2004) already suggested that general-domain statistical learning can be constrained by innate, domain-specific principles of linguistic structures, where UG could “instruct” the learner what cues (or what type of regularities) to attend. Note, however, that what is not so far clearly known are the sorts of statistical informations human listeners are able to use in natural linguistic settings. As Seidenberg, MacDonald, and Saffran (2002) pointed out:

> “Our understanding of the contribution of statistical learning is limited by incomplete knowledge of the kinds of statistics infants encode and whether these are the ones relevant to natural language. This view also leaves open the critical question of why only humans acquire language, as many other species are capable of simple forms of statistical learning.” (p. 554).
Therefore, although statistical learning is also observed in non-human primates (Hauser et al., 2001) and even in rats (Toro & Trobalón, 2005), the phylogenetic root of this mechanism does not imply that non-human species are able to extract the same regularities from a speech stream that human listeners do (e.g., Trout, 2001). Indeed, similar cross-species behaviors do not necessarily imply the same underlying computational abilities and units of analysis (Weiss & Newport, 2006).

**Cross-species differences in statistical computations**

Rats can segment an AL speech stream (Toro & Trobalón, 2005). However, they do so by using the overall frequency of co-occurrence among AL-items, a different kind of computation than that the one used by humans (who can rely on TPs: Aslin et al., 1998; Safran et al., 1996a).

On the contrary, cotton-top tamarins (like humans) are sensitive to TPs between adjacent syllables (Hauser et al., 2001) and also between non-adjacent ones (Newport, Hauser, Spaepen, & Aslin, 2004). These non-human primates are also sensitive to TPs between nonadjacent vowels but not between non-adjacent consonants (Newport et al., 2004). And here is where the distinction between statistical learning in humans and in other species probably lays.

In opposition to what is observed in other species, human listeners are able to extract statistical regularities both between nonadjacent consonants and between nonadjacent vowels (Newport & Aslin, 2004). Bonatti, Peña, Nespor, & Mehler (2005) have demonstrated that human listeners find it easier to track TPs between nonadjacent consonant than between nonadjacent vowels, in opposite to what was
found with tamarins (Newport et al., 2004). Therefore, consonants seem to be more suitable than vowels to parse speech streams using statistical dependencies (see Bonatti, Peña, Nespor, & Mehler, 2007; Mehler, Peña, Nespor, & Bonatti, 2006). This could suggest a particular and species-specific role of consonants in language processing.

**Vowels and Consonants in human statistical learning**

An emergent bulk of research has suggested vowel-consonant independent (structural and functional) treatment (e.g., Boatman, Hall, Goldstein, Lesser & Gordon, 1997; Caramazza, Chialant, Capasso, & Miceli, 2000; Cutler, Sebastián-Gallés, Soler-Vilagelu, & van Ooijen, 2000; Nespor, Peña, & Mehler, 2003), since an early phase in linguistic development (Nazzi, 2005; Nazzi & New, 2007). In particular, Nespor et al. (2003) noted that vowels are the main carriers of prosodic information, providing information at the prosodic and syntactic levels and defining some basic linguistic distinctions at rhythmic properties of languages (e.g., Ramus, Nespor, & Mehler, 1999). Consonants have a privileged role at lexical level, outnumbering the number of vowels in many languages, and performing a primary role in lexical items’ distinction.

In this particular demonstration of the different roles of vowels and consonants in speech processing, the ALL has revealed itself an important tool. The compatibility between particular types of statistical learning and natural languages can suggest that the structure of natural languages may be formed, at least in part, by the constraints and selectivity of what human learners find easy to acquire.

In sum, although humans share a basic statistical learning mechanism with
other species (particularly other primates), it differs in what regards the more complex types of computations humans are able to perform.

The study of the representations underlying statistical computations in humans can provide clues for the role of different linguistic units (e.g., consonant, vowels, syllables) in speech processing. ALL is therefore an additional, complementary paradigm for exploring and testing hypotheses about language evolution, language acquisition and sublexical segmentation procedures in adulthood. It also enables the study of speech processing of human listeners in different stages of linguistic development, as well as cross-species comparative studies (e.g., Hauser et al., 2001; Newport et al., 2004).

### 7.4. The ALL paradigm revisited

The AL speech stream used in ALL experiments is clearly an extreme case: all words are usually of the same length, with no short word embeddings in AL-words, and with a reduced artificial “lexicon”. Thus AL obviously deviates from natural languages. However, this does not necessarily diminish its impact or usefulness to the study of language processing.

An additional fact confirming the ALL importance is the fact that the ALL paradigm provides a highly controlled situation in which the available sources of information can be systematically manipulated. However, the generally used two-alternative forced choice task, which is an indirect (off-line) measure of ALL and hence of speech segmentation, is the Achilles’ heel of this paradigm. In the present study (experiments 7, 8 and 9; see chapter 6), we have demonstrated that the ALL
paradigm can be combined with conventional experimental techniques, to gather more fine-grained (and on-line) information about speech segmentation processing. Note, however, that the present study is just a first demonstration that ALL can be combined with other techniques and thus further work should extend the present results recurring to other experimental tasks.

As a matter of fact, it would be interesting to evaluate whether similar results to the lexicalization effects found in Experiment 8 would also be observed using pause detection tasks (Mattys & Clark, 2002). Contrary to the lexical decision task (used in the present study), pause detection seems to be less prone to strategic biases. Pause detection offers a measure of the overall level of lexical activity in the absence of any explicit linguistic judgment, since listeners are merely asked to detect periods of silence in speech.

It would also be interesting to use the visual word paradigm (Tanenhaus & Spivey-Knowlton, 1996), to evaluate whether the discrepancy in lexicalization effects found in congruent and in incongruent cues condition (Experiments 8 and 9 of the present study) would be replicated. This would allow an evaluation of the lexicalization process in a more fine-grained manner (see section 1.2.3. of chapter one). Since this eye-tracking task is now beginning to be used with visual words instead of pictures of referents (e.g., Huetting & McQueen, 2007; McQueen & Viebahn, 2007), it could be a sensitive tool for the study of word learning in the absence of semantic contexts in adulthood.

In future work with the ALL paradigm it would also be important to analyze directly the segmentation procedures adopted by listeners during the AL-familiarization phase. This could possibly be achieved by adopting a variant of ALL
similar to the one designed by Shukla et al. (2007), combined with on-line measures such as the ones used in Serial RT tasks (see Cleeremans et al., 1998). This adaptation could be used to evaluate the role of different available sublexical cues during an AL familiarization period, considering the time course. For example, it could be evaluated whether consistent AL speech segmentation (i.e., listeners’ systematic detection of AL-word boundaries) driven by a particular cue could occur faster than when it is driven by another cue. It could also be used to evaluate how different cues are integrated in speech processing, during the time course of AL exposition.

### 7.5. Thoughts for future research

How is lexical activation modulated during speech perception processing within a continuous flow of speech? There are two inter-related aspects to this question (cf. McQueen, Dahan, & Cutler, 2003). The first concerns the parameters which determine whether a given word should enter or leave the competitor set. The second concerns the metric which is used to compute the goodness of fit of any given word candidate to the speech input.

The present study considered only sublexical sources of information and their role in speech segmentation in different listening conditions. A next step in this research would be to evaluate the interaction between sublexical and high-level (lexical and post-lexical) sources of information in lexical activation.

In fact, Mattys & Melhorn (2007) already demonstrated a graded interaction between top-down and bottom-up information in speech segmentation. In a forced
choice task (i.e., deciding which one of two words; e.g., pie vs. eye, was heard), on near-homophonic phrases (e.g., “scum pie” vs. “scump eye”), listeners adopted compensatory segmentation strategies whereby reliance on high-level (sentential/lexical) information was inversely related to the strength (strong vs. mild acoustic junctures) of the acoustic cues. Although lexical effects (e.g., a bias in favor of “scum pie” instead of “scump eye”) were accompanied with a reduction in phonetic acuity, sentential effects were not. Thus, only lexical information affected listeners’ sensitivity to acoustic cues.

Therefore, whereas segmental match is critical for lexical access, the impact of mismatching is not restricted to segmental information, since subsegmental (e.g., Davis et al., 2002; Gow & Gordon, 1995; Salverda et al., 2003; Spinelli, Welby, & Schaegis, 2007) and suprasegmental information (e.g., Salverda et al., 2007; Soto-Faraco et al., 2001) also seem to have a role in lexical activation. Indeed, we could manipulate the congruency of available sublexical cues in a continuous speech stream compatible with different lexical parsing (e.g., /selæfiʃ/ for the target /fiʃ/ and /æfiʃ/, Spinelli et al., 2007). It could also be evaluated whether the role of sublexical cues on lexical activation is modulated by listening conditions: Does their weighting in speech segmentation (cf. Mattys et al., 2005) extend to the modulation of the activation of different lexical candidates? Would cognitive noise have a particular role in the weighting of different speech-specific cues?

The present study showed that cognitive noise such as a perceptual attentional load can seriously compromise the computation of statistical information (i.e., TPs between adjacent syllables). However, in conditions in which only speech-specific information is available it could be useful to instigate different types of cognitive
noise, namely a perceptual (passive and peripheral) load and a working memory (active and central) load. In these conditions, it is possible that the weight changing observed in these two cognitive noise conditions could differ, in particular when both sublexical and lexical information are available. Further work should investigate the differential impact of different types of cognitive noise in speech-specific cues, as well as whether and how physical and cognitive noises interact with each other.

7.6. In sum

In the study reported in this thesis we evaluated the role of different sublexical sources of information in speech segmentation and in the lexicalization process in the context of a full mature speech perception system. We have also provided evidence regarding the usefulness of the ALL paradigm, particularly when combined with conventional experimental designs.

According to the present results and in line with an emergent bulk of research, spoken word recognition models have two important challenges to equate. On the one hand, models need to be able to accommodate the role of sublexical informations in speech segmentation and their interaction with high-level information (Mattys & Melhorn, 2007; Mattys, Melhorn, & White, 2007). On the other hand, the present data show that these models also need to incorporate the lexicalization processes that occur in the context of a continuous speech stream, as well as the role that the congruency between the available segmentation cues plays on both speech segmentation and word learning, and thus not only in on-line processing but also in long-term structural changes.
At last, I will end this thesis adapting (and not adopting) the enunciation of the *Law of Conservation of Mass* of Antoine-Laurent Lavoisier (1743-1794) to speech segmentation domain. Therefore, in speech segmentation:

« rien ne se perd, rien ne se crée, tout se transforme ». 
REFERENCES


Fernandes, T., Ventura, P., & Kolinky, R. (*in preparation*). What is the impact of attention on Statistical Learning? Amount of Familiarization and Intention to Learn in Artificial Language Learning.

Fernandes, T., Kolinky, R. & Ventura, P. (*in preparation, a*). Cognitive Noise is also Noise: The Impact of Attentional Resources on the use of Statistical Information and Coarticulation as Speech Segmentation Cues.


Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The Faculty of Language: What is it, who was it, and how did it evolve? *Science, 298*, 1569-1579.


References


References


References


APPENDICES
9.1. Appendix I

TP-words and part-words of the AL used in Experiments 1 - 7 of the present study, represented in their phonological form according to IPA. The “#” defines a TP-boundary within part-words.

<table>
<thead>
<tr>
<th>AL stimulus type</th>
<th>TP-words</th>
<th>Part-words</th>
</tr>
</thead>
<tbody>
<tr>
<td>/bəbuku/</td>
<td>/bə#kile/</td>
<td></td>
</tr>
<tr>
<td>/bukələ/</td>
<td>/bi#lufe/</td>
<td></td>
</tr>
<tr>
<td>/luʃəbə/</td>
<td>/ ku#buke/</td>
<td></td>
</tr>
<tr>
<td>/keʃubi/</td>
<td>/ke#lefu/</td>
<td></td>
</tr>
<tr>
<td>/fuʃibu/</td>
<td>/fibu#lu/</td>
<td></td>
</tr>
<tr>
<td>/kilebu/</td>
<td>/lebu#fu/</td>
<td></td>
</tr>
</tbody>
</table>
9.2. Appendix II

The pictures of Snodgrass and Vanderwart (1980) by semantic category (i.e., animals; fruit and vegetables; transport vehicles) used in Experiments 4 – 6 in the cognitive noise (attention-load) conditions.

*Animals:*

*Fruits & vegetables:*

*Transport vehicles:*
9.3. Appendix III

The text and the four questions presented to participants in the low-attention load (weak cognitive noise) conditions in incidental ALL – Experiments 4 - 6 of the present work.

A TERMODINÂMICA

A Termodinâmica é o estudo das transformações de energia de uma forma para outra e da interacção da energia com a matéria. Pode ser considerada como uma ciência deductiva baseada em duas leis principais, leis essas que têm a forma de negação, estabelecendo restrições do comportamento da energia. Tais leis correspondem ao "Primeiro Princípio da Termodinâmica" e ao "Segundo Princípio da Termodinâmica".

O Primeiro Princípio estabelece que: "A energia não pode ser criada ou destruída". Este enunciado -que é uma negação está contido numa lei maior, natural, o chamado Princípio da Conservação de Energia. O Segundo Princípio também tem carácter negativo: "O calor não pode fluir espontaneamente de um corpo frio para um corpo quente". O Segundo Princípio gerou sempre discussões de carácter histórico e filosófico, fora do contexto mecânico/termodinâmico que lhe deu origem. Ambos os princípios estão de acordo com a experiência humana, não se tendo encontrado, até o momento, qualquer facto ou evidência experimental que os refute. Inclusivamente, passaram a merecer a qualificação de "leis naturais".

No domínio da termodinâmica teórica, as duas leis são as premissas ou axiomas, que formam o alicerce desta ciência. A tarefa mais notória dos investigadores deste campo é a determinação do maior número de consequências logicamente necessárias
decorrentes das duas leis que foram assumidas. Esse parece ter sido o caminho percorrido pelo matemático (nascido na Grécia mas que viveu na Alemanha) Constantin Carathéodory no início do Século XX. Um outro caminho seria a construção do edifício da termodinâmica a partir de uma outra forma das duas leis acima mencionadas, acrescentando um conjunto de regras gramaticais e lógicas, de uma maneira semelhante à utilizada por Euclides no desenvolvimento da geometria. Obviamente, o valor prático deste tipo de trabalho seria dado, em última análise, pela extensão e alcance das suas consequências. Um empreendimento de reconstrução da termodinâmica deste tipo, foi tentado pelo engenheiro inglês Joseph Keenan nos anos 1940, com o objectivo de tornar mais claros os textos desta ciência.

Os princípios da Termodinâmica constituem restrições gerais que – muito provavelmente! - a Natureza impõe às transformações de energia. Esses princípios não podem ser deduzidos a partir de enunciados mais básicos, são portanto primitivos. Além do mais, requerem o emprego de conceitos que são eles próprios primitivos, como, por exemplo, energia. Neste contexto, o conceito de energia corresponde a uma abstracção matemática cuja existência nunca ultrapassa o âmbito da sua relação funcional com variáveis ou coordenadas que podem ser medidas e que dispõem de uma interpretação física mais concreta. Portanto, o conceito de energia é primitivo e o Primeiro Princípio da Termodinâmica, que expressa o facto da energia se conservar, é igualmente um enunciado primitivo. A origem deste princípio é a Mecânica, onde a sua aplicação a corpos rígidos em movimento, na ausência de atrito, fornece uma associação entre formas externas de energia e trabalho mecânico. A sua verificação experimental é relativamente simples e relaciona-se com a Segunda Lei do movimento de Newton.

A ideia central contida na transição de um princípio de conservação limitado à Mecânica para uma lei global da Termodinâmica tem origem no reconhecimento do
facto do calor ser uma forma de energia e do facto da energia interna ser uma propriedade intrínseca da matéria. Desta forma, o Primeiro Axioma que corresponde ao Postulado da Energia Interna pode ser formulado da seguinte forma: "Existe uma forma de energia (conhecida como energia interna) que é uma propriedade intrínseca de um sistema e que está relacionada funcionalmente com as coordenadas mensuráveis que o caracterizam. Este postulado afirma a existência de uma função, chamada energia interna, e fornece uma relação que a associa a grandezas mensuráveis. Contudo, a equação que expressa este enunciado não define implicitamente o que é a energia interna, uma vez que não existe tal definição, mas fornece um meio de calcular as variações de função e não o valor absoluto.

**Questão 1:** Defina o que é a Termodinâmica.

**Questão 2:** Quais foram os dois caminhos percorridos pelos investigadores desta área, tendo como ponto de partida os dois princípios da Termodinâmica?

**Questão 3:** O que entende por enunciado e conceito primitivos?

**Questão 4:** Descreva as relações entre a Mecânica e a Termodinâmica.
9.4. Appendix IV

Carrier sequences used in the word-spotting task in Experiment 7. The sequences are given in their phonological form, according to IPA. The “#” marks the statistical boundary between AL stimuli and the residual syllable in the string, both for initial carriers (in which the AL stimulus was embedded at the onset of the four-syllabic carrying string) and for final carriers (in which the AL stimulus was embedded at the end of the four-syllabic carrying string). The TP-words and part-words of the AL are underlined. In part-words, the dot marks TP boundaries.

<table>
<thead>
<tr>
<th>AL stimulus type</th>
<th>TP-words</th>
<th>Part-words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial carriers</td>
<td>Final carriers</td>
</tr>
<tr>
<td></td>
<td>/bēbuku#drō/</td>
<td>/drō#bēbuku/</td>
</tr>
<tr>
<td></td>
<td>/bukel#tre/</td>
<td>/tre#bukel/</td>
</tr>
<tr>
<td></td>
<td>/lufebe#grō/</td>
<td>/grō#lufebe/</td>
</tr>
<tr>
<td></td>
<td>/kēfubi#nēs/</td>
<td>/nēs#kēfubi/</td>
</tr>
<tr>
<td></td>
<td>/fūfibu#rōs/</td>
<td>/rōs#fūfibu/</td>
</tr>
<tr>
<td></td>
<td>/kīlebu#get/</td>
<td>/ger#kīlebu/</td>
</tr>
</tbody>
</table>
9.5. Appendix V

Lists A and B of the base-words used for creating the ALs of Experiments 8 and 9, and used as critical trials in the lexicalization test (i.e., the lexical decision task). Phonological version of the stimuli according to IPA is presented and the English translation is in parentheses.

<table>
<thead>
<tr>
<th>Base-Words</th>
<th>List A</th>
<th>List B</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIVELA (buckle)</td>
<td>/fi'velə/</td>
<td>FINITO (finite)</td>
</tr>
<tr>
<td>LECTIVO (school year)</td>
<td>/le'tivu/</td>
<td>LATADA (trellis)</td>
</tr>
<tr>
<td>PIRRAÇA (teasing)</td>
<td>/pi'rasə/</td>
<td>PIJAMA (pajamas)</td>
</tr>
<tr>
<td>ROLETA (roulette)</td>
<td>/ru'letə/</td>
<td>RAVINA (ravine)</td>
</tr>
<tr>
<td>SACOLA (knapsack)</td>
<td>/se'kɔlə/</td>
<td>SAMARRA (sheepskin coat)</td>
</tr>
<tr>
<td>TACADA (cupful)</td>
<td>/te'kada/</td>
<td>TISANA (tisane)</td>
</tr>
<tr>
<td>SONATA (sonata)</td>
<td>/su'nata/</td>
<td>SOSSEGO (calm)</td>
</tr>
<tr>
<td>FONEMA (phoneme)</td>
<td>/fu'nema/</td>
<td>FARRAPO (rag)</td>
</tr>
<tr>
<td>GASOSO (gaseous)</td>
<td>/ge'sozu/</td>
<td>GAVETA (drawer)</td>
</tr>
<tr>
<td>CHICHARRO (horse-mackerel)</td>
<td>/ʧiʃarə/</td>
<td>CHALAÇA (jest)</td>
</tr>
</tbody>
</table>
9.6. Appendix VI

AL stimuli presented in phonological form according to IPA, by AL (AL\textsubscript{A2} and AL\textsubscript{B2}, according to the two lists of base-words) and stimulus type (TP-word and part-word) used in Experiment 8 and in Experiment 9. All TP-words of Experiment 8 (that constituted the part-words of Experiment 9) differed from the base-words presented in Appendix B by their last syllable.

<table>
<thead>
<tr>
<th>Artificial Languages (AL)</th>
<th>AL\textsubscript{A} (AL\textsubscript{A2})</th>
<th>AL\textsubscript{B} (AL\textsubscript{B2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 2</td>
<td>TP-words</td>
<td>Part-words</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>Part-words</td>
<td>TP-words</td>
</tr>
<tr>
<td>/fi\v\nu/</td>
<td>/kufone/</td>
<td>/fim\n/</td>
</tr>
<tr>
<td>/li\v\nu/</td>
<td>/zetaka/</td>
<td>/letavu/</td>
</tr>
<tr>
<td>/pi\v\unu/</td>
<td>/durule/</td>
<td>/pi\v\unu/</td>
</tr>
<tr>
<td>/rulenu/</td>
<td>/nus\v\nu/</td>
<td>/reviku/</td>
</tr>
<tr>
<td>/sek\v\nu/</td>
<td>/mugezo/</td>
<td>/semadu/</td>
</tr>
<tr>
<td>/teka\v\nu/</td>
<td>/zus\v\nu/</td>
<td>/tiz\v\nu/</td>
</tr>
<tr>
<td>/sunabi/</td>
<td>/bip\v\nu/</td>
<td>/suset\v/</td>
</tr>
<tr>
<td>/funetu/</td>
<td>/tule\v/</td>
<td>/fu\v\unu/</td>
</tr>
<tr>
<td>/ge\v\vu/</td>
<td>/me\v\vu/</td>
<td>/ge\v\vu/</td>
</tr>
<tr>
<td>/ji\v\vu/</td>
<td>/de\v\vu/</td>
<td>/jelatu/</td>
</tr>
</tbody>
</table>