the Phonetician

A Publication of ISPhS International Society of Phonetic Sciences

NUMBER 107-108 2013/I-II

ISPhS
International Society of Phonetic Sciences

President: Ruth Huntley Bahr

Secretary General:
Mária Gósy

Honorary President:
Harry Hollien

Vice Presidents:
Angelika Braun
Marie Dohalská-Zichová
Mária Gósy
Damir Horga
Heinrich Kelz
Stephen Lambacher
Asher Laufer
Judith Rosenhouse

Past Presidents:
Jens-Peter Köster
Harry Hollien
William A. Sakow †
Martin Kloster-Jensen†
Milan Romportl †
Bertil Malmberg †
Eberhard Zwirner †
Daniel Jones †

Honorary Vice Presidents:

A. Abramson	P. Janota †	H. Morioka	M. Shirt
S. Agrawal	W. Jassem	R. Nasr	E. Stock
L. Bondarko	M. Kohno	T. Nikolayeva	M. Tatham
E. Emerit	E.-M. Krech	R. K. Potapova	F. Weingartner
G. Fant †	A. Marchal	M. Rossi	R. Weiss

Auditor:
Angelika Braun

Treasurer:
Ruth Huntley Bahr

Affiliated Members (Associations):

American Association of Phonetic Sciences	J. Hoit & W.S. Brown
Dutch Society of Phonetics	B. Schouten
International Association for Forensic Phonetics and Acoustics	A. Braun
Phonetic Society of Japan	I. Oshima & K. Maekawa
Polish Phonetics Association	G. Demenko

Affiliated Members (Institutes and Companies):

KayPENTAX, Lincoln Park, NJ, USA	J. Crump
Inst. for Advanced Study of the Communication Processes, University of Florida, USA	H. Hollien
Dept. of Phonetics, University of Trier, Germany	J.-P. Köster
Dept. of Phonetics, University of Helsinki, Finland	A. Iivonen
Dept. of Phonetics, University of Zürich, Switzerland	S. Schmid
Centre of Poetics and Phonetics, University of Geneva, Switzerland	S. Vater

International Society of Phonetic Sciences (ISPhS) Addresses

www.isphs.org

President:

Professor Dr. Ruth Huntley Bahr
President's Office:
Dept. of Communication Sciences and Disorders
University of South Florida

4202 E. Fowler Ave., PCD 1017
Tampa, FL 33620-8200
USA
Tel.: ++1-813-974-3182
Fax: ++1-813-974-0822
e-mail: rbahr@usf.edu

Secretary General:

Prof. Dr. Mária Gósy
Secretary General's Office:
Kempelen Farkas Speech Research Laboratory,
Research Institute for Linguistics,
Hungarian Academy of Sciences
Benczúr u. 33
H-1068 Budapest
Hungary
++36 (1) 321-4830 ext. 172
++36 (1) 322-9297
e-mail: gosy.maria@nytud.mta.hu

Guest Editors:
Dr. Robert Mannell
Guest Editor's Office:
Department of Linguistics
Australian Hearing Hub
Level 3 North, Room 522
16 University Avenue
Macquarie University, NSW 2109
Australia
Tel.: +61-2-9850 8771
e-mail: robert.mannell@mq.edu.au

and

Prof. Dr. Mária Gósy
Guest Editor's Office:
Kempelen Farkas Speech Research Laboratory,
Research Institute for Linguistics,
Hungarian Academy of Sciences
Benczúr u. 33
H-1068 Budapest
Hungary
Tel.: ++36 (1) 321-4830 ext. 171
Fax: ++36 (1) 322-9297
e-mail: gosy.maria@nytud.mta.hu

Review Editor:
Prof. Dr. Judith Rosenhouse
Review Editor's Office:
Swantech
89 Hagalil St
Haifa 32684
Israel
Tel.: ++972-4-8235546
Fax: ++972-4-8327399
e-mail: swantech@013.net.il

Cover by András Beke

Technical Editor:
Dr. Tekla Etelka Gráczi
Technical Editor's Office:
Kempelen Farkas Speech Research Laboratory,
Research Institute for Linguistics,
Hungarian Academy of Sciences
Benczúr u. 33
H-1068 Budapest
Hungary
Tel.: ++36 (1) 321-4830 ext. 171
Fax: ++36 (1) 322-9297
e-mail: graczi.tekla.etelka@nytud.mta.hu

INTRODUCING THE GUEST EDITORS

Robert Mannell has been at Macquarie University since 1982. Before becoming a linguist and phonetician he had been a hospital biochemist, drug detection chemist and computer programmer. He joined Macquarie University's Linguistics Department as a part time PhD student whilst working full time in the same department as a computer programmer for the Macquarie Dictionary and as a researcher and programmer for a series of projects on speech synthesis and phonetics with Professor John Clark. His main research interests over the years have particularly focussed on speech synthesis, speech perception, psychoacoustics and other aspects of phonetics. In the 1990's he established, with Jonathan Harrington, Macquarie University's very successful Bachelor of Speech and Hearing Sciences (now the Bachelor of Speech, Hearing and Language Sciences) which leads, via additional postgraduate study, to careers in Audiology, Speech and Language Pathology and Linguistics. For about 10 years now this degree has also been equally co-convened by Felicity Cox, and the degree currently has an enrolment in excess of 300 students. He has supervised numerous PhD students and has had 5 successful PhD completions in the past two years, including Yoshito Hirozane who has a paper in this edition of the Phonetician, based on part of his PhD work. Current research funding includes the Hearing Cooperative Research Centre and the ARC Centre of Excellence in Cognition and its Disorders. In this funded research he is particularly focusing on a model of the auditory processing of speech and in particular is examining the perception and cognition of resonances and antiresonances in voiced speech sounds.

Mária Gósy graduated from Eötvös Loránd University, Budapest, with an MA in Hungarian and Russian linguistics. She received her PhD in phonetics in 1986 and her DSc (Doctor of the Hungarian Academy of Sciences) in psycholinguistics in 1993. She received a postdoctoral fellowship at MIT in 1987, and was a visiting professor at the Institute of Perception Research in Eindhoven in 1991. She was teaching at Boston University in 1987 and at University of Vienna (1991–1993). She has supervised 22 PhD students, including a foreign student, who successfully defended their dissertations so far. Currently she is a professor at Eötvös Loránd University and head of the Phonetics Department both of the University and the Research Institute for Linguistics of the Hungarian Academy of Sciences. She was the representative of Hungary in the

European Committee of the International Reading Association (1992–1997), and a board member of ESCA (later ISCA) between 1997 and 2000. She has been the Secretary General and Vice-President of the ISPhS from 2003. She was a member of the Presidency of the Hungarian Academy of Sciences (2004–2007). In 2011 she was elected as a board member of the Council of the IPA. She has received eleven national awards including the Order of Merit of the Hungarian Republic, Officer's Cross in 2012. She has participated in ten research projects including two international ones, and has been the leader of five of them. Her research areas are phonetics, psycholinguistics and applied research in speech sciences. She has been asked to give plenary talks at three international conferences. Her current work focuses on the speech production process and on the phonetic aspects of spontaneous speech. She has published 8 books, more than 160 papers (in Hungarian and English); is chief editor of the Beszédkutatás (Speech Research) Hungarian journal, a board member of several international journals, as well as chief organizer of four international congresses. She was appointed a member of the Hungarian Academy of Sciences in 2013.

the Phonetician

A **Peer-reviewed** Journal of ISPhS/**International Society of Phonetic Sciences**

ISSN 0741-6164

Number 107-108 / 2013-I-II

CONTENTS

ON THE SIMILARITY OF TONES OF THE ORGAN STOP *VOX HUMANA* TO HUMAN VOWELS[1]

Fabian Brackhane[2] and Jürgen Trouvain[3]
[2]Institut für Deutsche Sprache (IDS), Mannheim, Germany, [3]Computational Linguistics and Phonetics, Saarland University, Saarbrücken, Germany
e-mail: [2]brackhane@ids-mannheim.de, [3]trouvain@coli.uni-saarland.de

Abstract

In mechanical speech synthesis from the 18th up to the 20th century, reed pipes were mainly used for the generation of the voice and the organ stop vox humana was central in this process. This has been described in different historical documents which report that the vox humana in some organs sounded like human vowels. In this study, tones of four different voces humanae were recorded to investigate their similarity to human vowels. The acoustical and perceptual analysis revealed that some, though not all, tones show a high similarity to selected vowels.

1 Introduction

Many authors of the 18th and 19th century consider the organ stop *vox humana* as the prototype for a mechanical speech synthesiser or, more specifically, as the prototype for a vowel synthesiser. In this view, the task would be to develop the vowel-like features of the *vox humana* to a "speech organ" as Euler (1773: 246) suggested.

However, evidence for a real similarity to vowels is either missing or does not hold up under today's standards. Based on personal experience, the resemblance of the sound of modern and historical *voces humanae* and human vowels does not seem to be very close. For this reason, we performed a study including an acoustic analysis, as well as perception tests, to verify the historical descriptions of the *vox humana* and its similarity to human vowels.

2 The mechanism and use of the organ stop vox humana

The organ stop *vox humana*, consisting of reed pipes, has been described since the middle of the 16th century (Eberlein, 2007: 817). An organ stop is a set of organ pipes with different pitches but constructed in the same way. It can be switched "on", i.e., admitting the pressurised air to the pipes of this stop, or "off", i.e., stopping the air. Organs usually have multiple stops (often between 25 and 30 and not all of their pipes are visible from the outside). The majority of the stops are flue

[1] A shorter version of this article was published under the title “The organ stop 'vox humana' as a model for a vowel synthesizer” in the proceedings of the 14th Interspeech (Lyon) 2013, pp. 3172-3176.

pipes (see Fig. 1 bottom), although reed pipes are also common (see Fig. 1 top). A characteristic feature of the reed pipes used in a *vox humana* is the *resonator* that is of a relatively constant size independent of the pitch of the pipe. This means that there are possibly slight differences with respect to the size of the *resonators* because each pipe of a given *vox humana* stop is hand-made. For almost every other organ stop consisting of reed pipes, the length of the *resonator* decreases successively with the increasing pitch of the pipes. In case of the *vox humana*, the *resonators* act as a filter in such a way that formants can be observed that are similar to those found in human vowels (Lottermoser, 1936: 48; Lottermoser, 1983: 135).

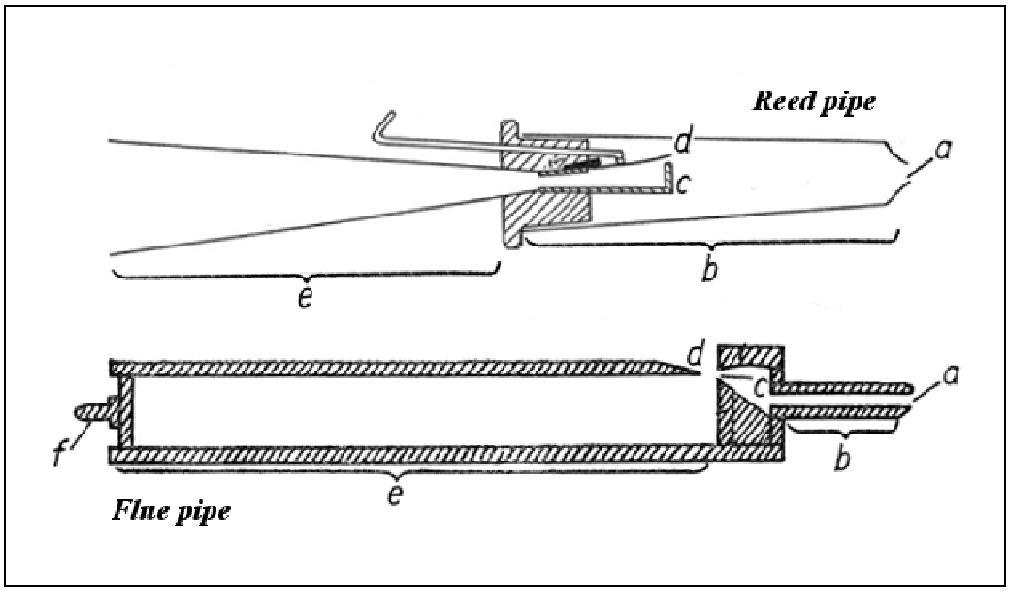

Figure 1. Schematic drawing of a reed pipe (top, re-drawn after Lottermoser 1936: 15) and a flue pipe of the type stopped diapason (bottom, redrawn after Adelung 1982: 43). The air flows into the pipes (a) passing the socket or boot (b). The air in the reed pipe (top) will be excited by the reed tongue (d) that lies on the shallot (c). The excitation of the air in the flue pipe (bottom) is possible by an increased air pressure at the windway (c) and the continuation towards the upper lip (d). The resonator (top e) and the body (bottom e) act as acoustic filters. (f) represents the cap needed for stopped flue pipes.

The term *vox humana* originates from the use of an organ reed stop with proportionally short resonators which substitute for the human singing voice. For this reason, it was never used solo but was usually played together with the so-called *tremulant* and the stopped flue stop *bourdon* (also called *stopped diapason*) of the same pitch. The *tremulant* changes the pressure of the air streaming to the pipes in brief intervals. The resulting sound, which resembles the vibrato of a human singing voice, has been named *vox humana*. Thus, the organ stop *vox humana* has been used

as a substitute for the human singing voice, however, it was not considered to be an *imitation* of the human voice.

However, knowledge about the original meaning of the term *vox humana* has been lost over time and was considered as an *imitation* rather than as a *subsititution*. For these historical reasons, there is not only one construction type but various ones. Nearly every organ builder of the 18th century intended to invent a really natural sounding *vox humana*. Thus, the name *vox humana* can be considered as a programmatic title rather than as a technical term. Numerous historical documents attested that these pipes clearly sounded like vowels (e.g. Greß, 2007: 27).

This new usage made organ builders (e.g. Joseph Gabler), as well as researchers such as Leonhard Euler (1707-1783) and Christian Gottlieb Kratzenstein (1723-1795), to consider the *vox humana* as the prototype of speech synthesis.

3 Recordings and acoustic analysis of various *voces humanae*

3.1 Data

It was our aim to test the historical statements concerning the similarity of the *vox humana* sound to those of human vowels. This required recordings of those organs where the stops are historically authentic (and not re-constructed). The research question was whether pipes of a *vox humana* really displayed formant structures similar to those of human vowels. More specifically, we were interested in determining whether certain vowel qualities could be recognised reliably by human listeners.

The first author recorded selected tones from the originally preserved *vox humana* stops of four different organs from the middle of the 18th century. Three of these were located in churches in the southwest Germany: Abteikirche Amorbach, Schlosskirche Meisenheim and Stadtkirche Simmern (AMO, MEI, SIM henceforth). These organs were built between 1767 and 1782 by craftsmen from the same family of organ builders (Stumm), and all three organ stops had the same construction style and sizes (see Fig. 2).

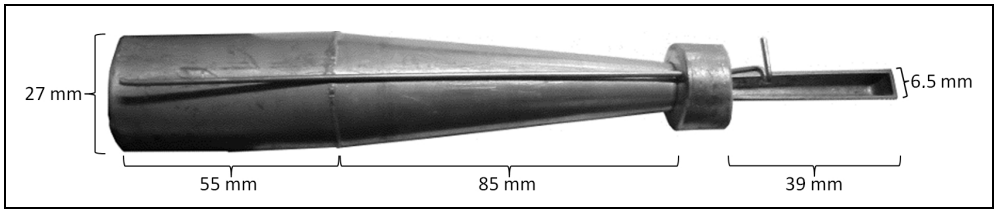

Figure 2. Reed pipe from the *vox humana* (Tone g0) from Amorbach (1782), without the boot (cp. (b) Fig. 1 top) and without the reed tongue (cp. (d) Fig. 1 top).

In addition, *the vox humana* of the organ of the Stadtkirche in Waltershausen (Thuringia, Eastern Germany) was recorded (WAL henceforth) at a later time. This organ stop is a copy of the *vox humana* from the great organ at the monastery in Weingarten (1750) which is famous because of its constructor, Joseph Gabler, who

attempted to build pipes with a sound that resembled human singing voices in a particular way. The *resonators* of these pipes were adapted to human larynges.

The tones C, G, c0, g0, c1, g1, c2, g2 and c3 (in an alternative notation C, G, c, g, c', g', c'', g'', c''') were recorded from the *voces humanae* of all four organs (9 tones * 4 organs = 36 recordings in total). In SIM, we also recorded the historical (i.e., not reconstructed) reed pipe stops *trumpet* and *crumhorn* (for C and g0). These two stops substantially differ from the *vox humana* in their construction styles and they were recorded for comparison with the *vox humana* of the same organ and the measurements found in Lottermoser (1983). Only two tones were selected: C as the lowest one and g0 because it has been described as particularly vowel-like (see e.g., Frotscher 1927: 54). Thus, the total number of recorded tones increased to 40.

The tones in AMO, MEI and SIM were played solo for the recordings, i.e., as pure tones and for this reason without the additional stops *stopped diapason* and *tremulant*, which are typically used in musical tradition. The *vox humana* in WAL could not be played solo for technical reasons; consequently the tones here were played in combination with the flue pipes of the *stopped diapason*, but without the *tremulant*.[2]

The microphone was placed at a distance of about half a metre above the resonators to produce comparable recordings in the acoustically different churches and to reduce the echo and filter effects of the rooms as much as possible (although the influence of the acoustic conditions of the churches can never be completely excluded). All recorded tones were about 5 seconds in duration. This length is due to the fact that the reed pipes need a relatively long time to reach the stationary phase.

The acoustic analysis of the data included the measurement of F_0 and the first three formants. For each 5-sec tone, the first and last 5% of the duration were ignored and from the remainder of the tone, 10 equidistant values were taken. The analysis was performed with the phonetic standard freeware Praat (version 5.3.19).

3.2 Results

The values for the fundamental frequency show that all four organs differed in their F_0 for virtually all tones (see Table 1). For example, the tone G, comparable to a bass voice, ranged from 98 Hz in AMO to 105 Hz in MEI. In the following sections, only the results for SIM and WAL are reported due to a high level of comparability of the stops in AMO, MEI and SIM.

The spectra of all *voces humanae* tones showed clear formant structures. This is also true for the additional stops; *trumpet* and *crumhorn* (see Fig. 3). However, the formant shapes of the *voces humanae* showed more similarity to the formants of typical human speech. Interestingly, in all four organs, the values for F_0 and F_1

[2] Unfortunately, the recordings of the three Stumm organs in AMO, SIM and MEI were already finished when we surprisingly had the opportunity to record the organ in WAL. Thus, the recordings from WAL are not fully comparable to the recordings of the other three *voces humanae.*

converged or even merged for the two and sometimes three highest tones, which made a visible distinction nearly impossible.

Table 1. F_0 values in Hz of all tones of all voces humanae.

Tone	AMO	MEI	SIM	WAL
C	66	70	69	69
G	98	105	102	103
c0	132	141	136	139
g0	197	210	205	208
c1	263	281	274	277
g1	395	411	408	415
c2	527	562	548	554
g2	790	844	818	832
c3	1054	1124	1093	1108

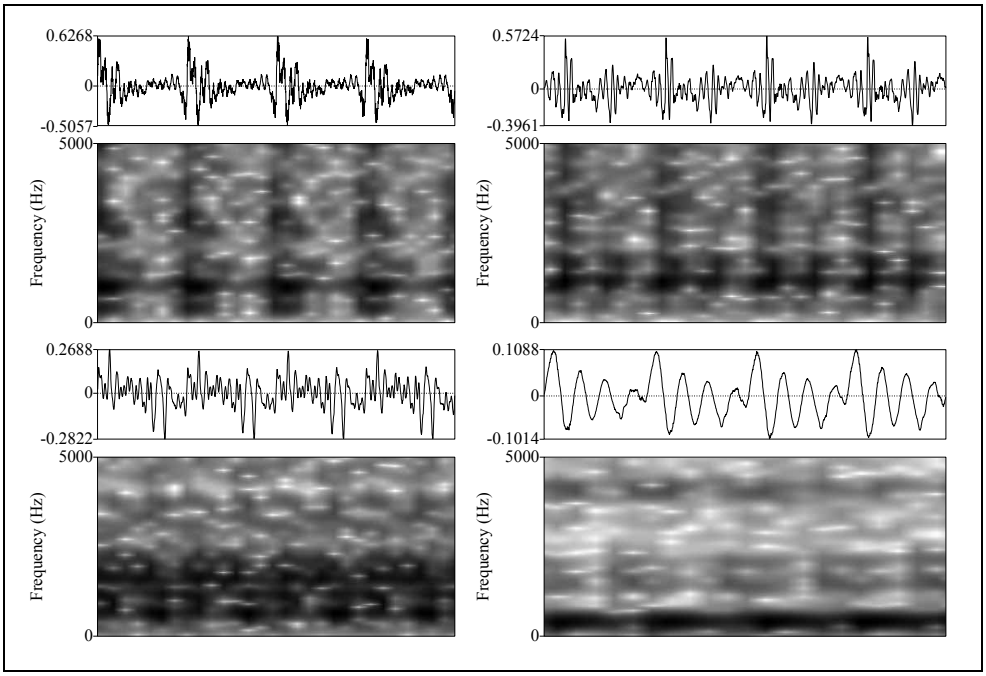

Figure 3. Waveforms and spectrograms of sections with four periods taken from the tone C of the stops in SIM: *vox humana* (top left), crumhorn (top right) and trumpet (down left) (duration: 60 ms) and from the vowel /ø/ of a male German speaker (down right; duration: 42 ms, F_0: 97 Hz).[3]

[3] Recordings of the tone G which would be more comparable to the human voice were not available for all stops.

In Figure 4 (a-c), the spectral distribution of the *vox humana* from SIM (from Fig. 3) and WAL are compared with the spectrum of the human vowel that is also shown in Figure 3. The decline of the spectral slope was much more intense for the human vowel than for the organ-generated tone.

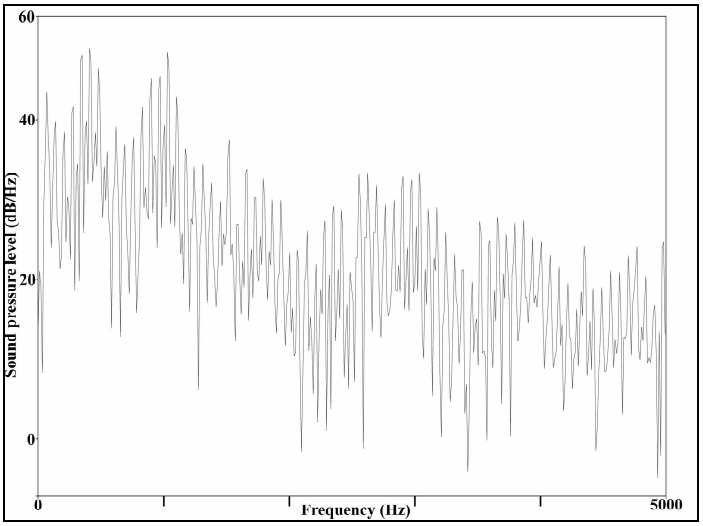

Figure 4a. Spectrum for the middle part of tone C of the stop vox humana from SIM from 0 to 5 kHz.

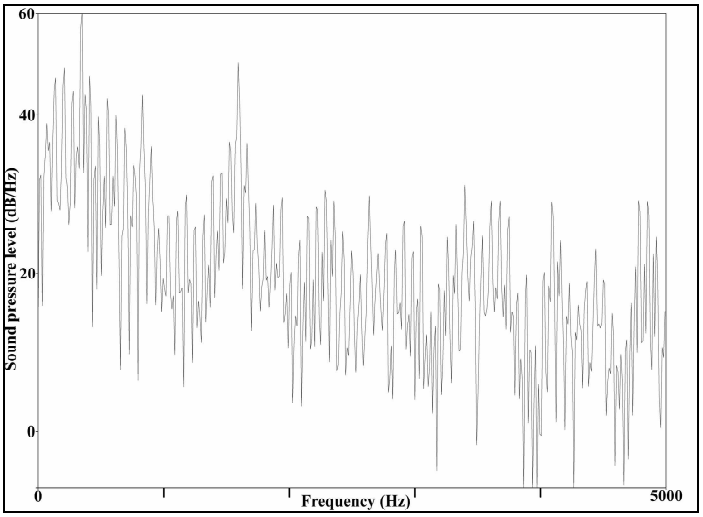

Figure 4b. Spectrum for the middle part of tone C of the stop vox humana from WAL from 0 to 5 kHz.

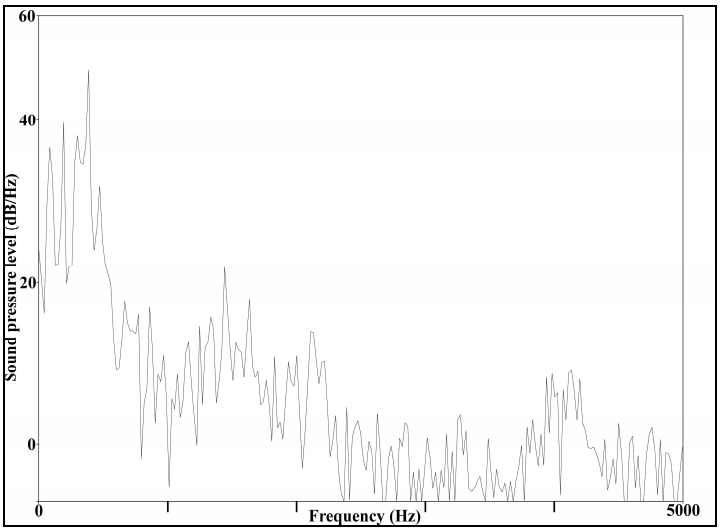

Figure 4c. Spectrum for the middle part of vowel /ø/ of a human male voice from 0 to 5 kHz.

But there are also differences in the spectral distributions of the *voces humanae*. Figure 4 (a and b) displays the harmonic distribution of energy for the C tones of the *voces humanae* in SIM and WAL, with the latter having a much higher level of intensity.

Figure 5 (a and b) displays the locations of F_1, F_2 and F_3 for the *voces humanae* of SIM and WAL. One can see that the formant distribution of the WAL tones mainly reflected changes in F_1 (from 400 to 1300 Hz), whereas the tones from the SIM organ showed a larger variation in F_2. Compared to the formant space of human (male) voices (German speakers producing long, tense vowels, taken from Simpson, 1998), both organs generated a smaller vowel space. In addition, the organs' vowel spaces had higher average formant values than the human vowel space. This formant shift is illustrated in the very small overlap of the spaces for the SIM *vox humana* and the human voice.

Inspection of the F_3 values revealed a much wider formant range for the organs compared to a male voice. For instance, F_3 of the SIM organ ranged between 1900 and 2800 Hz, WAL between 2000 and 3000 Hz, whereas the F_3 of the human voice ranges between 2200 and 2500 Hz.

For two tones, c1 from MEI and SIM, respectively, maximal energy was found on the 7^{th} harmonic (at around 1970 Hz). This is in line with a previous study by Lottermoser (1983: 135) on the acoustics of reed pipes for the tone C. However, the maximal energy of all other tones from AMO, MEI and SIM were irregularly distributed on other harmonics. The tones for WAL could not be considered because the additional labial pipes changed the energy distribution in a substantial way (cp. the differences in the harmonic distribution in Fig. 4b).

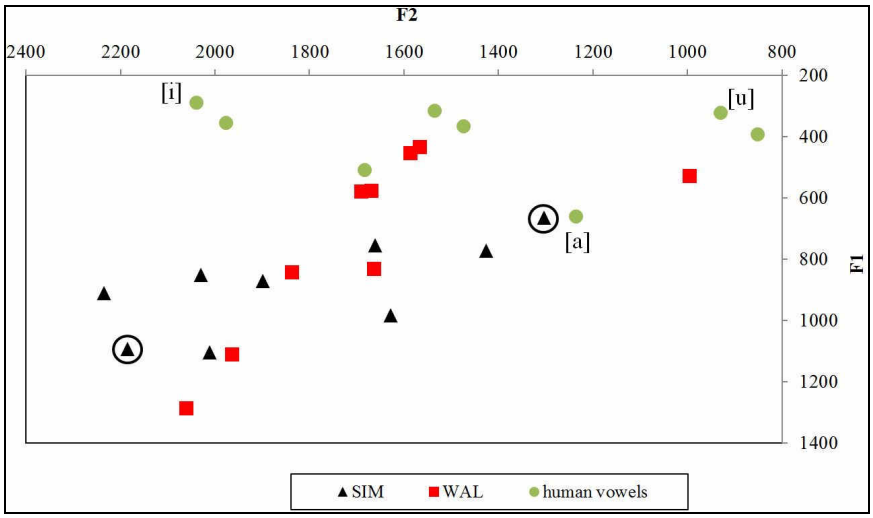

Figure 5a: Values for F_1 and F_2 of the tones of the *vox humana* in SIM (black triangles) and WAL (red squares) as well as standard values for the German long vowels of male voices (Simpson, 1998) (green dots). The encircled triangles indicate well recognised vowel qualities: [i] on the left, [ø] on the right.

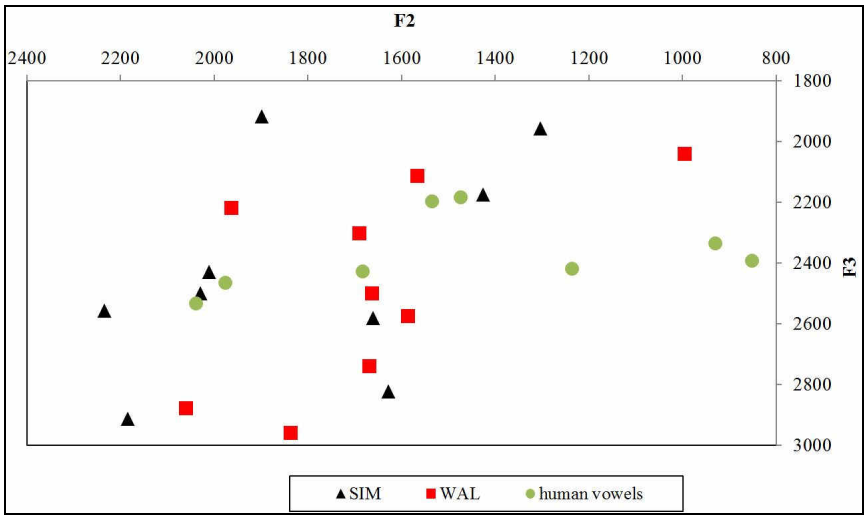

Figure 5b. Values for F_2 and F_3 of the tones of the *vox humana* in SIM (black triangles) and WAL (red squares) as well as standard values for the German long vowels of male voices (green dots) after Simpson (1998).

3.3 Discussion

The differences in fundamental frequencies for the same tone across organs can be explained by the fact that in the 18th century, the fundamental frequency had not yet been standardised with a fixed value (in contrast to today). Thus, the tuning of the tones could vary according to the region and to the size of the organ.

These data suggest that a central feature of the reed pipes from different *voces humanae* is a spectral distribution with a clear ″formant-like″ structure, illustrated by differentiated frequency bands of higher intensity. Formants could also be found for the reed stops *crumhorn* and *trumpet* (cp. Fig. 3), however, the distribution of these formants seemed to be less similar to those of the human voice. It is unclear whether further stops, especially flue stops, which account for two thirds of all pipes in a typical organ, also show formants[4]. Moreover, it is unclear which reed pipes show the largest similarity to the formants of human vowels.

The formant values of the *voces humanae* produced a vowel space that was smaller in size and with more upshifted formant values in comparison to a human speaking voice (cp. Fig. 5). This possibly could be explained by the smaller "vocal tract" of the investigated *voces humanae* in comparison to a human vocal tract which is given in text books with 17.0 cm x 4.5 cm (e.g., Pompino-Marschall, 2009: 160). This size is also in contrast to measures of the *voces humanae* from organs built by Stumm with pipe sizes of 14.0 cm x 2.7 cm for the tone g0. Fitch and Giedd (1999) reported average values of vocal tract length (based on MRI data) for young male adults (aged 19 to 25) of only 15.0 cm, whereas the values for 13-16 year old male children were 14.0 cm. The latter correlates with the length of the resonator of a *vox humana* pipe. This agreement is interesting when considering that the use of the *vox humana* stop once was a substitution for church boys' choirs.

The one-directional variation of the vowel space in the formant plane can also be explained with the resonator of the *vox humana* as a single-opened conic tube without any constrictions. It is usually assumed that vowels produced in a human vocal tract need two cavities, a back cavity and a front cavity. For a single cavity, as in the *vox humana* pipe, the higher formants should just be multiples of the first formant. However, this is not exactly the case when we compare the formant values in Table 2.

Future experiments with formant synthesis could show whether the measured formants from the organs can generate acoustic patterns which sound like humanoid vowels to the listener. Formant synthesis could also be used for experimentation with spectral tilt, which is reduced in the *vox humana* compared to a human voice. One explanation for this reduction in spectral tilt is the absorbing characteristics of the human oral cavity, which are not found in metal organ pipes.

The spectral distribution of the *vox humana* is partially different from that of human vowels. As already described in Lottermoser (1983: 135), the maximal energy of the highest tones can be found in the 7th harmonic of the solo played voces humanae. On the other hand, the fundamental frequency was hardly ever found to be the strongest harmonic (2 exceptions out of 27 tokens from AMO, MEI and SIM). The *vox humana* in WAL was played in combination with another stop which, in this case, caused the fundamental frequency to be the strongest harmonic.

[4] It is planned for future studies to record flue pipes as well. This would allow a comparison to reed pipes with respect to formant structures.

Table 2: Values of all tones for SIM and WAL for (measured) F_1, doubled F_1, (measured) F_2, three times F_1 and (measured) F_3.

	SIM					WAL				
Tone	**F_1**	**F_1*2**	**F_2**	**F_1*3**	**F_3**	**F_1**	**F_1*2**	**F_2**	**F_1*3**	**F_3**
C	982	1964	1628	2946	2823	453	906	1587	1359	2573
G	664	1328	1304	1992	1957	529	1058	996	1587	2040
c0	773	1546	1426	2319	2174	576	1152	1670	1728	2739
g0	754	1508	1661	2262	2580	434	868	1567	1302	2114
c1	871	1742	1899	2613	1919	843	1686	1838	2529	2959
g1	852	1704	2029	2556	2500	1286	1572	2061	3858	2878
c2	1104	2208	2011	3312	2430	578	1156	1690	1734	2301
g2	911	1822	2234	2733	2557	831	1662	1663	2493	2499
c3	1092	2184	2185	3276	2912	1110	2220	1964	3330	2218

4 Perception tests

The aim of the perception tests was to find out whether listeners could reliably associate the recorded tones to vowel categories. If so, it would be interesting to know more about the underlying nature of these perceptual impressions.

Two listening tests were performed. The first test could be seen as a pilot test, whereas the second test was a repetition of the first one, with substantial improvements. Since both tests were very similar, they are presented together.

4.1 Method

There were 40 stimuli for the first test consisting of the 36 *vox humana* tones plus the four tones from the stops *crumhorn* and *trumpet*. Each stimulus had a duration of 5 seconds. Twenty German linguists served as participants. The stimuli were presented via headphones in a randomised order and could be played as often as the participant wished. The participants were asked to indicate the vowel quality of each stimulus, if possible in terms of IPA cardinal vowels. There was also the option to say "no vowel". The answers were given in spoken form directly to the experimenter.

The second test was similar to the first one but with some modifications. This time, the experiment was performed using a web-based platform for the perception tests (with the help of Draxler, 2011) in order to test more participants (with German as their first language). In total, there were 29 participants, including linguists and non-linguists. The number of stimuli was reduced to 18 (using the *voces humanae* in SIM and WAL), plus the four tones from the stops *trumpet* and *crumhorn*. Each stimulus occurred three times, resulting in 66 stimuli presented in randomised order. Since the *voces humanae* from SIM and WAL showed the most contrasting results in the first test, these were selected for the second test. Each stimulus was shortened to 400 ms (taken from the middle part) in order to make it comparable to a long vowel in German. The vowel categories in the second test were the letters representing all long, tense vowels in German: I, Ü, E, Ö, Ä, A, O, U, which represent the vowels /i, y, e, ø, ɛ, a, o, u/. The first test revealed that only three out of twenty participants were able to use the IPA system, so letters were used to permit more consistent answers. The answer "no vowel" was not possible this time. For

technical reasons, one stimulus was not correctly played (c0 from WAL). Consequently, the corresponding results will not be presented.

4.2 Results

The results (see Table 3) for both *voces humanae* clearly indicated the correlation between the fundamental frequency and the vowel category. In other words, the higher the F_0, the more /i/-like the selected vowel and the lower the F_0, the more /o/-like the vowel.

The tones at the periphery (in terms of F_0, as well as F_1, F_2 and F_3) were assessed more consistently than those in the middle region. This is obvious for instance for the SIM tones in the second experiment, which revealed a stable c3 for /i/ (84%), but far less consistency for the next lower tone, g2 (between /i/ and /e/, with a tendency to /i/). The tone c1 is more or less equally distributed between the qualities of /e/, /ɛ/, /ø/ and /a/. For the corresponding tone of WAL, the listeners largely preferred /a, ɑ/ (test 1) or even /u/ (test 2).

The tones from the comparative stops *crumhorn* and *trumpet* showed less consistent answers than those of the *voces humanae*, especially for the tone g0. Comparing the results of the organs of SIM and WAL, it was evident that the tone-vowel correspondences of SIM showed a higher level of consistency than those of WAL (except for C and the maverick answer for c1).

The general tendencies of the first perception test were confirmed by the second, but on a more reliable basis. The results were sometimes clearer (e.g., for c0 and g1 in SIM) and often led to a higher level of consistency for SIM as well as for WAL.

Table 3. Percentages of answers for the stimulus tones of SIM and WAL for both perception experiments. The values for F_0 and the formants are in Hz. The stops were voces humanae (VH), crumhorn (CR) and trumpet (TR). Vowel categories in experiment 1 were clustered according to the German vowel letters. The most frequent answer for each tone is given in bold. Grey-shading of cells according to numbers: 100-80% (darkest grey), 79-60%, 59-40%, 39-20% (lightest grey), 19-0% (no shading).

Experiment 1																	Experiment 2								
no V	i	y	e	ɛ	ø/œ	a/ɑ	ɑ/ɔ	u	Σ	stop	tone	F_0	F_1	F_2	F_3	i	ü	e	ä	ö	a	o	u	Σ
40	**0**	**0**	**0**	**0**	25	**10**	25	**0**	100	VH Simonsen	C	69	982	1628	2823	**0**	**2**	**2**	**8**	31	**7**	38	11	100
5	**0**	**0**	**0**	10	**80**	**0**	**0**	**5**	100		G	102	664	1304	1957	**0**	**1**	**2**	**8**	**85**	**2**	**1**	**0**	100
5	**0**	**0**	**5**	35	35	**10**	**10**	**0**	100		c^0	136	773	1426	2174	**0**	**1**	**14**	**15**	64	**3**	**2**	**0**	100
20	**0**	**0**	20	35	**10**	**10**	**5**	**0**	100		g^0	205	754	1661	2580	**0**	**0**	**15**	**36**	41	**7**	**1**	**0**	100
15	**0**	**0**	**15**	25	**15**	20	**5**	**5**	100		c^1	274	871	1899	1919	**2**	**2**	20	21	29	20	**3**	**3**	100
20	**0**	**0**	35	**20**	**0**	20	**5**	**0**	100		g^1	408	852	2029	2500	**5**	**6**	59	**16**	**6**	**6**	**0**	**3**	100
25	**5**	**5**	20	**5**	**0**	35	**0**	**5**	100		c^2	548	1104	2011	2430	**15**	**13**	37	**2**	**8**	23	**1**	**1**	100
30	50	**0**	**10**	**0**	**0**	**5**	**0**	**5**	100		g^2	818	911	2234	2557	51	**7**	25	**2**	**3**	**7**	**1**	**3**	100
15	**80**	**0**	**0**	**0**	**0**	**5**	**0**	**0**	100		c^3	1093	1092	2185	2912	**84**	**9**	**3**	**0**	**0**	**3**	**0**	**0**	100
15	**0**	**0**	**0**	**5**	40	**15**	**20**	**5**	100	VH Wahrenhausen	C	69	453	1587	2573	**0**	**1**	**1**	**2**	33	**1**	40	21	100
10	**0**	**0**	**0**	**0**	20	30	35	**5**	100		G	103	529	996	2040	**1**	**1**	**6**	**3**	40	**9**	32	**7**	100
15	**0**	**5**	**10**	**0**	65	**0**	**0**	**5**	100		g^0	208	434	1567	2114	**1**	26	**15**	**1**	32	**3**	**9**	11	100
10	**0**	**5**	**10**	**5**	**10**	35	**5**	20	100		c^1	277	843	1838	2959	**2**	**9**	**10**	**2**	**10**	**5**	**5**	56	100
25	**10**	**15**	20	**5**	**5**	20	**0**	**0**	100		g^1	415	1286	2061	2879	**9**	29	28	**2**	**13**	**6**	**5**	**9**	100
35	20	35	**5**	**0**	**0**	**5**	**0**	**0**	100		c^2	554	578	1690	2301	33	21	28	**1**	**7**	**2**	**0**	**8**	100
35	**15**	35	**0**	**0**	**0**	**5**	**0**	10	100		g^2	832	831	1663	2499	39	22	**3**	**1**	**1**	**8**	**3**	22	100
30	35	**5**	**0**	**5**	**5**	**15**	**0**	**5**	100		c^3	1109	1110	1964	2218	75	11	**2**	**1**	**0**	**9**	**0**	1	100
30	**0**	**0**	**0**	**5**	**5**	40	**15**	**5**	100	CR	C	69	1097	1688	3014	**0**	**0**	**3**	**13**	28	30	**17**	**9**	100
10	**0**	**0**	**0**	**15**	35	40	**0**	**0**	100		g^0	205	941	1434	1989	**1**	**0**	28	20	34	**14**	**3**	**0**	100
35	**0**	**5**	**0**	**0**	**10**	**15**	30	**5**	100	TR	C	69	695	1494	2001	**0**	**1**	**1**	**6**	20	**14**	33	25	100
35	**5**	**10**	30	**0**	**0**	**10**	**0**	**10**	100		g^0	205	1576	1747	2631	**6**	**16**	22	**2**	**13**	**10**	**8**	23	100

4.3 Discussion

The perception experiments demonstrated that some, though not all, tones were reliably associated with vowels. This was definitively the case for the tones c3 as /i/ and G as /ø/ in SIM. The association rates for these two tones as human vowels were similar to the recognition rates of human vowels produced in CV and VC English syllables (Weber and Smits, 2003), where some vowels reached recognition rates as low as 45%. This is particularly evident for vowels that are not at the periphery of the vowel system. This finding can be compared with some of our results, for instance that of c1 for SIM. In both listening tests, this particular tone was associated with an even distribution between /e/, /ɛ/, /ø/ and /a/ as an area covering non-high and non-back vowels. Interestingly, all but one of about twenty visitors to the poster presentation at the Interspeech conference (with various language backgrounds) associated c1 of SIM with /ɛ/ when listening to it via headphones.

The tones from the *vox humana* in WAL produced less consistent association rates than the SIM tones and those from the other two churches (not reported here). In WAL, the tones were recorded in combination with the flue pipes from the *stopped diapason* leading to a different spectral distribution: the lower harmonics and the fundamental frequency were quite strong compared to the other organs. The two different stops merge to a new synthetic tone colour and were not perceivable separately.

There was a very strong relationship between the F_0 of the tones and their perceived vowel quality, which can be traced back to sound symbolism (Ohala, 1994). However, F_0 alone cannot explain these results. Obviously, the formant structure also plays a role. For instance, in SIM, the tone c3, reliably associated with /i/, showed very high values for F_2 and F_3; whereas G, heard as /ø/, possessed the lowest values for these formants.

It is striking to see that other stops with reed pipes, in our case *crumhorn* and *trumpet,* did *not* show as consistent results as the *voces humanae*, although F_0 and formant structure were also present there. Obviously, a *vox humana* was able to produce more human vowel-like sounds than *crumhorn* and *trumpett* the other organ stops that also use a reed pipe.

5 Conclusion

We could partially replicate the historically documented enthusiastic impression of the *vox humana* as an instrument with which it is possible to play human-like vowels. Although it is not clear how to explain this effect, we could show that *voces humanae* differ from other organ stops with reed pipes in terms of similarity to the human voice. This is interesting because von Kempelen (1791) used an excitation mechanism similar to a reed pipe in his famous speaking machine (see Kempelen (1791) for the original text and e.g. Brackhane (2011) for a historical reception).

Since we focused on isolated tones in this project, we cannot say anything about the influence of temporal and intensity dynamics, which can possibly explain the *vox humana* as a vowel synthesiser to a certain degree. The desire to generate

isolated vowels with the help of separate *vox humana*–like pipes was also required in the second part of the prize question of the St. Petersburg academy in 1780: "1) Qualis sit natura et character litterarum vocalium a, e, i, o, u tam insigniter inter se diversorum. 2) Annon construe queant instrumenta ordini toborum organicorum, sub termin vocis humanae noto, similia, quae litteratum vocalium a, e, i, o, u, sonos exprimant. (What is the nature and character of the vowels a, e, i, o, u, which are so different from each other? Is it possible to construct an instrument like the organ pipes called *vox humana* that can produce the vowels a, e, i, o, u?)" [translation of the authors from Kratzenstein (1781)]. Kratzenstein won the prize by producing /a, e, o, u/ according to the principles of the *vox humana* with a small organ consisting of four reed pipes, but for /i/ he used a flue pipe. Our study shows that more vowels than those can be convincingly produced with a *vox humana* in an organ, including an /i/.

The *vox humana* is definitively a fascinating musical instrument, which is partially able to generate human speech. However, the *vox humana* is not a genuine mechanical vowel synthesiser as hoped in historical times.

6 Acknowledgements

The authors thank Christoph Draxler for his support with the second perception test as well as Bernd Möbius, Eva Lasarcyk, Peter Birkholz, Coriandre Vilain and John Ohala for feedback on this research. Our thanks go also to the visitors of the poster presentation at Interspeech 2013 in Lyon. We are also grateful to the anonymous reviewer whose comments helped to improve this paper as well as to Ruth Huntley-Bahr.

References

Adelung, W. 1982. *Einführung in den Orgelbau*. Wiesbaden: Breitkopf.

Brackhane, F. 2011. Die Sprechmaschine Wolfgang von Kempelens – von den Originalen bis zu den Nachbauten. *Phonus 16* (Reports in Phonetics, Saarland University), pp. 49-148.

Draxler, Chr. 2012. Percy - An HTML5 framework for media rich web experiments on mobile devices. *Proc. 12th Interspeech*, Florence, pp. 3339-3340.

Eberlein, R. 2007. Vox humana. In: H. Busch and M. Geutig, (eds) *Lexikon der Orgel*. Laaber: Laber-Verlag.

Euler, L. 1773. *Briefe an eine deutsche Prinzessin über verschiedene Gegenstände aus der Physik und Philosophie: Aus dem Französischen übersetzt. Band 2*. Leipzig: Junius.

Fitch, W.T. and J. Giedd 1999. Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *Journal of the Acoustical Society of America 106*(3), pp. 1511-1522.

Frotscher, G. 1927. *Die Orgel*. Leipzig: Weber.

Greß, H. 2007. *Die Orgeln Gottfried Silbermanns*. Dresden: Sandstein.

Kempelen, W. v. 1791. *Wolfgangs von Kempelen Mechanismus der menschlichen Sprache nebst Beschreibung seiner sprechenden Maschine*. Wien: Degen.

Kratzenstein, Chr.G. 1781. *Tentamen resolvendi problema ab academia scientiarium imperiali petropolitana ad annum 1780 propositum*. St. Petersburg: Academia Scientiarum.

Lottermoser, W. 1936. *Klanganalytische Untersuchungen an Orgelpfeifen*. Berlin: Junker &

Dünnhaupt.
Lottermoser, W. 1983. *Orgeln, Kirchen und Akustik. Bd. 1*. Frankfurt/ Main: Bochinsky.
Ohala, J.J. 1994. The frequency code underlies the sound symbolic use of voice pitch. In: L. Hinton, J. Nichols and J. J. Ohala (eds): *Sound Symbolism*. Cambridge: Cambridge University Press, pp. 325-347.
Pompino-Marschall, B. 2009. *Einführung in die Phonetik* (3rd edition). Berlin: de Gruyter.
Simpson, A. 1998. *Phonetische Datenbanken des Deutschen in der empirischen Sprachforschung und der phonetischen Theoriebildung*. (Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK) 33). Kiel.
Weber, A., and R. Smits 2003. Consonant and vowel confusion patterns by American English listeners. *Proc. 15th International Congress of Phonetic Sciences*, Barcelona, pp. 1437-1440.

EFFECTS OF RHYTHM ON ENGLISH RATE PERCEPTION BY JAPANESE AND ENGLISH SPEAKERS

Yoshito Hirozane[1], and Robert Mannell[2]
[1]Mejiro University, Japan
[2]Macquarie University, Australia
e-mail: hirozane@mejiro.ac.jp, robert.mannell@mq.edu.au

Abstract

An experiment was conducted with Japanese speakers to test the hypothesis that the stress-timed rhythm can be a source of their perception of faster English rate. The results did not strongly support the hypothesis. Another experiment was conducted with English speakers to examine the possibility that the results of the first experiment had been affected by the knowledge of the English phonological structure acquired by the Japanese participants. The results implied that they might have been.

1 Introduction

We often feel that foreign languages are spoken quickly (Roach, 1998). Native speakers, however, do not seem to find their mother tongue to be as fast as nonnative speakers do.

Grosjean (1977) had a French passage read in five different tempos to native French speakers and native English speakers who had no knowledge of French, and asked them to evaluate perceptual tempos using the magnitude estimation method, which requires subjects to estimate the magnitude of the stimuli by assigning numerical values proportional to the stimulus magnitude they perceive. For all tempos, the evaluation by the native English speakers who had no knowledge of French was that it was perceived as faster than was the case for the native French speakers.

Schwab and Grosjean (2004) had short French passages read in three different tempos to native and nonnative French speakers and asked them to judge the rate using the magnitude estimation method. Non-native speakers of French perceived French passages as faster than did the native speakers of French. The difference in rate perception between native and non-native speakers became greater as the tempo increased. The comprehension level of the passages was correlated with the rate evaluation. The lower the comprehension level, the faster was the rate evaluation.

Pfitzinger and Tamashima (2006) conducted an experiment similar to Grosjean (1977) under symmetric conditions. They had native German speakers and native Japanese speakers listen to German and Japanese spontaneous speech spoken in different tempos, and asked them to evaluate these rates perceptually. The native

Japanese speakers evaluated the German speech rate as 7.47% faster than the native German speakers. On the other hand, the native German speakers evaluated the Japanese speech in rate as 9.13% faster than the native Japanese speakers. These findings suggest that the perceptual tempo of speech is not the same between native and non-native speakers, and that non-native speakers tend to perceive speech as faster than native speakers.

It seems that the same holds true for Japanese speakers listening to English. The average speech rate of British English is 230–280 syllables per minute (Tauroza and Allison, 1990), which is equivalent to 3.8–4.7 syllables per second. According to Griffiths (1992), Japanese learners of English at the lower intermediate level begin to find it hard to understand English when its speech rate exceeds 3.8 syllables per second. Although comprehension difficulty is not always correlated with perceptually faster rate, it is quite likely that Japanese learners of English perceive most of the English utterances, which native speakers of English consider normal in rate, as fast. Why is it that exactly the same speech is perceived as different in rate between native and non-native speakers of English?

The rate perception of a foreign language cannot be independent of the level of the listener's competence in that particular language as long as the listener tries to understand what is being said. It is assumed that the more you understand a foreign language, the slower you perceive it to be. Such being the case, anything which interferes with comprehension could be a source of faster perceived rate. The recognition of speech sounds and the mapping of sound to meaning are two of the major components of the speech perception process (Cutting and Pisoni, 1978; Massaro, 1975; Pisoni, 1975). Hindrance to the functioning of either component would lead to poor comprehension. In L2 listening, the phonetic features of a foreign language, which are quite different from those of L1, could hinder the recognition of speech sounds, and thus lead to faster perceived rate.

Among such phonetic features is a language characteristic rhythm. Roach (1998) assumes that "syllable-timed speech sounds faster than stress-timed to speakers of stress-timed languages". His assumption is based on his speculation that "… if a language with a relatively simple syllable structure like Japanese is able to fit more syllables into a second than a language with a complex syllable structure like English or Polish, it will probably sound faster as a result." In other words, he assumes that Japanese sounds faster than English to the ear of English speakers because, for structural reasons, more Japanese syllables tend to be produced per second than English syllables. His explanation of Japanese being perceptually faster than English is reasonable. But how would he explain the fact that English is perceptually faster than Japanese to the ear of Japanese speakers? Our hypothesis is that a mere difference in rhythm could be a source of faster perceptual rate for a foreign language. It is possible that Japanese speakers' perception of English as fast is caused by the stress-timed rhythm which is characteristic of English.

If English sounds fast to the ear of Japanese listeners because of its characteristic rhythm, which is quite different from the one used in Japanese, the perceptual rate

will be reduced by eliminating the characteristic rhythm from English and approximating it instead to that of Japanese. What is referred to as the characteristic rhythm of English and Japanese here is the stress-timed rhythm and the mora-timed rhythm respectively.

In languages that are said to have stress-timed rhythm, stressed syllables tend to occur at relatively regular intervals (Roach, 2009). On the other hand, in languages that are said to have mora-timed rhythm, all mora syllables tend to have equal durations. Stressed syllables are more prominent than unstressed syllables due to four main factors (Morton and Jassem, 1965): loudness, length, pitch, and vowel quality. If these parameters are appropriately controlled, the prominence will be leveled out and the stress-timed rhythm will disappear. Syllables ideally controlled to have equal prominence to each other should have equal loudness, duration, and vowel quality to each other.

It would be difficult to control these parameters of natural speech. But those of the synthetic speech can be controlled comparatively easily. The Festival Speech Synthesis System (Black and Clark, 2003) (referred to as Festival hereafter) has a function which controls stress and intonation with ToBI annotation. By editing scripts, you can add or remove stress without worrying about fine-tuning the parameters. In Experiment 1, pairs of English sequences synthesized by Festival, one of which was a sequence with stress-timed rhythm and the other was a sequence approximated to a mora sequence by removing stress from the first one, were presented to Japanese speakers to test if there is any significant perceptual difference in rate caused by the stress-timed rhythm.

2 Experiment 1

2. 1 Methods

2.1.1 Participants. Twenty-three native Japanese speakers (4 males and 19 females) participated in the experiment. They were all undergraduate students of Mejiro University in Tokyo, Japan. All of them majored in English. Before they entered university, they had had at least 6 years of English education since junior high school. Their English skills were at a lower intermediate level on average. None of the participants had any hearing loss or hearing impairment.

2.1.2 Stimuli. The present experiment used a pair of English tokens, one of which had stress-timed rhythm and the other mora-timed rhythm. It was not very easy, however, to realize mora-timed rhythm in English.

A major characteristic of mora syllables is that they each have, not exactly, but approximately equal durations. Producing every syllable with approximately equal duration is possible in the case of Japanese because its syllable structure is very simple. Most of the mora syllables are in the form of CV or V, and small numbers of them are CjV as in *kyo* “home”, mora nasal N as in *pa-N* “bread”, and geminate Q as in *ka-Q-ta* “bought”. Mora syllables consist of one (V, N, or Q) to three (CjV) segments. It is not hard to pronounce each of them within the same amount of time.

English syllable structure, on the other hand, is more complex than that of Japanese. The basic structure of the English syllable is (CCC)V(CCCC). One syllable can consist of one to as many as eight segments. Variation in syllable length is so great that it would be much more difficult to pronounce English syllables so that each syllable would have approximately the same duration. Even if successfully pronounced, whether they can be said to have mora-timed rhythm or not is open to question.

Since the difficulty of keeping syllable durations constant in English arises from its complex syllable structure, one solution to overcome this is to confine all the syllables to the Japanese basic syllable structure, namely CV. The tokens were easier to make with sequences of nonwords rather than meaningful sentences of real words because it is very difficult to generate meaningful English sentences made up solely of CV words. The nonsense CV syllable chosen for the tokens used in this experiment was /da/.

The stress rhythm tokens were synthesized so that they represented four of the typical meters of English: iambic, trochaic, anapaestic, and dactylic (referred to as WS, SW, WWS, SWW respectively hereafter). The mora rhythm tokens were synthesized so that they had F_0 contours similar to those of the corresponding stress rhythm tokens.

Since the making of these tokens was a little complicated, it is explained below in more detail.

Festival 1.4.3 was used for the synthesis. There were three major steps needed to synthesize speech with Festival: 1) write a script with Scheme and 2) input the script to Festival, and then 3) Festival returns synthesized speech. Synthesized speech can be controlled by editing scripts. For the control of stress and intonation, ToBI annotation is available in Festival.

Table 1: Base sentences and their corresponding sequences as a result of replacement of each syllable by /da/. 'da' represents unstressed /da/ and 'DA' represents stressed /da/.

Meter	**Base sentences**	**Sequences as a result of replacement of each syllable by /da/**
WS	We go to school by bus at eight.	da DA da DA da DA da DA
SW	Every girl was crying sadly.	DA da DA da DA da DA da
WWS	The police have arrested the thief on the spot.	da da DA da da DA da da DA da da DA
SWW	Everyone thought it was anything but wonderful.	DA da da DA da da DA da da DA da da

Four meaningful sentences having one of the four typical English meters were synthesized with Festival (see Table 1). The script for the sentence with the WWS meter is shown below as an example. Appendix 1 shows all the scripts.

(set! utt1 (Utterance Words (the (police ((accent H*)(tone H-H%))) have (arrested ((accent L*)(tone L-))) the (thief ((accent L*)(tone L-))) on the (spot ((accent H*)(tone L-L%))))))

The synthetic speech returned from Festival represented the WWS meter very accurately as shown in Figure 1. The stressed syllables are the longest among the adjacent three syllables in either direction and are often accompanied by pitch movement.

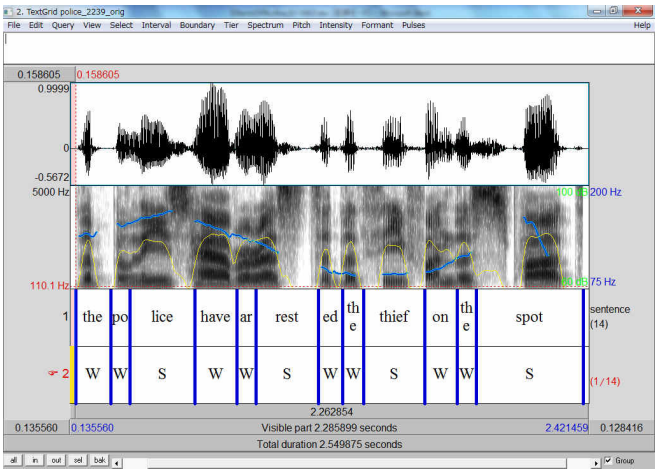

Figure 1. Waveform, spectrogram, F_0 movement (blue lines over the spectrogram), and intensity (yellow lines over the spectrogram) of the "police" sentence, "The police have arrested the thief on the spot."

By modifying the scripts for the base sentences, sequences of /da/ were synthesized. Below is the script used for the synthesis of the WWS sequence. Appendix 2 shows all the scripts.

(set! utt1 (Utterance Words (da da (daa ((accent H*)(tone H-H%))) da da (daa ((accent L*)(tone L-))) da da (daa ((accent L*)(tone L-))) da da (daa ((accent H*)(tone L-L%))))))

Compare the script with the one for the base sentence. In this script, all of the unstressed syllables of the base sentence were replaced with the unstressed /da/, which is represented by 'da', and all of the stressed syllables were replaced with the stressed /da/, which is represented by 'daa'. The stressed /da/ could be just 'da'. However, 'daa' was chosen as a preferable sound because it was longer in duration and different in vowel quality than the unstressed /da/, which would help make the syllable more prominent than the other unstressed ones (See Figure 2).

Since ToBI annotations were not modified at all, the sequence returned from Festival retained almost the same pattern of F_0 contour. (Compare Figures 1 and 2.) By modifying the script the same way, three other sequences of WS, SW, SWW meter were obtained.

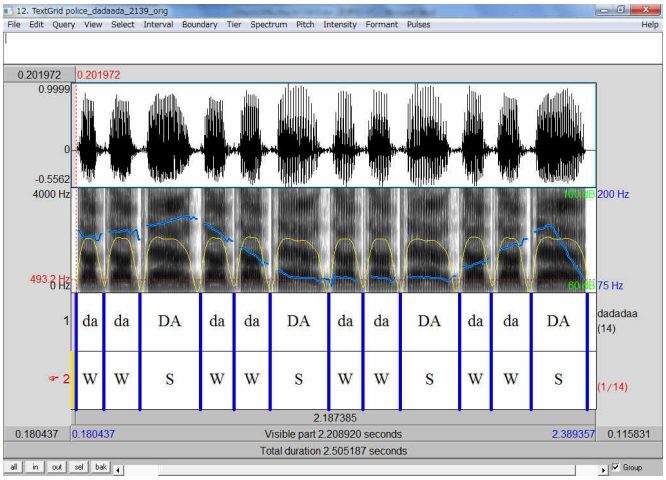

Figure 2. Waveform, spectrogram, F_0 movement, and intensity of the stress rhythm sequence "DA" in this diagram is equivalent to "daa" in the Festival scripts in the text of this paper.

The sequences with mora rhythm were also produced by modifying the scripts for the base sentences. This time each syllable of the base sentence was replaced by the unstressed /da/, which is indicated by 'da' in the script. The script for the sequence corresponding to the base sentence with WWS meter is shown below as an example. Appendix 3 shows all the scripts.

```
(set! utt1 (Utterance Words (da da (da ((tone H-H%))) da da (da ((tone L-)))
da da (da ((tone L-))) da da (da ((accent H*)(tone L-L%))) )))
```

The difference from the script of the stress rhythm sequence is that all the markups for stress, such as (accent H*), except the one for the last syllable were deleted, and every daa was replaced by da.

The sequences of /da/ returned from Festival after running the script had almost the same F_0 contour pattern as the corresponding stress rhythm sequence. (Compare Figures 2 and 3). Perceptually, none of the /da/ syllables were more prominent than the others and the sequences all had the appropriate rhythm, which was quite similar to the Japanese mora-timed rhythm. A total of eight sequences (4 meters x 2 rhythms) were obtained by synthesis.

Since the purpose of the experiment was to investigate the effects of rhythm on rate perception, parameters other than rhythm, especially the physical rates, had to be kept identical within each pair of the stimuli. However, the sequences obtained so far were still different in terms of physical rate.

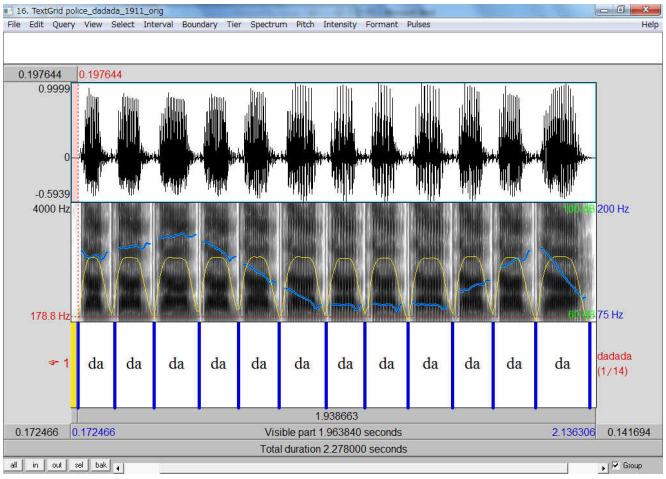

Figure 3. Waveform, spectrogram, F_0 movement, and intensity of the mora rhythm sequence

Physical rate can be defined as the number of linguistic units produced per unit time. The linguistic units counted for rate measurement could be words, syllables, moras, phonemes etc., depending on the purpose of the measurement. Both of the paired sequences obtained so far had the same array of segments (simple succession of /da/) and the same number of units (the same number of /da/), but they were different in duration. For all meters, the stress rhythm sequence was longer than the mora rhythm sequence. For the physical rates to be identical between the paired sequences, they had to have the same length. So, the stress rhythm sequences were compressed to the length of the corresponding mora rhythm sequences. We did not choose to extend the mora rhythm sequences because extension often makes resultant sequences sound like they are spoken by a drunken or tired person and such connotations may affect rate perception. The actual adjustment was done by the duration manipulation function of Praat (Boersma and Weenink, 2009).

After the adjustment, the paired sequences had the same physical rate and a very similar F_0 contour but different rhythms. These sequences could now serve as the tokens for the present experiment.

2.1.3 Procedures

The whole experiment was conducted through Praat. The participant was seated in front of a computer screen showing three rectangles lined up horizontally which were labeled "1st", "same", "2nd," from left to right in this order. Four pairs of sequences, each of which was composed of the stress rhythm sequence and the mora rhythm sequence elicited from the identical English sentence, were randomly presented four times to the participant over the headphones. In other words, eight different sequences were randomly presented in pairs on the condition that each of the paired sequences had been elicited from the identical English sentence. The order of presentation of the stimulus pairs was counterbalanced across the participants. The participant was asked to indicate which sequence of a given pair

sounded faster or if both sounded the same in terms of rate by clicking one of the three rectangles on the screen. The experiment was conducted with one participant at a time. The total number of trials was 16.

2.2 Results and Discussion

Table 2 shows the number of times Japanese speakers indicated that the stress or mora sequence was faster or both sounded the same in rate. A binomial test[1], at an alpha level of 0.05, revealed that, except for the WWS meter, there was not a significant difference between the number of times that Japanese speakers perceived the stress sequence as faster and the number of times that they perceived the mora sequence as faster.

A chi-square test of goodness-of-fit[2] was performed at an alpha level of 0.05 to determine whether the stress and mora sequences were equally judged to be faster. This test revealed that the responses were not equally distributed in the population, χ^2 (3, N = 311) = 9.29, $p < .05$. As a whole, Japanese speakers judged the stress sequences as faster than the mora sequences.

Table 2. Number of times Japanese speakers indicated the stress or mora sequence was faster or the same after hearing pairs of stress and mora rhythm sequences

Meter	Stress	Mora	Same	Binomial test results
WS	43	31	18	
SW	34	44	14	
WWS	47	31	14	* (p = .044)
SWW	48	33	11	
Total	172	139	57	

Approximating the rhythm of English to the rhythm of Japanese did not greatly slow down the perceived rate of the sequences by Japanese speakers. The stress sequences appeared to be only slightly faster than the mora sequences for Japanese speakers. The hypothesis that the stress rhythm could be a source of Japanese perception of English as fast was not strongly supported.

Are these results enough to conclude that the stress rhythm of English does not affect the rate perception by Japanese? Before making a conclusion, there are a couple of things to consider. English was not a totally unfamiliar language to the Japanese participants who had been learning English for years. These results may reflect part of their learning outcome. It is possible that they might have familiarized themselves with the apparently rapid tempo of English attributed to stress rhythm at some earlier stage of their learning. Neither of the two rhythms, stress and mora, may have sounded faster than the other because they had overcome the apparent rapidity caused by the stress rhythm which once was quite unfamiliar for them. If

[1] The “same” responses were excluded from the binomial test.
[2] The “same” responses were excluded from the chi-square test of goodness-of-fit.

the present experiment is conducted with Japanese speakers to whom stress rhythm was totally unfamiliar, there is still a chance that they might perceive the stress sequences as faster than the mora sequences.

Another thing to note in the present experiment is that the experimental task did not require lexical access because the stimulus tokens were all nonwords. The participants had to evaluate the rate solely on the basis of the phonetic information contained in the stimuli. Most of the time, however, the evaluation of English rate by Japanese, especially by English as Foreign Language (EFL) learners, is accompanied by lexical access. In this respect, the rate evaluated by the participants in the present experiment was not exactly the same as the English rate commonly perceived by Japanese listeners. Rate evaluation without lexical access would be less affected, but not unaffected, by one's knowledge of the language than would be the case for rate evaluation with lexical access because the former could be made without any knowledge of the language. The two types of evaluation should be strictly distinguished in the evaluation of rate perception.

To find out more about the effects of stress rhythm on rate perception by Japanese listeners without lexical access, the same experiment ideally ought to be conducted with Japanese speakers to whom stress rhythm is totally unfamiliar. Such people, however, are hard to find these days in Japan where English is taught as a compulsory subject in junior high school and will soon be compulsory in elementary school as well. It is much easier to find English speakers to whom mora rhythm is totally unfamiliar. Conducting the same experiment with English speakers instead of Japanese speakers would tell us, though indirectly, whether the knowledge of a second language (L2) can affect rate perception or not.

If the results of Experiment 1 do not reflect the Japanese participants' knowledge of English acquired over the years, English speakers with no knowledge of Japanese should also perceive the stress and mora sequences as the same in rate. If, on the other hand, the results of Experiment 1 do reflect the Japanese participants' knowledge of English because of instruction, English speakers with no knowledge of Japanese should perceive the mora and stress sequences as different in rate.

3 Experiment 2

In Experiment 1, we found that the two different rhythms, stress and mora, did not induce Japanese speakers to perceive differences in rate which physically did not exist in the sequences. The purpose of Experiment 2 is to test if English speakers' rate perception of the sequences is also unaffected by the rhythmic difference.

3.1 Methods

3.1.1 Participants. Twenty-five native English speakers (25 females) participated in the experiment. They were all undergraduate students of Macquarie University in Sydney, Australia. None of them had studied Japanese as a foreign language before the experiment. None of the participants had any hearing loss or hearing impairment.

3.1.2 Stimuli. The stimuli were the same as those used in Experiment 1.

3.1.3 Procedures. The experiment was conducted in small groups in the speech perception laboratory at Macquarie University, Sydney, Australia. No computer software was used. Four pairs of sequences, each of which was composed of the stress rhythm sequence and the mora rhythm sequence elicited from the identical English sentence, were randomly presented four times to the participants over the headphones by way of a CD player. The stimulus pairs were randomly presented but were not counterbalanced across the participants. After listening to each stimulus pair, the participants were asked to indicate their responses on a sheet of paper by circling one of the three options ("1st faster", "same", "2nd faster") printed on it. The three options corresponded to the three response rectangles presented to the Japanese speakers. The total number of trials was 16.

3.2 Results and Discussion

Table 3 shows the number of times English speakers indicated the stress or mora sequence was faster or both sounded the same in rate.

Table 3. Number of times English speakers indicated the stress or mora sequence was faster or the same after hearing pairs of stress and mora rhythm sequences

Meter	Stress	Mora	Same	Binomial test results
WS	33	52	15	* ($p = .025$)
SW	29	59	12	** ($p < .001$)
WWS	20	60	20	** ($p < .001$)
SWW	20	62	18	** ($p < .001$)
Total	102	233	65	

A binomial test, at an alpha level of .05, revealed that for all meters, there was a significant difference between the number of times that English speakers perceived the stress sequence as faster and the number of times that they perceived the mora sequence as faster. English speakers perceived the mora sequences as faster than the stress sequences.

A chi-square test of goodness-of-fit was performed, with an alpha level of 0.05, to determine whether the stress and mora sequences were equally judged to be faster. The test revealed that the responses were not equally distributed for this population, χ^2 (3, N = 335) = 55.99, $p < .001$. Not only for individual meter but also as a whole, English speakers judged the mora sequences as faster than the stress sequences.

The results showed that the mora sequences sounded faster than the syllable sequences even when both had the same number of syllables per second and the only difference was the rhythm. What affected the rate perception in this particular experiment was not the physical rate, but the rhythmic difference.

A rhythmic type usually has nothing to do with rate. It seems illogical, then, that the rhythmic difference alone, without any physical difference in rate, affected rate perception. What the listeners based their rate judgments on could be more than what had been physically input through their senses. How they process the input seems to have more relevance.

When people listen to L2, they usually use the same segmentation strategy as they use to listen to their native language (L1) (Cutler et al. 1986, 1993). The same thing could hold true when people listen to sequences with unfamiliar rhythm. In this case, the English speakers had never studied Japanese and the mora rhythm was totally unfamiliar to them. It is likely that they applied their L1 segmentation strategy to the mora sequences.

Suppose English speakers used the Metrical Segmentation Strategy (Cutler, 1990). They would look for strong syllables for segmentation as they listen to the mora sequences, as well as the stress sequences. But they would never find them within the mora sequences because none of the mora syllables were more prominent than others. Instead, they would only find mora syllables, which are more similar to weak syllables in that both are low in prominence.

To their ears, the entire mora sequence would be very similar to successions of weak syllables, which in English tend to be pronounced more quickly than strong syllables. Without mora syllables corresponding to English strong syllables, they could not find among a series of mora syllables any foot which serves as the basis for counting beats in English poetry. Unable to recognize a foot, they may have been at a loss for what to do to decide the rate of the mora sequences. Perhaps the processing of the mora syllables went away before they could do anything to evaluate their rate.

In the stress sequences, on the other hand, they could find strong syllables from time to time, because they tended to be longer in duration and helped slow down the rate, at least momentarily. With the strong syllables. they could easily recognize feet which helped them count beats, as they usually do with meaningful English sentences. This could be how English speakers evaluated the mora sequences as faster than the stress sequences, even when their physical rates were exactly the same.

4 General discussion

Since English speakers with no knowledge of Japanese perceived the mora sequences as faster than the stress sequences, it is assumed that Japanese speakers with no knowledge of English would perceive the stress sequences as faster than the mora sequences. If this assumption is the case, the results of Japanese speakers in this experiment reflected the outcome of English learning by the Japanese participants. So, what have the Japanese learned to do through years of English learning? Cutler and colleagues (1989) showed that even bilingual listeners who had acquired English and French, despite their full command of both languages, could only use one of the two differing segmentation procedures: stress-based or syllable-based. The procedure available to them appears to depend on which language is dominant for them. French-dominant bilinguals used a syllable-based and English-dominant bilinguals used a stress-based segmentation procedure. They were no different from monolinguals in that they could not switch from one segmentation procedure to another depending on the rhythmic type of the language they listened

to. However, they differed from monolinguals in that the French-dominant bilinguals were able to suppress the application of a syllable-based segmentation procedure when they listened to English, because it would be inefficient to process English using a syllable-based segmentation procedure.

The Japanese participants, in this project, were not bilinguals. However, they had been sufficiently exposed to English to the extent that they knew the mora-based segmentation procedure, which is suitable for Japanese, did not help much in extracting words from a continuous flow of English. It is likely, then, that they had learned to suppress the application of the mora-based segmentation procedure when they listened to English, although the degree of suppression might not be equal to that of the bilinguals. Whatever the degree, the suppression of the unsuitable segmentation procedure will increase the efficiency of speech processing, thus contributing to the slowing down of the perceived rate of speech.

The stimulus tokens were all non-words. If the segmentation procedure was affecting rate perception, what kind of units were the participants extracting from the sequence which contained no real words? According to Ingram (2007), when native speakers of English were asked to indicate how many 'words' they heard in the nonce phrase "flant nemprits kushen signortle" spoken with a stress and intonation contour appropriate for "French-language teaching instructions", the most frequent response was four. They were also asked to indicate where the 'word' boundaries were and the most popular sites were those indicated by blank spaces. This example demonstrates that segmentation, which is part of language processing, is possible, even at the prelexical level, if one is familiar with the prosodic and phonological characteristics of the language.

The size of the units the listener can divide the sequence into depends on how much knowledge they have about the prosody and phonology of the language: the more knowledge the listener has, the larger the segmentation units. The larger the size of the segmentation units, the smaller the number of units the listener would recognize per unit time, which would lead to slower perceived rate.

English speakers could divide the stress sequences into units equivalent to words or higher-level constituents, but not the mora sequences because they were totally unfamiliar with the prosody and phonology of Japanese. They could recognize units larger than the syllables in the stress sequences, but they could not recognize units larger than the moras in the mora sequences. Since the paired mora and stress sequences had the same duration, the mora sequences sounded faster than the stress sequences to the English speakers.

The Japanese participants, on the other hand, could divide not only the mora sequences, but also the stress sequences, into units larger than individual syllables or moras because they had some knowledge of English prosody and phonology, as well as that of Japanese. This could be why they evaluated the rates of the stress and the mora sequences as the same. If they had never studied English before, they might have perceived the stress sequences as faster than the mora sequences, just as the English speakers perceived the mora sequences as faster than the stress sequences.

5 Conclusion

The results of Experiment 1 appear to show that the rhythm per se does not affect the perceived rate of English speech by Japanese speakers. But the results of Experiment 2 imply that rhythm might have influenced their perception before they began to learn English. They may have overcome the difficulty of accepting the differing rhythm from their native language as they were exposed to more and more English. In order to verify this hypothesis, further research is required.

References

Black, A.W., and R. Clark 2003. *The Festival Speech Synthesis System* (Version 1.4.3). Retrieved from http://www.cstr.ed.ac.uk/projects/festival/

Boersma, P., and D. Weenink 2009. *Praat: doing phonetics by computer* (Version 5.1.05). Retrieved from http://www.praat.org/

Cutler, A. 1990. Exploiting prosodic probabilities in speech segmentation. In G. Altmann (ed.): *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*. Cambridge, MA: MIT Press. pp. 105-121.

Cutler, A., J. Mehler, D. Norris, and J. Segui 1986. The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, 25(4), 385-400.

Cutler, A., J. Mehler, D. Norris, and J. Segui 1989. Limits on bilingualism. *Nature*, 340(6230), 229-230.

Cutler, A., J. Mehler, T. Otake, and G. Hatano 1993. Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language*, 32, 258.

Cutting, J.E., and DB. Pisoni 1978. An Information-Processing Approach to Speech Perception. In JF. Kavanagh and W. Strange (eds.): *Speech and Language in the Laboratory, School and Clinic*. Cambridge: MIT Press. pp. 38-73.

Griffiths, R. 1992. Speech Rate and Listening Comprehension: Further Evidence of the Relationship. *TESOL Quarterly*, 26(2), 385-390.

Grosjean, F. 1977. The perception of rate in spoken and sign languages. *Attention, Perception, and Psychophysics*, 22(4), 408-413.

Ingram, J.C. 2007. *Neurolinguistics: an introduction to spoken language processing and its disorders.* Cambridge: Cambridge University Press.

Massaro, D.W.E. 1975. *Understanding Language: An Information-Processing Analysis of Speech Perception, Reading, and Psycholinguistics.* New York: Academic Press.

Morton, J., and W. Jassem 1965. Acoustic Correlates of Stress. *Language and Speech*, 8(3), 159-181.

Pfitzinger, H.R., and M. Tamashima 2006. Comparing perceptual local speech rate of German and Japanese speech. Paper presented at the *3rd International Conference on Speech Prosody*, Dresden, Germany.

Pisoni, D.B. 1975. Information processing and speech perception. In G. Fant (ed.): *Speech Communication* Vol. 3. New York: John Wiley. pp. 331-337.

Roach, P. 1998. Some Languages are Spoken More Quickly Than Others. In L. Bauer, and P. Trudgill (eds.): *Language Myths.* London*:* Penguin. pp. 150-158.

Roach, P. 2009. English phonetics and phonology: a practical course (4th ed.). Cambridge: Cambridge University Press.

Schwab, S., and F. Grosjean 2004. La perception du débit en langue seconde. *Phonetica*, 61(2-3), 84-94.

Tauroza, S., and D. Allison 1990. Speech Rates in British English. *Applied Linguistics*, 11(1), 90-105.

Appendix 1 Festival scripts for Experiment 1 and 2 (base sentences)

WS

(set! utt1 (Utterance Words ((we((accent L*))) (go((accent H*)(tone H-H%))) to (school((accent L*)(tone L-))) by (bus((accent L*))) at (eight((accent H*)(tone L-L%))))))

SW

(set! utt1 (Utterance Words ((every((accent H*)))(girl((accent L*)))was(crying((accent L*)))(sadly((accent H*)(tone L-L%))))))

WWS

(set! utt1 (Utterance Words (the (police((accent H*)(tone H-H%))) have (arrested((accent L*)(tone L-))) the (thief((accent L*)(tone L-))) on the (spot((accent H*)(tone L-L%))))))

SWW

(set! utt1 (Utterance Words ((every((accent H*))) one (thought((accent L*))) it was (anything((accent H*))) but (wonderful((accent L*)(tone L-L%))))))

Appendix 2 Festival scripts for Experiment 1 and 2 (stress-timed da sequences)

WS

(set! utt1 (Utterance Words ((da((accent L*))) (daa((accent H*)(tone H-H%))) da (daa((accent L*)(tone L-))) da (daa((accent L*))) da (daa((accent H*)(tone L-L%))))))

SW

(set! utt1 (Utterance Words ((daa((accent H*))) da (daa((accent L*))) da(daa((accent L*))) da(daa((accent H*)(tone L-L%)))da)))

WWS

(set! utt1 (Utterance Words (da da (daa((accent H*)(tone H-H%))) da da(daa((accent L*)(tone L-))) da da (daa((accent L*)(tone L-))) da da (daa((accent H*)(tone L-L%))))))

SWW

(set! utt1 (Utterance Words ((daa((accent H*))) da da (daa((accent L*)(tone L-))) da da (daa((accent H*))) da da (daa((accent L*)(tone L-L%)))da da)))

Appendix 3 Festival scripts for Experiment 1 and 2 (mora-timed da sequences)

WS

(set! utt1 (Utterance Words ((da((accent L*))) (da((accent H*)(tone H-H%))) da (da((accent L*)(tone L-))) da (da((accent L*))) da (da((accent H*)(tone L-L%))))))

SW

(set! utt1 (Utterance Words ((da((tone H-)))da(da((tone L-))) da (da((tone L-))) da(da((tone H-))) (da((tone L-))))))

WWS

(set! utt1 (Utterance Words (da da (da((tone H-H%))) da da(da((tone L-))) da da (da((tone L-))) da da (da((accent H*)(tone L-L%))))))

SWW

(set! utt1 (Utterance Words ((da((accent H*))) da da (da((accent L*)(tone L-))) da da (da((accent H*))) da da (da((accent L*)(tone L-L%)))da da)))

CROSS-LINGUISTIC STUDY OF FRENCH AND ENGLISH PROSODY
F_0 Slopes and Levels and Vowel Durations in Laboratory Data

Katarina Bartkova and Mathilde Dargnat
University of Lorraine, ATILF, France
e-mail: {katarina.bartkova, mathilde.dargnat}@atilf.fr

Abstract

Prosody conveys linguistic and extralinguistic information through prosodic features which are either language dependent or language independent. In addition, each speaker has unique physiological characteristics of speech production and speaking style, and thus speaker-specific characteristics are also reflected in prosody. Distinguishing the language-specific and speaker-specific aspects of prosody using acoustic parameters is a very complex task. Therefore, it is very challenging to extract and represent prosodic features which can differenciate one language from the other or one speaker from the other. The goal of our study is to investigate whether the prosody of isolated sentences in French and English is determined by their shared syntactic structures and whether the prosodic features used by the two languages are different or similar. In our cross-linguistic comparison of the prosodic parameters, two approaches are used. First, F_0 slopes measured on target words in the sentences are analyzed by fitting mixed linear regression models (R package *lme4*). Secondly vowel duration and F_0 values for each syllable are prosodically annotated using an automatic prosodic transcriber and the symbolic and numeric values are used in a more qualitative comparison of our data. It appears from the analyzed data that the observed F_0 curves in our corpus do not always correspond to linguistic theory and that the output of the automatic prosodic transcriber provides relevent information for a cross-linguistic study of the prosody.

1 Introduction

Prosody is an important component of oral communication for transferring linguistic, pragmatic and extralinguistic information and gives the speech signal its expressiveness mainly through melody, intensity and sound duration. Variation of the prosodic parameters allows a listener to segment the sound continuum, and to detect emphasis on the speech signal (i.e., accent of words or expressions). The prosodic component of speech conveys the information used for structuring the speech message, such as emphasis on words and structuring the utterance into prosodic groups.

However the prosodic component of the speech signal is less easy to process than its segmental part as there are few constraints in the realization of its parameter values. Yet, prosodic information is difficult to add into the manual transcription of speech corpora, or other automatic speech processing. Hence, it is important to investigate automatic approaches for recovering such information from speech material.

Even if not perfect, the use of an automatic approach for prosodic annotation of the speech would be very useful especially as the agreement on manually annotated prosodic events (boundary levels, disfluences and hesitation, perceptual prominences) between expert annotators is quite low (68%). Even after training sessions, the agreement does not exceed 86% (Lacheret-Dujour et al., 2010) and the task can be considered even more difficult and complex when manual coding of pitch level is to be carried out. In fact, it is difficult for human annotators not to be influenced by the meaning of an utterance; annotators can be tempted to associate a prosodic boundary at the end of a syntactic boundary or at the end of a semantic group instead of focusing solely onto the prosodic events. Moreover, there can be a discrepancy between the parameter values and their perception by a human annotator. For instance, an acoustic final rise can be perceived as a fall depending on the preceding F_0 curve (Hadding-Koch and Studdert-Kennedy, 1964). Moreover the same F_0 contours can have non-standard occurrences (F_0 rises can be found at the end of declarative sentences) and a human transcriber may be influenced by what he considers as being the norm, and standardize the transcription of prosodic phenomena, ignoring what he sees and what he hears.

A further advantage of an automatic processing is that, once the values of the parameters are normalized, they are then always compared to the same threshold values. This process is extremely difficult to follow when human (hence subjective) annotation is concerned.

The goal of the present study is to test an automatic approach for prosodic labeling in a cross-linguistic study of speech prosody in French and English. We use an automatic system, PROSOTRAN, in this study. This program is well adapted for annotation of languages, such as French, in which the syllable duration is one of the major parameters of stress. PROSOTRAN is able to annotate the prosody of sentences in French and English containing the same syntactic structures.

2 Prosodic annotation

Prosodic parameters are subject to a prosodic coherence governing parameter values across the prosodic group. It was observed in automatic speech synthesis (in diphone and data driven approaches) that a sudden unjustified change in F_0 or sound duration (beyond stressed syllables or prosodic junctures), is perceived either as a corruption of the speech signal or as an occurrence of a misplaced contrastive stress (Boidin, 2009). Most of the time transcribers focus on the transcription of parameter values of syllables considered as linguistically prominent, carrying pertinent linguistic information. The other syllables, linguistically non-prominent, remain

generally uncoded, although their prosody contributes to an overall perception of a correct pattern. Therefore, in order to keep a faithful prosodic transcription of the speech signal, all syllables should receive annotation of their different parameters. Moreover, some F_0 changes that can be perceptually crucial may not be transcribed in an appropriate way. Thus, a final F_0 rise generally indicates a question, an unfinished clause, or an exclamation, but it can also occur at the end of statements in spontaneous speech. A phonological transcription should avoid using one and the same symbol for these cases (for example, H%), as these types of rises, which may sometimes correspond to the same F_0 contours, are perceptually distinguished (Fónagy and Bérard, 1973).

Prosodic annotation is a complex and difficult task and linguists and scientists working in speech technology address this issue from various angles. A distinction can be made between phonological approaches (Silverman et al., 1992; Hirst, 1998; Delais-Roussarie, 2005; etc.) and acoustic-phonetic prosodic analysis (Beaugendre et al., 1992; Mertens, 2004). Most of the prosodic transcription systems capture levels (*extra high, high, mid, low, extra low*) and movements of the F_0 values (*rising*, *falling*, or *level)*, or integrated F_0 patterns (*Hat pattern*,…).

The prosodic transcription system, **ToBI** (Tone and Break Indices) (Silverman et al., 1992; Beckman et al., 2005), is often considered as a standard for prosodic annotation. However, ToBI appears to be a somewhat hybrid system. It is based on Pierrehumbert's abstract *phonological* description of English prosody (Pierrehumbert, 1980), but is often considered as a *phonetic* transcription, using the perception of the melody for its symbolic coding and the visual observation of the evolution of F_0 values.

INTSINT (an INternational Transcription System) is a production-oriented system. This system is a relatively language independent one; it has been used for the description of F_0 curves in several languages (Hirst and Di Cristo, 1998). A limited number of symbols are used to transcribe relevant prosodic events. These include absolute (*Top, Mid, Bottom*) or relative (*Higher, Lower, Same, Upstepped, Downstepped*) designations. The limitations of the system stem from the use of the F_0 values alone.

Other approaches should be included to complete our short overview of prosodic annotations. The syntactic-pragmatic approach of French intonation integrates a morphological approach, where the intonation is built from sequences of prosodic morphemes, (Focus, Theme, Topic…) (Rossi, 1999). Another interesting approach to prosody is an abstract representation of relational "holistic gestalts", which integrated tonal and temporal whole word profiles, with pitch range variations. This type of system is well adapted to the representation of attitudinal patterns (Aubergé et al., 1997).

3 Cross linguistic study

The use of prosodic parameters is common in all the languages, but some of the uses are language independent. There are universal tendencies (Bolinger, 1978), but

also distinctions in intonational structure between languages ("semantic", "phonotactic", “pragmatic”…) (Ladd, 1996; Crystal, 1969). The comparison of the prosodic parameters among languages is very challenging precisely because of the universality and language specificity of prosody. This is especially true for Germanic (e.g., Dutch, English, German) and Romance languages (e.g., French, Italian, Spanish) (Hirst and Di Cristo, 1998; Ladd, 1996). Therefore, in order to conduct multi-language comparisons, several kinds of prosodic transcription should to be used: an acoustic-phonetic one (broad and narrow), a perceptual transcription for the perceptually relevant events in duration, intensity and melody, a phonological transcription, and a functional transcription.

3.1 French & English prosody

French uses a combination of segmental and tonal cues to signal prosodic phrases, and differs in this respect from a language like English, which relies almost exclusively on tonal boundaries (Gussenhoven, 1984). In French, lexical stress is mostly quantitative (Delattre, 1938), and the final syllable is the one which undergoes a potential lengthening. However, lengthening of the last syllable in a French word corresponds to final (pre-boundary) lengthening, which affects rhythm, and is not an accentual lengthening as in English (Campbell, 1992).

French is generally considered as a language with mostly ‘rising’ F_0 patterns accompanied by a lengthening of final syllables. According to Vaissière (2002), the French ear is trained to perceive rising *continuation* F_0 patterns at the end of prosodic phrases: each prosodic phrase inside a sentence tends to end with a high rise (Delattre’s *continuation majeure*), or a smaller rise (Delattre’s *continuation mineure*). In Delattre’s theory of French intonation, a categorical difference in intonation patterns is expected between minor and major continuation patterns, which are syntax-dependent. Furthermore, according to Delattre, major continuation patterns are only rising, whereas minor continuations can show rising or falling patterns. Prominence is not lexically driven in French (i.e., there is no lexical stress), but it is determined by prosodic phrasing (Delais-Roussarie, 2000).

3.1.1 F_0 contours. French and English intonations are sometimes described by a set of contours. Delattre (1966) identified 10 basic contours that can describe the most frequent intonation patterns in French. Post (2000) also listed 10 contours although these contours differ from those proposed by Delattre. As far as English is concerned, 22 pertinent intonation contours are proposed by Pierrehumbert (1980) to describe English intonation.

It is common to use the term *assertion intonation* or *question intonation* to refer to falling or rising contours. Falling contours are associated with assertion or assertiveness (Bartels, 1999), whereas rising contours are associated with questions or aspects of questioning (uncertainty, ignorance, call for a response or feedback from the addressee, etc.). Although prototypical assertions are uttered with a falling contour and that prototypical confirmation or verifying questions are uttered with a rising contour, occurrences of assertions with a rising contour and occurrences of

confirmation or verifying questions with a falling contour are far from rare in everyday conversations (Beyssade et al, 2003).

In the following paragraphs, F_0 contours in French and English sentences are measured and compared.Their difference was statistically evaluated.

3.2 Corpus

The corpus used in this study was recorded as a part of project Intonal, which focuses on intonation in French and English. The project was conducted by the University of Nancy2 and the LORIA research laboratory (2009-2012). The recorded corpus contains 40 short sentences belonging to 8 syntactic categories which were recorded by 20 French and 20 English native speakers. In a previous study, two prosodic parameters associated with F_0 slope were calculated for some target words in sentences. These words are bolded and underlined in the following sentences:

- (CAP). Continuative configuration at the end of the first clause in a two clause sentence, without any coordinating conjunction: “Il dort chez Maria, il va finir tard. / He'll sleep at Maria's, he'll finish late.”
- (CAO). Continuative configuration at the end of the first clause in a two clause sentence, with a coordinating conjunction: “Il dort chez Maria car il finit tard. / He'll sleep at Maria's because it’s too late.”
- (CIS). Continuative configuration on a subject NP: “Les agneaux ont vu leur mère. / The lambs have seen their mother.”
- (CIA). Continuative configuration on a NP subject in the first clause of a two clause sentence:
  “Nos amis aiment Nancy parce que c’est joli. / Our friends really like Nancy because it’s pretty.”
- (QAS). Question configuration at the end of a clause: “Il dort chez Maria? / Will he sleep at Maria’s?”
- (QIS). Interrogative configuration on a simple subject NP: “Qui a appelé? Nos amis? / Who has phoned? Our friends?”
- (DIS). Short declarative sentence “Nos amis. / Our friends”.
- (DAS). Longer declarative sentence: “Il dort chez Maria. / He’ll sleep at Maria’s”.

Two kinds of non-conclusive F_0 slope configurations were studied here at two levels. First, on the syntactic level: the slope of the final segment of a subject NP in a declarative sentence, followed (CIA) or not (CIS) by another sentence. Second, on the discourse level: the slope of the final segment of A in a two clause utterance AB, where A and B are declarative clauses connected by a discourse relation, marked (CAO) or not (CAP) by a conjunction.

These sentences were used to investigate whether the intonation of the target words is realized in a similar manner in both English and French and whether:

- there is a significant difference between major continuation curves (expected in CAO and CAP sentences) and minor continuation curves (expected in CIA and CIS sentences).

- continuative rising slopes (expected in sentences CAO, CAP, CIA & CIS) are different from interrogative slopes (measured in QIS & QAS) sentences
- continuative falling slopes (measured in CIA and CIS types of sentences) are different from declarative slopes (measured on declarative sentences DIS & DAS).

3.3 Segmentation and annotation of the speech signal

In order to segment our speech data, knowing the orthographic transcriptions, a text-to-speech forced alignment was carried out using the CMU sphinx speech recognition toolkit (Mesbahi et al., 2011). This provided an automatic segmentation of the speech signal at the phoneme level. The automatic segmentation of each speech signal was then manually checked by an expert phonetician using signal editing software. Intonation slopes were computed as regression slopes (RslopeST) using F_0 values in semitones, which were estimated every 10 ms. Slopes were calculated on the last two syllables of the target segments (in underlined bold characters in 3.2) of every sentence.

3.1.1 Statistical analysis. F_0 slope data are analyzed by fitting mixed linear regression models (R package *lme4*). Using this approach, one can contrast the different configuration types and show the differences that are significant and those that are not (function *glht*, package *multcomp*).

The statistical analysis showed that in French, sentences where we expect minor F_0 patterns, continuation patterns (CIA-CIS sentence types) are mostly rising (95%). The major continuation sentence types (CAP-CAO) also have rising F_0 slopes (59%); but there is a significant difference between sentences with coordinating conjunctions (CAO), containing 73% of rising F_0 slopes, and paratactic (CAP) sentences containing only 46% of rising F_0 slopes.

In the English data, the F_0 slopes measured in minor continuation (CIA-CIS) sentence types can rise (53%) and fall (47%) equally. In major continuation (CAP-CAO) sentence types, F_0 slopes are seldom rising (21%) and there is no marked difference between F_0 slopes in sentences with coordinating conjunctions (CAO, 18% of rising patterns) and F_0 slopes in paratactic sentences (CAP, 24 % of rising patterns).

In the French corpus, slopes measured on minor continuation (CIS-CIA) sentence types are not significantly different from juxtaposed sentence types where major continuation slopes (CAO) are expected, although they are significantly different from slopes measured on sentences with coordinating conjunctions (CAP) [see Figure 1 (left)]. Neither is there a significant difference between slopes measured on these two sentence types (CIA-CIS) (where minor continuation slopes are expected). However, the slopes of the latter are significantly higher than the slopes measured on short declarative sentences (DIS) and significantly lower than the slopes measured on simple subject NP questions (QIS). On the other hand, slopes measured on juxtaposed sentences (CAP) are significantly lower than those measured on sentences with a coordinating conjunction (CAO).

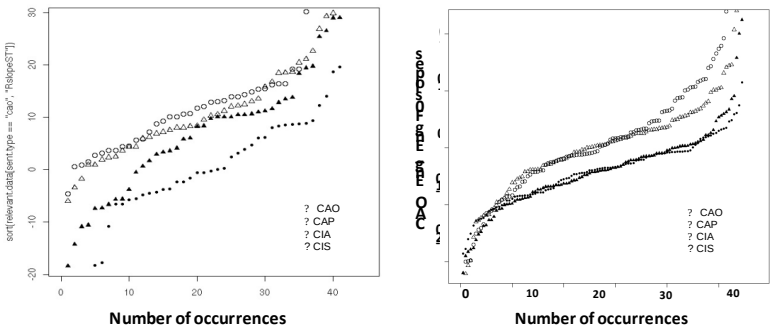

Figure 1. F_0 slope values for the French (left) and English (right) corpora in 4 sentence types. The *Y axis* corresponds to RslopeST value (RslopeST = slope of the regression line of the pitch data points in semitones) and the *X axis* to increasing ordering of observations (each point is an observation).

In the data recorded by English speakers, slopes of minor continuation sentence types (CIA-CIS) are significantly higher than slopes measured on major continuation sentence types (CAO-CAP) and are also significantly higher than slopes measured on short declarative sentences (DIS). However, no significant difference was found between minor continuation slopes (CI) and slopes measured on short questions (QIS). English speakers do not utter juxtaposed sentences (CAP) differently than sentences containing coordinating conjunctions (CAO) (see Figure 1 (right)). Furthermore, major continuation slopes (CAP-CAO) are not significantly different from slopes measured on longer declarative sentences (DAS) and interrogative (QAS) sentences (Bartkova et al., 2012).

3.4 Additional analyses using automatic annotations

As it appears from the previous analysis of the obtained results, the syntactic differences among the sentences studied are not necessarily marked, as expected by theory (Delattre, 1966) or by prosodic means, and there are not systematic and significant differences among the rising and falling F_0 slopes used. However, pertinent prosodic differences among these syntactic structures can be scattered all along the utterances and they are not necessarily concentrated on the final syllables of the target words alone. In order to compare the different syntactic structures and their prosody in a more precise way, and to conduct a deeper cross linguistic comparison of the prosody among French and English sentences, a subset of the data was annotated by our PROSOTRAN automatic annotation tool and the results of the obtained annotations were analyzed and discussed in the paragraph below. The corpus used was comprised of one sentence for each sentence type uttered by about 10 French speakers (as not all the speakers uttered all the sentences) and about 20 English speakers (all speakers uttered all sentences).

3.4.1 Speech data processing. The speech data processing used in this part of our study had 4 different stages. During the first stage, prosodic parameters are extracted from the speech signal. In the second stage, prosodic annotations are yielded by our

annotation tool PROSOTRAN using the extracted parameters and these parameters are hand checked by phoneme segmentation, as in our previous speech data processing (see 3.3). In order to check whether our annotation is faithful or not, the third processing stage recalculates the numerical F_0 values from the prosodic annotation and during stage four, the prosody of the speech signal is resynthesized using Praat (and the PSOLA technique). The resynthesis of the melody allows for checking whether or not the quality of the obtained signal was corrupted by the previous prosodic parameter manipulations.

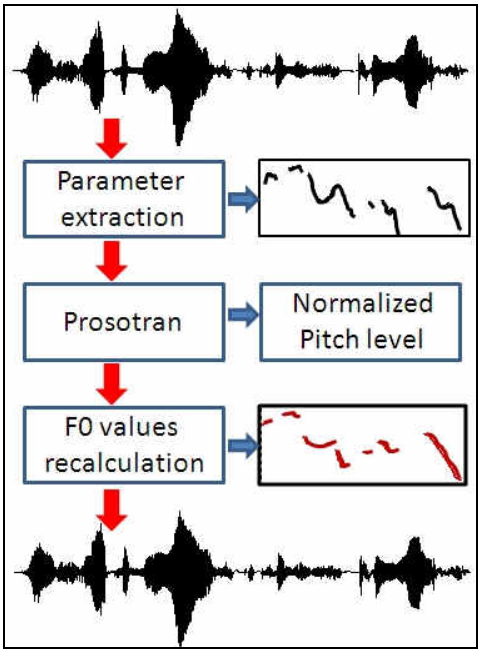

Figure 2. Illustration of the 4 stages of our prosodic processing: (1) parameter extraction, (2) prosodic labeling with PROSOTRAN, (3) F_0 value recalculation, and (4) resynthesis with the recalculated F0 values.

3.4.2 Parameter extraction. Acoustic parameters, such as F_0 in semi-tones and log energy, are calculated from the speech signal every 10 ms with the Aurora front-end (Speech Processing, 2005). The forced alignment between the speech signal and its phonetic transcription provides phoneme durations, as well as the duration of the pauses. Synchronization between the phoneme units and their acoustic parameters (F_0 and log energy values) is carried out and prosodic parameters are calculated for every relevant phoneme.

3.4.3 PROSOTRAN. Our annotating tool, PROSOTRAN, is a system enabling automatic annotation of prosodic patterns. Since all linguistically relevant prosodic events are realized at the phonetic level by some sort of changes in the prosodic parameters, PROSOTRAN assigns a symbolic label to every syllabic nucleus for each prosodic parameter separately. The resulting annotation is multitiered, with

each tier being associated with a single parameter. PROSOTRAN encodes vowel duration, vowel energy, F_0 slope movement, F_0 level, delta F_0 values and some more information concerning the F_0 curve either symbolically or numerically. However, as for our cross linguistic study, only vowel duration and vowel F_0 levels are used, therefore only the calculation and coding of these parameters are explained in the following paragraphs (for more information about PROSOTRAN, see Bartkova et al., 2012).

3.4.3.1 Duration. Although the temporal axis of the speech signal is represented by all sound durations, PROSOTRAN uses only vowel durations in its prosodic annotation. This avoids the issue of syllabic structure variability, and vowel duration is considered to be more homogeneous and therefore more representative of speech rate variation than syllable duration (Di Cristo, 1985). Moreover, vowel nuclei constitute the salient part of the syllable and are hence the most important speech element used to convey the prosody (Segui, 1984).

In the French corpora processing, each vowel duration was compared to the mean duration and associated standard deviation of the vowels occurring in non-final positions (i.e. not at the end of a word nor before a pause) when measured on the speech data uttered by the same speaker. This way, stressed vowels whose duration is lengthened (vowel duration is one of the major prosodic parameter of French stressed vowel) are discarded from the calculation of the mean and standard deviation values. In the English corpora processing, the vowel durations are compared to the mean duration and standard deviation of all the vowels of all the speech material produced by the same speaker.

To represent sound durations, symbolic annotations are used, representing duration from extra short duration (Voweldur----) to extra long duration (Voweldur++++).

3.4.3.2 F_0 range and levels. In order to represent the speech melody, a melodic range was calculated between the maximum and the minimum values of the F_0 in semi-tones. For each speaker, all speech material was used to build a histogram of the distribution of the F_0 values. To avoid extreme, often wrongly detected F_0 values, 6% of the extreme F_0 values (3% of the highest and 3% of the lowest ones) were discarded. The resulting range was then divided into several zones (9 in our case) and coded into levels (from 1 to 9). F_0 slopes were calculated for vowels and semi-vowels.

Results of the annotation are stored in text files and also in TextGrid files to make possible visualisation by Praat (see Figure 3 for annotation examples).

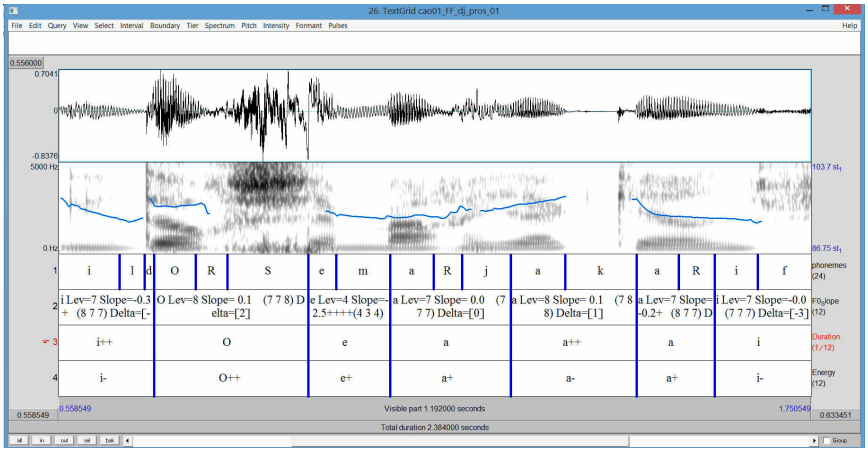

Figure 3. Example of the prosodic labeling provided by the PROSOTRAN tool

3.4.3.3 F_0 level normalization. In order to compare the F_0 patterns of our French and English data, the F_0 level annotation produced by PROSOTRAN was used. However, to minimize the overall range differences among the speakers for a sentence type, F_0 level normalization of the different speakers was carried out. To obtain normalized F_0 level values, the F_0 pattern of one of the speakers was taken as a reference, and all other speaker F_0 patterns were adjusted in order to minimize the Euclidean distance between the individual speaker F_0 pattern and the reference pattern. Normalized F_0 levels were computed for each sentence and for each speaker.

Once the F_0 levels for all vowels were normalized by sentence type, a mean F_0 level value was calculated for each sentence type syllable to yield one representative F_0 level pattern of per sentence type (see Figure 4). Using this single representative F_0 level pattern per sentence enable us to compare the F_0 patterns of the French and the English sentence types and to carry on our cross linguistic study of the prosody.

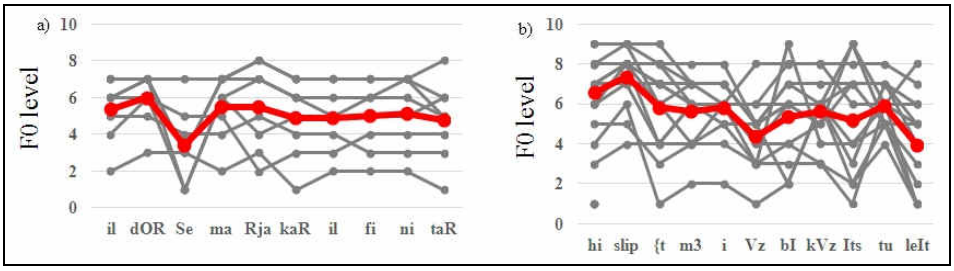

Figure 4. Calculation of a representative F_0 level pattern for a French (a) and an English (b) sentence.

As mentioned before, the duration of each vowel was annotated symbolically. Using these symbolic annotations, a numeric coefficient was calculated expressing the degree of vowel lengthening produced by different speakers. Thus the coefficient value α indicates that the duration of a given vowel is on avrage equal to the mean

duration value plus α times the standard deviation. A low value coefficient indicates that the vowel was largely lengthened by only a few speakers or that the vowel was lengthened slightly by a large number of speakers.

3.5 Result analysis and discussion

The following figures contain the representative F_0 level patterns for the different sentence types. The circles indicate the prominent F_0 levels and the numbers show the vowel lengthening coefficient. Coefficients are indicated only for vowels whose duration was longer than the mean duration and greater than one times the standard deviation.

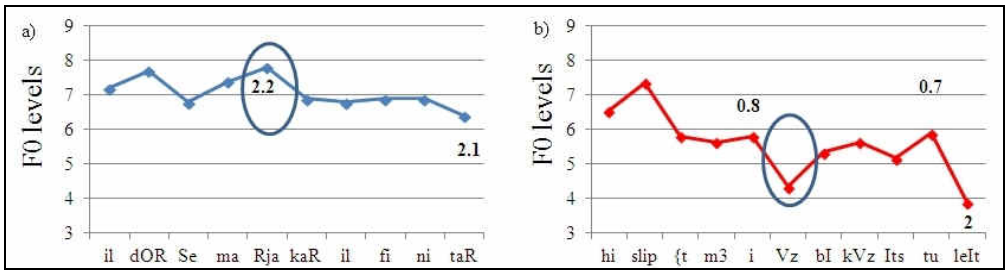

Figure 5. CAO - Continuative configuration at the end of the first clause in a two clause sentence, with a coordinating conjunction: (a) *Il dort chez **Maria** car il finit tard.* (b) *He'll sleep at **Maria's** because it's too late.*

For the continuative sentence types (Figure 5) French speakers marked the continuation with a rising F_0 while English speakers prosodically coded the same syntactic boundary with a lowering F_0. In French, the general rising tendency of the F_0 was not very high but the prosodic boundary also was indicated with a lengthened vowel duration (high duration coefficient). On the other hand, the downwards movement of the F_0 in English was more important but there is no vowel lengthening in the final syllable. The sentence final F_0 movement was falling in the both languages but the slope was steeper in English than in French.

In French paratactic sentences (Figure 6), the mean F_0 level pattern contained a slight F_0 rise on the prosody boundary and the vowel duration was lengthened (even more than in the previous sentence) in the boundary final syllable. French speakers give preference to upward (though moderate) movement of the F_0 on the prosodic boundary, while the majority of the English speakers favor downward movement of the F_0 curve. In French, the inter-utterance prosodic boundary was marked by a lengthening of vowel duration, while in English the utterance final F_0 level was very low and the vowel duration was very clearly lengthened.

In two clause sentences with a continuative configuration (Figure 7), most French and English speakers realized a high level F_0 at the end of the noun phrase subject. But neither French nor English speakers used vowel duration to highlight the prosodic boundary. However, the second prosodic boundary of the sentence, although marked with a lower F_0 level, contained lengthened vowel durations. In English, the final boundary F_0 level was very low (level 3) and the vowel duration

was strongly lengthened. In French, the final prosodic boundary had a relatively high F_0 level (level 7), but the final vowel lengthening was moderate.

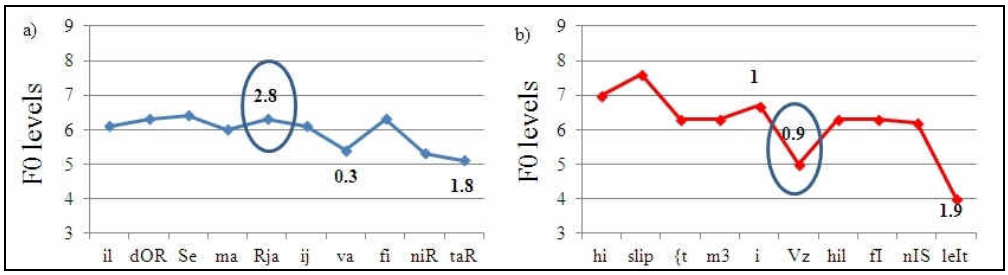

Figure 6. CAP - Continuative configuration at the end of the first clause in a two clause sentence, without any coordinating conjunction: (a) *Il dort chez **Maria**, il va finir tard.* (b) *He'll sleep at **Maria's,** he'll finish late.*

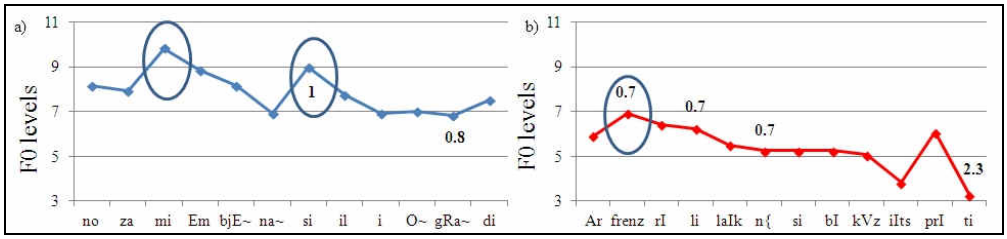

Figure 7. CIA - Continuative configuration on a NP subject in the first clause of a two clause sentence: (a) ***Nos amis** aiment Nancy ils y ont grandi.* (b) ***Our friends** really like Nancy because it's pretty.*

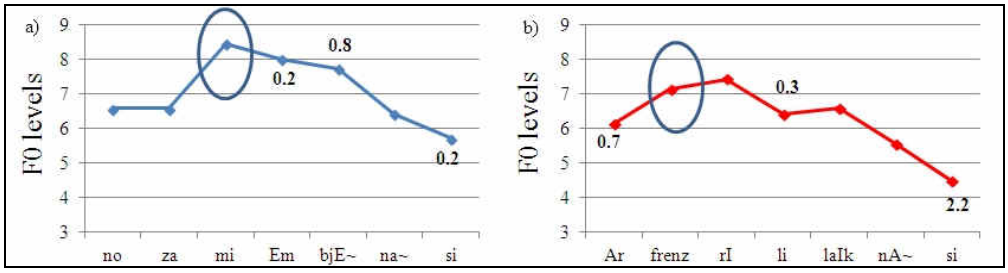

Figure 8. CIS - Continuative configuration on a subject NP: (a) *Nos **amis** aiment bien Nancy.* (b) ***Our friends** really like Nancy.*

In sentences with a continuative configuration on a subject NP (Figure 8), the same phenomena was observed as in the CIA sentences (Figure 7): both speaker groups favored a high F_0 level (corresponding to a rising F_0 curve). This level was again higher in French than in English and no vowel lengthening was used to strengthen the prosodic boundary. The final F_0 level was low in both languages (although lower in English than in French) and the final vowel was significantly lengthened in English, while moderately lengthened in French.

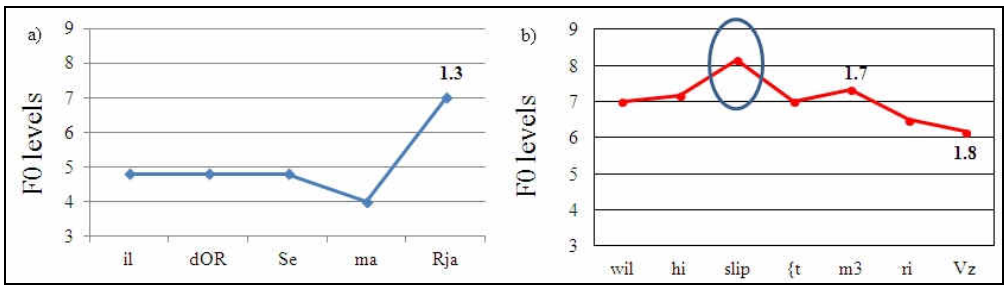

Figure 9. QAS - Question configuration at the end of a clause: (a) *Il dort chez **Maria**?* (b) *Will he sleep at **Maria's**?*

In French, the yes/no question configuration (Figure 9) of F_0 levels is similar to the configuration found in QIS type sentences (Figure 10): a huge level rise preceded by a rather flat F_0 level. The pattern in English sentences contained a lowering of the F_0 level at the end of the sentence as the interrogative character was expressed here by syntactic means (subject-verb inversion); therefore there was no need for prosodic marking.

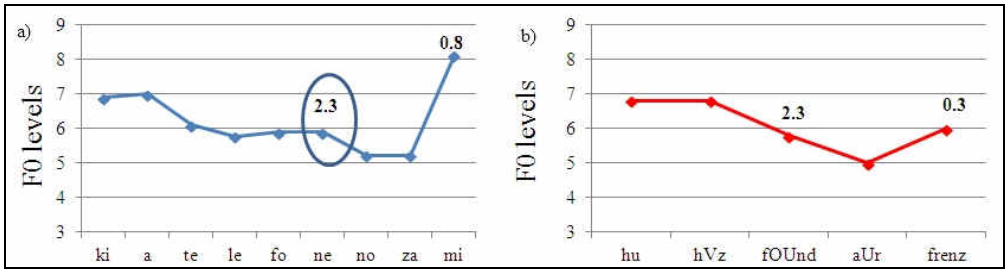

Figure 10. QIS - Interrogative configuration on a simple subject NP: (a) *Qui a téléphoné? **Nos amis**?* (b) *Who has phoned? **Our friends**?*

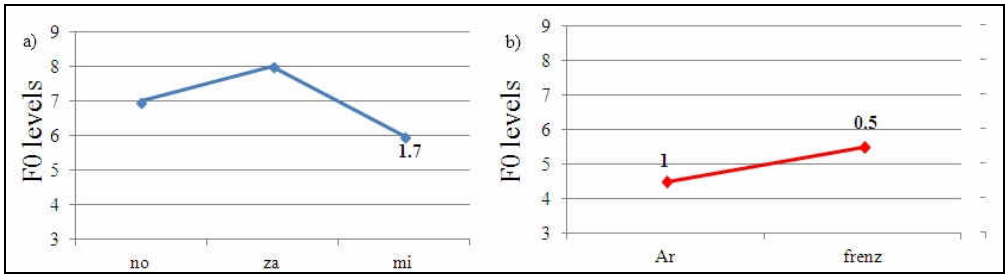

Figure 11. DIS - Short declarative sentence (a) ***Nos amis***. (b) ***Our friends***.

The French and English versions of the previous sentences contained final F_0 rise (high F_0 level), however the level was much higher in French sentences than in English. The first part of the sentence contained a clause containing an interrogative pronoun and its occurrence explained the falling pattern of the F_0 levels. The vowel

duration was used in both sentences to mark the prosody boundary in the first part of the sentence.

The short declarative sentence had a falling F_0 (low F_0 levels) in French pronunciations. However, in the English realization of the sentence, the pattern was slightly rising. In both sentences (French and English), the final vowel duration was also lengthened and used as a boundary marker.

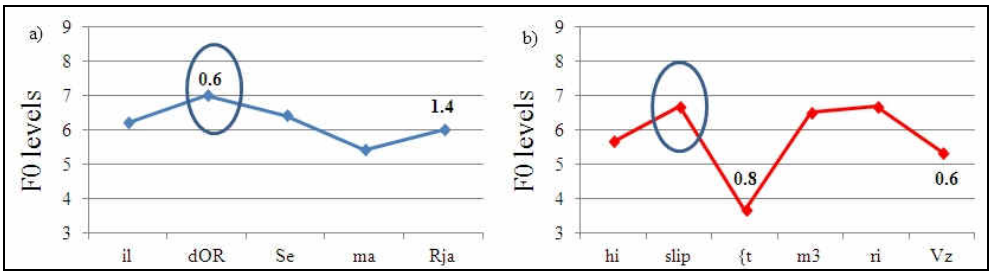

Figure 12. DAS - Longer declarative sentence: (a) *Il dort chez **Maria**.* (b) *He'll sleep at **Maria's**.*

In the longer declarative sentence, the F_0 level of the last vowel was low in English (falling movement) and slightly rising in French. In both cases, the vowel duration was lengthened and marked the prosodic boundary, while the first prosodic boundary was marked by slightly higher F_0 level.

3.6 General discussion

In French, the F_0 level was high at a major prosodic boundary. In fact, the level was higher than in English, especially in yes/no questions. English speakersused falling F_0 patterns to mark major continuation prosodic boundaries and strongly falling patterns to mark the end of declarative sentences. The duration of the last vowel was often lengthened in English and was used to mark the prosodic boundary.

In French declarative sentences, the F_0 range was narrower (1.8 levels on average) than in English (3.5 levels on average). In interrogative sentences, the mean F_0 of the pattern values was 3 in French and 2 in English. In French, the F_0 was more strongly rising on prosodic boundaries than in English. The final F_0 movement in assertive sentences was more moderate in French (falls through 1.2 levels) than in English (falls through 2.1 levels).

The declarative sentences in French were uttered at a higher F_0 level (mean level value 7) than English sentences (mean level value 5.4). The level range used in English sentences was larger (the F_0 on average evolves through 3 levels) than in French sentences, where the mean level range used is 2.

Interrogative sentences in French were uttered at relatively lower range (5 and 6.2) compared to assertive sentences. English speakers used a relatively higher range level for interrogative sentences than assertive sentences (6.9 and 6.2 levels).

The general tendency for French intonation in the phrases studied here is as follows: in French, speakers gave preference to a more flat F_0 (narrower range of F_0

levels used), with mainly upward movement on prosodic boundaries. In English, the range of F_0 levels was broader with mainly downward F_0 movement.

Vowel duration wass used in both languages to indicate prosodic boundaries. In French, a slight F_0 movement on a prosodic boundary was completed by lengthened vowel duration, which indicated the boundary location and its depth. In English, vowel lengthening typically took place at boundaries where the F_0 movement was important. The lengthened vowel duration was used in both languages, however vowel durations were longer on non-final prosodic boundaries in French (mean coefficient value of vowel lengthening 1.8) than in English (mean coefficient value of vowel lengthening 0.8). Moreover, vowel duration was slightly more lengthened in English in sentence final syllables (followed by a pause) than in French. Indeed, in English, the mean vowel lengthening coefficient value was 1.4, while in French its value was 1.2.

3.7 Speech synthesis

In order to verify whether our approach to prosody representation and coding is correct, the F_0 pattern represented as a range of 9 levels was transformed to semitones values and these values were used to synthesize the melody of the sentences in our corpus. According to our preliminary perception tests, made by only 2 expert phoneticians (a French and an English native), all of the resynthesized sentences sounded very natural and there was very little difference between the modified and unmodified sentences. The listening tests were carried out by MOS (*Mean Opinion Score*) tests and the re-synthezided and natural sentences were judged on a 5 point scale (0-very bad, 5-excellent). According to this very preliminary test, the appreciation of naturalness in non-modified sentences was 4.4 out of 5 and the F_0 resynthesized sentences obtained a score of 4.2. Naturally, this very preliminary test will be completed in the future using more listeners in order to verify the validity of our preliminary tests.

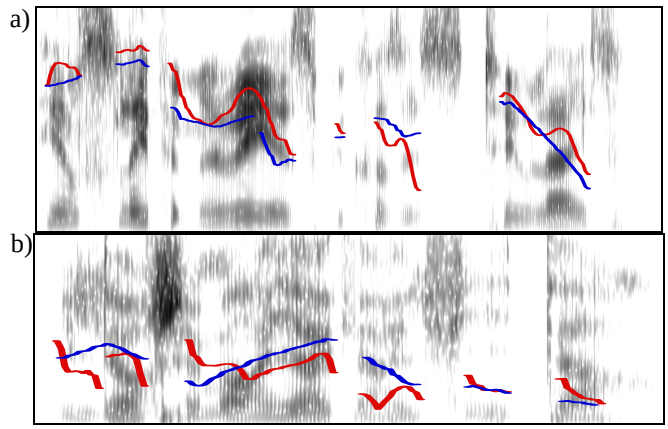

Figure 13. Examples of resynthesis of the melody (a) of an English and (b) of a French sentence. Natural melody curve in red and the synthesized melody curve in blue.

4 Conclusion

The goal of our study is to use an appropriate coding schema for prosody representation in a cross linguistic study of French and English prosody. The data used are laboratory data produced by a group of French and English native speakers and they contain sentences sharing the same syntactic structures in both languages. This syntactic specificity of the data base is well adapted to cross-linguistic study as it allows for comparison of prosodic phenomena relatively easily. However, a methodological problem remains: how to represent prosodic parameters in such a way that comparison would be pertinent.

Two approaches are tested in this study; the first is a general statistical analysis, which compares F_0 slopes measured on the last syllable of some of the words considered as pertinent from a prosodic point of view. This analysis showed that the prosody used in different syntactic structures is not necessarily supportive ofprevious prosodic theory (Delattre, 1966).

The second part of the study was dedicated to a more qualitative comparison of French and English prosody. Two prosodic parameters, vowel duration and F_0 values were coded by an automatic prosodic transcriber (PROSOTRAN), which provided symbolic and numeric annotations for use in our cross-linguistic study. The cross-linguistic comparison of these two parameters highlighted the same basic general differences or similarities on the use of prosody in these two languages. An attempt was also made here to verify how faithful the prosodic coding was by transforming the symbolic values of F_0 levels back to physical parameter values and then reconstructing the prosody of the sentences with F_0 synthesis. The preliminary results are very encouraging but further study is needed in order to get reliable perception test results.

References

Aubergé, V., T. Grepillat, and A. Rilliard 1997, Can we perceive attitudes before the end of sentences? The gating paradigm for prosodic contours. In *EuroSpeech'97.* Rhodes, Grèce, pp. 871-877.

Bartkova, K., A. Bonneau, V. Colotte, and M. Dargnat 2012. Productions of “continuation contours” by French speakers in L1 (French) and L2 (English). In *Proceedings of Speech Prosody.* Shangaï, China, 22-25 May 2012, pp. 426-429.

Bartkova, K., E. Delais-Roussarie, and F. Santiago-Vargas 2012. PROSOTRAN: a tool to annotate prosodically non-standard data, In *Proceedings of Speech Prosody.* Shangaï, China, 22-25 mai 2012, pp. 55-58.

Bartels, C. 1999. *The Intonation of English Statements and Questions.* New-York: Garland Publishing.

Beaugendre, F., Ch. d’Alessandro, A. Lacheret-Dujour and J. Terken 1992. A perceptual study of French intonation. In *ICSLP 92 Proceedings: 1992 International Conference on Spoken Language Processing*. Edmonton, Canada: Priority Printing, pp. 739-742

Beckman, M.E., J. Hirschberg and S. Shattuck-Hufnagel 2005. The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (ed.) *Prosodic Typology: The Phonology of Intonation and Phrasing.* Oxford: University Press, Chapter 2, pp. 9–54.

Beyssade, C., J.-M. Marandin, and A. Rialland 2003. Ground/Focus: a perspective from French. In R. Nunez-Cedeno et al. (eds): *A Romance perspective on language knowledge and use: selected papers of LSRL 2001.* Amsterdam/Philadelphia: Benjamins, pp. 83-98.

Boidin, C. 2009. *Modélisation statistique de l'intonation de la parole expressive*. PhD thesis, Rennes, published by University of Rennes 1.
Bolinger, D. 1978. Intonation across languages, Intonation across languages. In: *Universals of human language 2.* Stanford, Stanford UP, pp. 471-524.
Campbell, W.N. 1992. Syllable-based segmental duration. In G.Bailly, and C. Benoît (eds): *Talking machines: theories, models and design.* Amsterdam: Elsevier, pp. 211-224.
Crystal, D. 1969. *Prosodic systems and intonation in English.* Cambridge, Cambridge UP.
Delattre, P. (1938), "A comparative study of declarative intonation in American English and Spanish", Hispania XLV/2, pp. 233-241.
Delattre, P. 1938. L'accent final en français: accent d'intensité, accent de hauteur, accent de durée. *The French Review* 12(2), pp. 141-145.
Delattre, P. 1966. Les dix intonations de base du français. *The French Review* 40(1), pp. 1-14.
Delais-Roussarie, E. 2005. Interface Phonologie/Syntaxe: des domaines phonologiques à l'organisation de la Grammaire. In J. Durand, N. Nguyen, V. Rey and S. Wauquier-Gravelines (eds): *Phonologie et phonétique: approches actuelles.* Paris: Editions Hermès, pp. 159-183.
Delais-Roussarie, E. 2000. Vers une nouvelle approche de la structure prosodique. In *Langue Française* 126: Paris: Larousse, pp. 92-112.
Di Cristo, A. 1985. *De la microprosodie à l'intonosyntase.* Thesis. Université de Provence.
Di Cristo, A. 1998. Intonation in French. In A. Di Cristo and D. Hirst (eds): *Intonation systems: a survey of twenty languages.* Cambridge, Cambridge UP.
Di Cristo, A. 2010. *A propos des intonations de base du français.* Unpublished MS.
Fónagy, I., and E. Bérard 1973. Questions totales simples et implicatives en français parisien, Interrogation et Intonation. In A. Grundstrom and Léon P. (eds): *Studia Phonetica* 8, Ed. Paris: Didier. pp. 53-98.
Fónagy, I. 1980. L'accent français, accent probabilitaire: dynamique d'un changement prosodique. In I. Fónagy and P. Léon (eds): *L'accent en français contemporain. Studia Phonetica* 15. pp. 123-233.
Gussenhoven, C. 1984. *On the grammar and semantcis of sentence accents.* Dordrecht: Foris.
Hadding-Koch H., and M. Studdert-Kennedy 1964. An experimental study of some intonation contours. *Phonetica* 11, pp. 175-185.
Hirst, D. 1998. Intonation of British English. In D. Hirst and A. Di Cristo (eds): *Intonation Systems: A survey of twenty languages.* Cambridge: Cambridge University Press. pp. 56-77.
Hirst, D. and A. Di Cristo 1998. *Intonation systems: A survey of twenty languages.* Cambridge, Cambridge University Press
Ladd, R. 1996. *Intonational phonology*. Cambridge: Cambridge University Press.
Lacheret-Dujour, A., N. Obin, and M. Avanzi 2010. Design and evaluation of shared prosodic annotation for French spontaneous speech: from expert's lnowledge to non-experts annotations. In *Proceedings of the 4th Linguistic Annotation Workshop.* Uppsala, Sweden, pp. 265-274.
Mertens, P. 2004. The Prosogram: semi-automatic transcription of prosody based on a tonal perception model. In *Proceedings of Speech Prosody 2004*. Nara, Japan, pp. 549-552.
Mesbahi, L., D. Jouvet, A. Bonneau, D. Fohr, I. Illina, and Y. Laprie 2011. Reliability of non-native speech automatic segmentation for prosodic feedback. In *Proceedings Workshop on Speech and Language Technology in Education.* Venice, Italy, pp. 41-44.
Pierrehumbert, J. 1980. *The phonology and phonetics of English intonation.* PhD thesis, MIT. Distributed 1988, Indiana University Linguistics Club.
Post, B. 2000. *Tonal and phrasal structures in French intonation.* The Hague: Holland Academic Graphics.
Rossi, M. 1999. *L'intonation, le système français: description et modélisation.* Paris:

Editions Ophrys.
Segui, J. 1984. The syllable: A basic perceptual unit in speech processing? In H. Boumam and DG Bouwhuis (eds). *Attention and performance Vol.10: Control of language processes.* Hillsdale: Erlbaum. pp. 165-181.
Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg 1992. ToBI: a standard for labeling English prosody. *Proceedings of the Second Int. Conf. on Spoken Languages.* No 2, pp. 867-70.
Speech Processing, Transmission and Quality Aspects (STQ) 2005. Distributed speech recognition; extended advanced front-end feature extraction algorithm; compression Algorithms. *European Telecommunications Standards Institute, European Standards (ETSI ES).* pp. 202-212.
Vaissière, J. 2002. Cross-linguistic prosodic transcription: French vs. English. In N.B. Volskaya, N.D. Svetozarova, and P.A. Skrelin (eds.): *Problems and methods of experimental phonetics. In honour of the 70th anniversary of Pr. LV. Bondarko.* St Petersburg: St Petersburg State University Press. pp. 147-164.

ENGLISH MORPHONOTACTICS: A CORPUS STUDY

Katarzyna Dziubalska-Kołaczyk[1], Paulina Zydorowicz[2], and Michał Jankowski[3]
Adam Mickiewicz University
e-mail: [1]dkasia@ifa.amu.edu.pl, [2]zpaula@wa.amu.edu.pl, [3]mjank@wa.amu.edu.pl

Abstract

This contribution is devoted to the study of English morphonotactics. The term, proposed by Dressler and Dziubalska-Kołaczyk (2006), refers to the interface between phonotactics and morphotactics, and concerns consonant clusters which emerge as a result of morphological intervention. A distinction should be drawn between phonotactic clusters, which are phonologically motivated and occur within a single morpheme, e.g. /ld/ in *cold* and morphonotactic clusters, which may be triggered by concatenative (the case of English, e.g., /ld/ in *called*) and non-concatenative morphology (the case of Polish, cf. Dressler and Dziubalska-Kołaczyk, 2006). The goal of this paper is to investigate phonotactic and morphonotactic clusters occurring in the word-final position in English from the point of view of markedness. We hypothesize that phonotactic clusters tend to be less marked than morphonotactic ones. In this approach, markedness is defined on the basis of three criteria of consonant description; manner and place of articulation (MOA and POA) as well as voice (Lx). The verification of this hypothesis will be conducted within Beats & Binding phonotactics, which operates using the Net Auditory Distance principle (NAD) (Dziubalska-Kołaczyk, 2009). This model formulates universal preferences for optimal clustering, depending on the length of a cluster and its word position.

1 English phonotactics and morphonotactics

The scope of English phonotactics (at least in descriptive terms) is well-known from the works of Trnka (1966) and Gimson (1989). With respect to the word-initial position, English allows for double and triple clusters. The former ones usually consist of an obstruent followed by a sonorant (with the exception of the troublesome /s/ + plosive sequences). Triples are also restricted by phonetic content: the first position in a triple cluster must be filled by /s/; the second element is a fortis stop; the third element is either a liquid or a semivowel. All word-initial clusters in English are intramorphemic, and since they lack the morphological aspect, they will not be studied in the present contribution. Word-finally the following phonotactic possibilities are presented in Trnka (1966):

- final doubles /sp st sk ps ks ft pt kt dz mf mp mz nt nd ns nz nʃ ntʃ ndʒ ŋk (mb nθ nʒ) lf lv lp lb lθ lt ld lk ls lʃ ltʃ ldʒ lm (ln) jt jd js jz jn jl (jf) (jk)/

- final triples /mpt mps ŋkt ŋks lkt lks kst lst jst jnt (nts lts)/.[1]

The clusters presented above are monomorphemic ones. The maximal number of segments in a monomorphemic final cluster is three, and the content of final doubles and triples is much less restricted than that of initial sequences. In the word-final position we can find clusters which are unmarked, as traditionally they are said to have a falling sonority slope, e.g. /lt/ in *melt*, but also marked clusters, which have a rising or flat sonority profile, e.g. /ks/ in *box* or /kt/ in *act*, respectively. Longer clusters always imply the presence of a morphological boundary.

Table 1 below presents the inventory of English inflectional suffixes, which, when added to a stem ending in a consonant, lead to the emergence of morphologically complex clusters.

Table 1. Word-final inflectional suffixes triggering the emergence of morphologically complex clusters in English

Function	**Pronunciation**	**Examples**
plural {s}	/s/ /z/	cats dogs
possesive {s}	/s/ /z/	Kate's John's
3rd person singular {s}	/s/ /z/	walks loves
past simple {ed}	/d/ /t/	loved worked
past participle {ed}	/d/ /t/	loved worked
ordinal {th}	/θ/	sixth

Gimson (1989) provides the following chart of word-final consonant clusters. Tables 2 and 3 below present double and triple clusters, respectively, in the word-final position.

Quadruple clusters are relatively rare and include /mpts/ in *exempts*, /mpst/ in *glimpsed*, /lkts/ in *mulcts*, /lpts/ in *sculpts*, /lfθs/ in *twelfths*, /ksts/ in *texts*, /ksθs/ in *sixths*, and /ntθs/ in *thousandths*.

English also possesses a range of derivational affixes (both prefixes ending with a consonant and suffixes beginning with a consonant) which lead to the creation of word-medial morphologically complex clusters, e.g. /mp/ in *imperfect* or /lpf/ in *helpful*. Word-medial morphologically complex clusters also emerge as a result of compounding, e.g. /ndb/ in *handbag* (when unassimilated and unreduced), /lpr/ in

[1] As Trnka explains, the bracketed clusters are extremely rare. It is also noteworthy that clusters preceded by /j/ cast a shadow of doubt on their actual existence as the palatal approximant is in fact an offglide of the preceding diphthong. Consequently, doubles containing the reported /j/ are singleton consonants, whereas triples beginning with /j/ should be regarded as doubles.

foolproof, /fst/ in *beefsteak*, /tkr/ in *gatecrasher*, or /th/ in *sweetheart*. The shape of clusters present in compounds is rather liberal as far as their phonological make-up is concerned, including the emergence of geminate clusters which are impossible in monomorphemic words, e.g. *midday* vs *better*. Although derivation and compounding generate a wide range of medial clusters, this aspect of phonotactics is beyond the scope of the present contribution.

Table 2. Word-final doubles in English (adapted from Gimson, 1989)

C_1C_2	**p**	**t**	**k**	**b**	**d**	**tʃ**	**dʒ**	**m**	**n**	**f**	**v**	**θ**	**s**	**z**	**ʃ**
p		+										+	+		
t												+	+		
k		+											+		
b					+									+	
d														+	
g					+									+	
tʃ		+													
dʒ					+										
m	+				+					+		+		+	
n		+			+	+	+					+	+	+	
ŋ			+		+									+	
l	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
f		+										+	+		
v					+									+	
θ		+											+		
ð					+									+	
s	+	+	+												
z					+										
ʃ		+													
ʒ					+										

Table 3. Word-final triples in English (adapted from Gimson 1989)

ending in /s/	pts pθs tθs kts mps mfs nts nθs ʜks lps lts lks lfs lθs fts fθs sps sts sks
ending in /z/	ndz lbz ldz lmz lnz lvz
ending in /t/	pst tst kst dst mpt nst ntʃt ŋst ŋkt lst lpt lkt ltʃt spt skt
ending in /d/	ndʒd nzd ldʒd lmd lvd
ending in /θ/	ksθ ntθ ŋkθ lfθ

The area of interaction of phonotactics and morphotactics has been referred to as morphonotactics, which is a sub-branch of morphology (Dressler and Dziubalska-Kołaczyk, 2006). Thus a distinction should be made between phonotactic clusters

(also called *lexical*), which occur within a single morpheme and morphonotactic clusters which are triggered by morphology.[2]

Concatenation often leads to the creation of clusters, which at times converge with already existing lexical sequences. Such is the case of /n l r/ + preterit or past participle /d/ sequences in words such as *fined* vs *find*, *called* vs *bald* and *feared* vs *weird*.[3] The aforementioned sequences are unlikely to indicate the morphological status of a cluster. However, the same morphological operations, e.g. past tense suffixation may lead to the creation of clusters which do not normally occur within roots and whose status is often marked. The marked clusters may be of two types:

(i) clusters whose size exceeds the size of a monomorphemic cluster, i.e. the excessive length of the cluster indicates the probability of a morphological boundary

(ii) clusters which are marked in complexity, that is their phonetic make-up renders them marked.

Clusters may be placed on a continuum from purely morphonotactic to lexical ones. Thus the following groups of clusters have been distinguished (Dressler and Dziubalska-Kołaczyk 2006: 253):

(1) Clusters which occur only across morpheme boundaries. To provide several representative examples, final clusters /fs vz/ occur exclusively at morpheme boundaries due to the addition of plural {-s} as in *cuffs*, *wives*, third person singular {-s} *laughs*, *loves*, and Saxon genitive {-s} *wife's*, *Dave's*. These sequences are extremely marked since they occupy the same position on the sonority scale (as they possess the same manner of articulation). Additional examples of exclusively morphological clusters are /bz gz ðz mz md nz/ as in *pubs*, *eggs*, *clothes*, *names*, *climbed*, *tons*.

(2) Clusters which by default occur at morpheme boundaries, whose monomorphemic opponents are extremely rare (a strong default). The best examples are clusters /ts dz/ as they are almost always morphologically motivated and occur in such words as *cats*, *kids*, etc. and whose monomorphemic congeners are *adze* or relatively uncommon borrowings, such as *quartz*. The group of strong defaults would also include /ps/ despite its occurrence in *lapse* or *apse*. In the majority of cases /ps/ occurs in bimorphemic words, e.g. *caps* or *keeps*.

(3) Clusters which by default occur at morpheme boundaries, however, there are quite a few morphologically unmotivated examples (a weak default). A rather weak

[2] Dressler and Dziubalska-Kołaczyk (2006) distinguish two sources of morphonotactic clusters: concatenative, present in English, and non-concatenative, absent from English but present, for example, in Polish. Non-concatenative morphology may by illustrated by the rule of vowel ~ zero alternation, e.g. /ln/ in *lnu* 'linen'-GEN.SG. (from *len* 'linen'-NOM.SG.) or zero-Genitive-Plural formation, in which case a medial cluster changes into a final one, and as such is more difficult to pronounce, e.g. /-pstf/ in *głupstw* 'silliness'-GEN.PL. (from *głupstwo* 'silliness'-NOM.SG.) (Dressler and Dziubalska-Kołaczyk, 2006).

[3] /r/ constitutes an element of a cluster in rhotic accents.

default is illustrated by the cluster /ks/, which apart from occurring in morphological clusters, occurs in a handful of Latinate words such as *tax*, *sex*, *six*, *fix*, *mix*, *box*, and *flux*.

(4) Clusters which occur both across morpheme boundaries and within morphemes (the majority are morphologically motivated). An example is /rd/ and /ld/, which may occur both in monomorphemic words, such as *cord* and *cold* respectively, as well as in morphologically complex words, such as *cared* and *called*.

(5) Clusters which occur exclusively within morphemes. This category includes all clusters which do not contain /t d s z θ/ as the final element. Examples of these clusters are abundant, e.g. /ndʒ/ in *orange*, /lf/ in *shelf*, /mp/ in *lamp*, etc.

2 The framework

The theoretical framework for measuring cluster markedness is that of Beats-and-Binding phonotactics (cf., Dziubalska-Kołaczyk, 2002, 2009, in press). This theory specifies phonotactic preferences as well as a way to evaluate clusters within these preferences. The rationale behind this model of phonotactics is to counteract the preference for CV. Since CV is a preferred phonological structure and consonant clusters tend to be avoided across languages and in performance, there must be a phonological means to let them function in the lexicon relatively naturally. This is achieved by auditory contrast and its proper distribution across the word. It is believed that auditory (perceptual) distance can be expressed by specific combinations of articulatory features which eventually produce the auditory effect.

Any cluster in a structure which is more complex than CV is susceptible to a change resulting in CV, e.g. via cluster reduction (consonant deletion) CCV→CV or vowel epenthesis CCV→CVCV or at least vowel prothesis CCV→VCCV. Ways to counteract this tendency include increasing the perceptual distance between the consonants (CC of the CCV) and counterbalancing the distance between the C and the V (CV of the CCV). This distance will be expressed by Net Auditory Distance (NAD). Nevertheless, cluster size remains an obvious measure of cluster complexity: longer clusters are unanimously more complex than the shorter ones.

NAD is a measure of the distance between two neighbouring elements of a cluster in terms of differences in MOA (manner of articulation) and POA (place of articulation). A general NAD table includes MOAs and POAs, in which manners refer to the most generally acknowledged version of the so-called sonority scale, while places are taken from Ladefoged (2006: 258). For particular languages, more detailed tables can be devised, reflecting the differences between systems as well as including more detailed MOA and POA scales, as in the table for English (see Table 4). Tentatively[4], also voice value was included in the calculation (designated as Lx,

[4] Although the difference in voicing (Lx) has been considered, laryngeal features are non-redundant within subclasses of sounds only (e.g., they are non-redundant within obstruents and largely redundant within sonorants) and as such will have to be included in more refined,

with the values 0 for voiceless and 1 for voiced). The numbers in the table are arbitrary. The numbers for the MOAs are based on the sonority scale which assumes equal 'distances' between members starting with STOP through VOWEL. These are expressed by the distance of 1. Affricates and liquids receive special treatment due to their phonetic characteristics. Similarly, the numbers for POAs arbitrarily reflect the distances between sounds. Again, the judgments refer to their phonetic characteristics.[5]

Table 4. Distances in MOA and POA: English

OBSTRUENT			**SONORANT**						
STOP		**FRICATIVE**	**NASAL**	**LIQUID**		**GLIDE**	**VOWEL**		
	AFFRICATE			**lateral**	**rhotic**				
5.0	**4.5**	**4.0**	**3.0**	**2.5**	**2.0**	**1.0**	**0**		
p b			m			w	1.0	bilabial	LABIAL
		f v					1.5	labio-dental	
		θ ð					2.0	inter-dental	CORONAL
t d		s z	n	l			2.3	alveolar	
	ʧ ʤ	ʃ ʒ			ɹ		2.6	post-alveolar	
						j	3.0	palatal	DORSAL
k g			ŋ			w	3.5	velar	
							4.0		RADICAL
ʔ		h					5.0	glottal	GLOTTAL

The preferences concerning final doubles and triples are formulated below.

Double finals:

NAD (V, C_1) ≤ NAD (C_1,C_2)

The condition reads:

In word-final double clusters, the net auditory distance (NAD) between the two consonants should be greater than or equal to the NAD between a vowel and a consonant neighbouring on it.

Triple finals:

class-specific calculations in future research. In fact, we want to propose, rather than the Lx criterion, the S/O criterion, i.e. the difference between sonorant and obstruent, be set as 1 (see below for the discussion of the clusters /nd/ and /nt/). Another possibility would be to consider Basbøll's (in press) 'spread glottis' proposal as a replacement of the feature 'voice'.

[5] We realize that such arbitrariness may be hard to defend. The present values will be modified in two directions: on the one hand, we will set the weights for MOAs and POAs, as has been done by Bertinetto and Calderone (2013), for instance, who propose 1.0 for CV opposition, 0.8 for manners and 0.5 for places and voicing as input values to their probabilistic system. On the other hand, we will rely on more phonetic detail, for example on timing differences between initial and final clusters. All of the modifications indeed aim, as rightly noticed by the reviewer, at setting the values so that the calculations actually derive the predicted degrees of markedness.

NAD (V,C_1) ≤ NAD (C_1,C_2) > NAD (C_2, C_3)

The condition reads:

For word-final triple clusters, the NAD between the first consonant and the second consonant should be greater than or equal to the NAD between this first consonant and the vowel, and greater than the NAD between the second and the third consonant.

The calculation of distances in a number of final double clusters is illustrated below (the values for all the segments are taken from Table 4). The examples show how the scale between strongly marked and strongly unmarked clusters is built.

-VC_1C_2: NAD (V, C_1) ≤ NAD (C_1,C_2)

NAD VC_1 = |MOA V − MOA C_1| + |Lx V − Lx C_1|

NAD C_1C_2 = |MOA C_1 − MOA C_2| + |POA C_1 − POA C_2| + |Lx C_1 − Lx C_2|

-Vlk (as in *milk*)

NAD Vl: |2.5 − 0 | + |1 − 1| = 2.5

NAD lk: |2.5 − 5| + |2.3 − 3.5| + |1 − 0| = |-2.5| + |-1.2| + |1| = 4.7

So, the above preference is observed, since 2.5 < 4.7. This is a *strongly unmarked* cluster.

-Vlt (as in *cult*)

NAD Vl: |2.5 − 0 | + |1 − 1| = 2.5

NAD lt: |2.5 – 5| + |2.3 – 2.3| + |1 – 0| = 3.5

2.5 < 3.5 is true. This is an unmarked cluster.

-Vls (as in *else*)

NAD Vl: |2.5 − 0 | + |1 − 1| = 2.5

NAD ls: |2.5 – 4| + |2.3 – 2.3| + |1 – 0| = 2.5

2.5 = 2.5 is true (*2.5 < 2.5 is not true). This is a borderline unmarked cluster.

-Vkt (as in *act*)

NAD Vk: |0 – 5| + |1 – 0| = 6

NAD kt: |5 – 5| + |3.5 – 2.3| + |0 – 0| = 1.2

*6 < 1.2 is not true. This is a strongly marked cluster.

The above examples illustrate that NAD is a scalar measure which reflects a tendency of a cluster towards an unmarked or marked phonological status. Hence, a cluster may be relatively preferred or dispreferred phonologically. Below we classify clusters dichotomically as either preferred or dispreferred, which is a simplification for the sake of the comparison with morphonotactic clusters.

Phonotactic complexity is thus measured by NAD and cluster size, and responds to a particular position in a word. Even more complexity is created when a need to signal a morphological boundary overrides a phonologically driven phonotactic preference and, consequently, leads to the creation of a marked cluster. Therefore,

one expects relatively marked clusters across morpheme boundaries and relatively unmarked ones within morphemes.

3 Empirical study

3.1 Predictions

The aim of this paper is to investigate English word-final phonotactics and morphonotactics quantitatively. Three hypotheses were formulated. The first hypothesis concerned the relationship between cluster size and the morphological make-up of a cluster. This hypothesis is based on the universal that CV is a preferred syllable structure in the languages of the world. It is predicted that the longer a cluster becomes, the more probable a morphological boundary is. The second hypothesis predicts that the degree of cluster preferability correlates with morphological complexity. It is predicted that morphonotactic clusters will tend to be dispreferred (relatively marked) in terms of NAD, whereas phonotactic clusters will be phonologically preferred (relatively unmarked, natural). This assumption stems from the semiotic precedence/superiority of morphology over phonology: the morphological function may take over the phonological one. In phonotactics, the need to signal a morphological boundary may turn out to be stronger than obeying the phonological preferability of a cluster. Finally, the third hypothesis concerns the relationship between cluster preferability and its frequency in the corpus. It is predicted that the most frequent clusters in the corpus will be preferred in terms of NAD.

3.2 Resources

The resources used in this study were used previously in another project (Dziubalska-Kołaczyk et al., 2012), which also focused on clusters but was contrastive in nature. In the course of our research, we have found that a list of inflectional forms based on a well-established dictionary and a word frequency list based on a large, well-balanced corpus are resources sufficiently reliable for our studies. With regard to the choice of resources for the study of phonotactics, we agree with Vitevitch and Luce (2004: 484), who say that "a common word (or segment, or sequence of segments) will still be relatively common, and a rare word (or segment, or sequence of segments) will still be relatively rare, regardless of the source".

3.2.1 The wordlist

The wordlist used in the present study was based on the CUV2 lexicon compiled by R. Mitton (Mitton, 1992) in the Oxford Advanced Learner's Dictionary of Current English (Hornby, 1974). This volume contains approximately 70.5K items, including inflectional forms along with UK phonemic transcriptions. US transcriptions and an additional 13,8K items were added to the original CUV2 lexicon by W. Sobkowiak for his Phonetic Difficulty Index software (Sobkowiak, 2006). For the present study, this 84,5K lexicon was stripped of proper nouns and duplicate forms, which brought the total number of items down to approximately 66K. The UK transcriptions were analyzed.

3.2.2 Basic terminal (final) cluster statistics

The number of items with word-final clusters was 19,417, which is approximately 30% of the total. The clusters were considered in terms of their length (number of component consonants), which ranged from 2 to 4:

2: bz (*cabs*), ft (*raft*)

3: fts (*crafts*), nts (*students*)

4: ksts (*texts*), ksθs (*sixths*)

and in terms of type (*lexical* or *morphological*):[6]

lexical (LEX): nt (*client*), kst (*next*), etc.

morphological with one boundary (sgl_M): z|d (*used*), ft|s (*thefts*), lpt|s (*sculpts*), etc

morphological with two boundaries (dbl_M): f|θ|s (*fifths*), lf|θ|s (*twelfths*), etc

Some clusters were *lexical* in some words and *morphological* in others, e.g. /nd/ in *wind* and *pinned* or /nz/ in *lens* and *sins*.

The assignment of morphological boundaries was performed automatically or semi-automatically based on the set of inflection clues presented in Table 1. In some cases, for clusters such as /nz/ for example, the grammar codes presented in the COCA resource (see section 3.2.3 below) were used to automatically identify and categorize "lexical cluster" entries, such as *lens* (grammar code **nn1**) and "morphological cluster" entries, such as *sins* (grammar code **nn2** or **vvz**). In other cases, orthography was a sufficient clue, as in the case of /ft/, for example: *left*, *rift*, *soft*, etc (lexical) and *stuffed*, *sniffed*, *puffed*, etc (morphological).

Morphological boundaries were marked with a vertical bar sign (|), e.g. d|z (as in *woods*), lf|θ|s (as in *twelfths*).

3.2.3 The corpus

Frequency data for the items studied were extracted from a frequency list based on the 410 million word Corpus of Contemporary American English (COCA) (Davies 2011). In other words, the corpus was used solely as a source of word frequency information. This list contains approximately 500,000 word forms, along with their grammar codes[7], number of occurrences, and number of sources in which they appear.

[6] It is important to clarify that only clusters generated by productive morphology were classified as *morphonotactic*. Irregular past tense and past participle forms, such as *meant*, *felt*, *slept* as well as suppletive forms, e.g. *went* were treated as lexicalised ones and counted together with phonotactic clusters.

[7] To assign grammar codes to words of the COCA corpus, its creators used the CLAWS part-of-speech tagger (http://ucrel.lancs.ac.uk/claws/, (Davies, 2011)).

3.3 Results

Figures 1-7 present the results for the three hypotheses formulated in section 3.1 above, whereas Tables 5-11 show detailed quantitative data.[8]

Figures 1-3 present the results for Hypothesis 1. The data confirm the prediction that the probability of a morphological boundary increases along with cluster length. This holds true for cluster types in the wordlist (note the threefold division: lexical, morphonotactic and mixed cluster types), the number of unique words in the paradigm, as well as the tokens in the corpus.

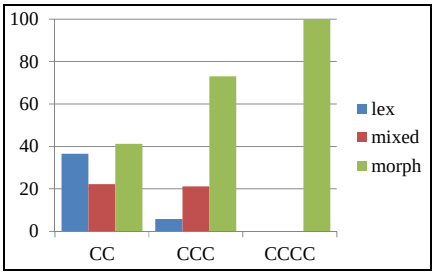

Figure 1. Cluster size vs morphology (cluster types in the wordlist, %)

Table 5. Cluster size vs morphology (cluster types in the wordlist)

	CC	CCC	CCCC	total
lex	23	3	0	26
mixed	14	11	0	25
morph	26	38	8	72
total	63	52	8	123

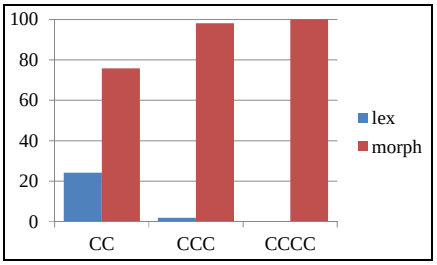

Figure 2. Cluster size vs morphology (unique words from the wordlist)

[8] The notation in the graphs and tables should be read as follows:
lex = lexical clusters
morph = morphonotactic clusters
mixed = cluster types which may have a morphologically simple or complex realisation
P = preferred clusters
D = dispreferred clusters
CC, CCC, CCCCC = double, triple, and quadruple clusters, respectively

Table 6. Cluster size vs morphology (unique words from the wordlist)

	CC	CCC	CCCC	total
lex	4,136	45	0	4,181
morph	12,928	2,292	16	15,236
total	17,064	2,337	16	19,417

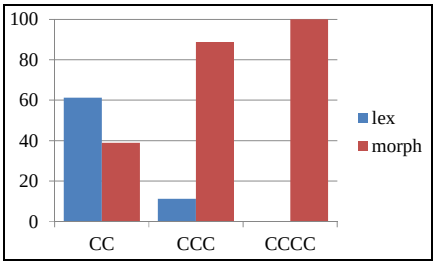

Figure 3. Cluster size vs morphology (corpus frequency, %)

Table 7. Cluster size vs morphology (corpus frequency)

	CC	CCC	CCCC	total
lex	3,2226,546	563,789	0	32,790,335
morph	2,0449,811	4,474,933	38,278	24,963,022
total	5,2676,357	5,038,722	38,278	57,753,357

Figuress 4-6 present the results for Hypothesis 2 which tested the relationship between cluster preferability and the morphonotactic status of a cluster. This hypothesis was partially confirmed: morphonotactic clusters are indeed dispreferred in terms of NAD (the proportion was 25 to 2); however, dispreferred clusters outnumbered the preferred ones among the lexical clusters (the proportion was 25 to 9).

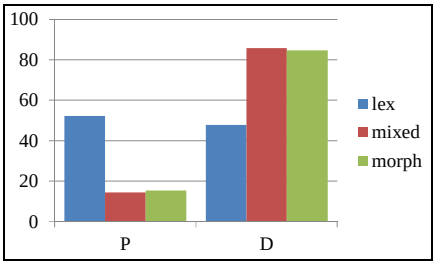

Figure 4. Cluster preferability vs morphology (cluster types from the wordlist, %)

Table 8. Cluster preferability vs morphology (cluster types from the wordlist)

	lex	mixed	morph	total
P	12	2	4	18
D	11	12	22	45
total	23	14	26	63

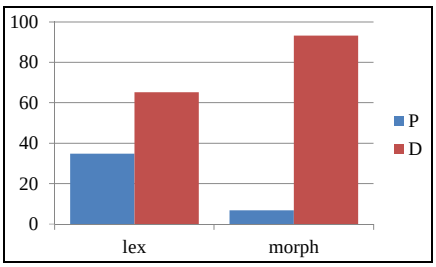

Figure 5. Cluster preferability vs morphology (unique words from the wordlist, %)

Table 9. Cluster preferability vs morphology (unique words from the wordlist)

	lex	morph	total
P	1,441	876	2,317
D	2,695	12,052	14,747
total	4,136	12,928	17,064

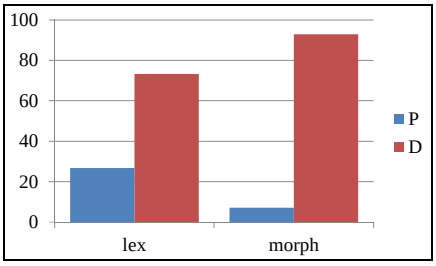

Figure 6. Cluster preferability vs morphology (corpus frequency, %)

Table 10. Cluster preferability vs morphology (corpus frequency)

	lex	morph	total
P	8,596,785	1,448,074	10,044,859
D	23,629,761	18,998,597	42,628,358
total	32,226,546	2,0446,671	52,673,217

In order to test hypothesis 3, we selected the five most frequent final double consonant clusters from the corpus, which included /nd st nt nz ts/. Out of these, only /nt/ was a preferred sequence according to NAD and, as Figure 7 demonstrates, it was an exclusively phonotactic sequence. The remaining 4 clusters were dispreferred according to NAD and all of them could occur across morpheme boundaries. /nz/ and /ts/ were heavily morphonotactic clusters; whereas /nd/ and /st/ could take either realisation (though in the corpus, they tended to occur intramorphemically).

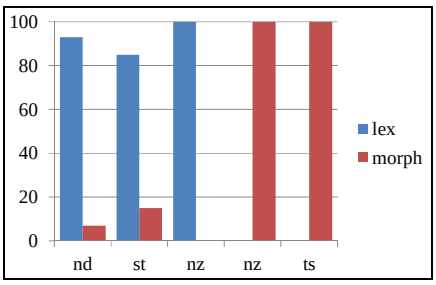

Figure 7. Cluster preferability vs corpus frequency (%)

As mentioned above, however, NAD is a scalar measure. /Vnt/ is borderline unmarked (3 = 3 is true) and /Vnd/ is weakly marked (*3 < 2 is not true). Also, as already noticed, voicing (Lx) is a disputable criterion. If we excluded it, and included the Sonorant vs. Obstruent difference instead (with the value of 1), then both /Vnt/ and /Vnd/ were borderline unmarked (both are sequences of Sonorant-Sonorant-Obstruent). In the future research, we aim to verify the criteria which account for phonotactics in the most optimal way[9].

/nz/ and /ts/ are both morphonotactic and marked. This is what we expected. As for /st/, it is well known that *initial* s+stop clusters are notoriously difficult to classify across models of phonotactic markedness. Sonority-based models generally admit *final* s+stops, since they include a slight sonority slope. NAD phonotactics disqualifies final s+stop clusters slightly less than initial ones: in both positions, these are strongly marked clusters. Thus, the NAD principle on its own cannot explain their occurrence. s+stops belong to the class of the so-called plateau clusters. These are generally problematic for phonotactic models. Baroni (in press), for instance, discusses the role of acoustic salience in the creation of such structures.

It is interesting to compare a cluster ranking based on type frequency (wordlist) with that based on token frequency (corpus). When a list of inflectional forms is considered, the number of individual word forms with a given cluster is the cluster's "frequency". On the other hand, when corpus data are considered, the cluster's "frequency" is the total number of occurrences of all the words with a given cluster in the corpus. As expected, the ranking of some clusters is influenced by the frequency of individual word forms which happen to be particularly frequent in texts, in which reflect actual use. As can be seen from Tables 12 and 13, the place of the /nd/ cluster in the token ranking is boosted by high frequency words, such as *and* (19.50% of all the wordforms considered!), *around* (0.50%), *found* (0.33%), etc. As a result, the cluster /nd/ moves from seventh place to first, and clusters /nz/ and /ts/ move from the top of the type ranking to fourth and fifth place respectively in the token ranking. It is also important to note that the words with the /nd/ cluster

[9] Among others, POA differentiation for vowels needs to be introduced. This will allow for more precise NAD calculations at the edges of vowels, which would be sensitive to vowel colour (palatal, labial, velar).

represent as many as 26.32% of the total number of the words studied and that the words with the top ten clusters represent as many as 70% of the total number of the words studied.

Table 12. Cluster ranking based on type (wordlist) frequency

1	nz	*means*
2	ts	*its*
3	st	*just*
4	lz	*schools*
5	nt	*want*
6	ks	*makes*
7	nd	*and*
8	dz	*kids*
9	ld	*world*
10	nts	*students*

Table 13. Cluster ranking based on token (corpus) frequency

		top frequency examples		**type rank**
1	nd	*and, around, found, find, end*	**26.33%**	**7**
2	st	*just, first, most, last, against*	**9.43%**	**3**
3	nt	*want, percent, president, student*	**8.53%**	**5**
4	nz	*means, questions, ones, plans*	**5.28%**	**1**
5	ts	*its, states, minutes, nights*	**4.99%**	**2**
6	ld	*world, old, told*	**3.71%**	**9**
7	ns	*since, once, sense*	**3.46%**	**–**
8	lz	*schools, officials, miles, calls*	**3.33%**	**4**
9	ks	*makes, six, looks, weeks*	**2.92%**	**6**
10	kt	*fact, looked, worked, effect*	**2.87%**	**–**
			70.85%	

4 Conclusion

The purpose of this contribution was to provide a quantitative and qualitative analysis of word-final consonant clusters in English with respect to the lexical and/or morphonotactic status of clusters, as well as their markedness. The results confirmed the prediction that the probability of a morphological boundary increases along with cluster length. Hypothesis 2, concerning the relationship between cluster structure (morphologically simple or complex) and markedness, found its partial confirmation in the data: morphonotactic clusters are indeed marked in terms of NAD. A fairly high percentage of marked clusters among the lexical ones, also visible in the results for Hypothesis 3, calls for further refinement of the measurement criteria for NAD. We have proposed to exclude Lx (contrast in voicing) and replace it with SO (Sonorant vs. Obstruent, contrast in obstruction). Another point to consider would be the difference in timing constraints in clusters depending on a position: initial clusters are much more rigid than the final ones.

We hope these data demonstrate that phonotactics need to be studied from both phonological and morphological perspective, as well as that the phonological perspective needs phonetic grounding. Importantly, we have shown the pivotal role of data source for frequency calculations.

5 Acknowledgements

This study is a part of a project Phonotactics and morphonotactics of Polish and English: description, tools and applications funded by National Science Centre (grant number: N N104 382540).

References

Baroni, A. (in press). On the importance of being noticed: The role of acoustic salience in phonotactics and casual speech. *Language Sciences*.

Basbøll, H. (in press). Syllable–word interaction: Sonority as a foundation of (mor)phonotactics. *Language Sciences*.

Bertinetto, P.M., and B. Calderone 2013. From Phonotactics to Syllables. A psycho-computational approach. [a talk delivered during the 46th *Societas Linguistica Europaea*].

Davies, M. 2011. Word frequency data from the Corpus of Contemporary American English (COCA). Downloaded from http://www.wordfrequency.info on January 24, 2011.

Dressler, W., and K. Dziubalska-Kołaczyk. 2006. Proposing Morphonotactics. *Rivista di Linguistica 18(2)* 249-266.

Dziubalska-Kołaczyk, K. 2002. *Beats-and-Binding Phonology*. Frankfurt/Main: Peter Lang.

Dziubalska-Kołaczyk, K. 2009. NP extension: B&B phonotactics, *PSiCL* 45(1), 55-71.

Dziubalska-Kołaczyk, K. (in press). Explaining phonotactics using NAD. *Language Sciences*.

Dziubalska-Kołaczyk, K., P. Wierzchoń, M. Jankowski, P. Orzechowska, P. Zydorowicz, and D. Pietrala 2012. *Phonotactics and morphonotactics of Polish and English: description, tools and applications.* [research project].

Gimson, A. C. 1989. *An Introduction to the Pronunciation of English.* (4th edition, revised by S. Ramsaran.) London: Edward Arnold.

Hornby, A.S. 1974. *Oxford Advanced Learner's Dictionary of Current English*, Third Edition. Oxford: Oxford University Press.

Ladefoged, P. 2006. *A course in phonetics*. Boston: Heinle & Heinle.

Mitton, R. 1992. *A description of a computer-usable dictionary file based on the Oxford Advanced Learner's Dictionary of Current English.* A text file bundled with the resource file.

Sobkowiak, W. 2006. PDI revisited: lexical co-occurrence of phonetic difficulty codes. In: W. Sobkowiak, and E. Waniek-Klimczak (eds.) 2006. *Dydaktyka fonetyki języka obcego. Neofilologia VIII. Zeszyty naukowe Państwowej Wyższej Szkoły Zawodowej w Płocku*. Płock: Wydawnictwo Państwowej Wyższej Szkoły Zawodowej. Proceedings of the Fifth Phonetics in FLT Conference, Soczewka, 25-27.4.2005. 225-238.

Trnka, B. 1966. *A phonological analysis of present-day standard English.* Alabama: University of Alabama Press.

Vitevitch, M.S., and P.A. Luce 2004. A web-based interface to calculate phonotactic probability for words and nonwords in English, *Behavior Research Methods, Instruments, & Computers* 36(3), 481-487.

TEMPORAL PATTERNS OF CHILDREN'S SPONTANEOUS SPEECH

Tilda Neuberger
Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
e-mail: neuberger.tilda@nytud.mta.hu

Abstract

The present study investigates the temporal properties of spontaneous speech during various stages of language acquisition. The analysis is synchronic-contrastive, including five age groups: 6-, 7-, 9-, 11-, and 13-year-old Hungarian-speaking, typically developing children. The occurrences, durations and distribution of speech turns, pause-to-pause intervals, silent and filled pauses, as well as the speech tempo were examined using Praat software. Statistical analysis was carried out using SPSS. Results showed that the temporal factors associated with speech, change with the age of the children in several ways: from the age of nine, pause-to-pause intervals lengthen, pauses shorten in spontaneous speech, and speech tempo increases. This finding can be explained by more developed cognitive skills and more established speech patterns, which allow quasi-simultaneous operations of speech planning and execution. Our empirical data supports the development of children's speech performance level from 6-13 years of age.

1 Introduction

During language acquisition, children develop linguistic competence that allows them to produce and comprehend spontaneous speech. In order to construct meaningful utterances, development is necessary in many aspects of language: phonological (e.g., Vihman, 1996), lexical (e.g., Nelson, 1973), morphological (e.g., Brown, 1983), syntactic (e.g., Bloom, 1970) and pragmatic development (e.g., Ninio and Snow, 1996).

Children must solve the segmentation problem for the first time, by dividing fluent speech into strings of discrete words. It is also necessary to recognize groupings of words and utterances in order to discern their syntactic organization (Jusczyk, 1997). The following factors may help the child to recognize the boundaries of speech units: semantic features, syntactic structures, and prosodic (suprasegmental) features, such as pauses, change of pitch, decrease of intensity, lengthening of the last syllable/word before a pause (Klatt, 1975; Lehiste et al., 1976; Streeter, 1978; Frazier et al., 2003; Carlson et al., 2005; Trainor and Adams, 2006).

In spontaneous speech, segmentation is more or less an automatic act, allowing speakers to divide their continuous speech into various units (e.g., paragraphs, sentences) (Lehiste, 1979; Kreiman 1982). Speech fluency is affected by various factors, e.g., physiological ones, the thoughts to be transformed, the type of the text, or the speech style (Duez, 1982; Kohler, 1983; Bortfeld et al., 2001). Children's spontaneous speech differs from that of adults, particularly in its complexity and fluency (Nippold, 1988). Observations have shown that around the age of three, children become able to produce shorter or longer fluent narratives.

The purpose of this study is to describe some temporal patterns in the spontaneous speech of children (e.g., frequency of occurrences of pauses, the length of pauses and utterances, and speech tempo). This cross-sectional study provides objective data on temporal organization of speech in children between the ages of 6 to 14 years. The basic assumption of the investigators is that the temporal organization of speech may provide insight into the covert processes of speech planning and execution. Thus, language development will be reflected in changes of temporal organization.

It is known that one of the most effective cues for boundary detection in speech production is pause. Silent pause is an interval of silence in the acoustic signal, i.e., a segment with no significant amplitude (Zellner, 1994). Silent pauses might have more than one function in speech: these pauses are present for physiological reasons (e.g., respiratory or intersegment pause), for intentionally marking major semantic and/or syntactic boundaries, for making the listener's job easier by aiding them to segment speech, or to give individuals ample time to parse the speech signal (Strangert, 2003; Harley, 2008). Previous research conducted on adults' narratives and conversations has indicated that silent pauses are more likely to occur at boundaries of coherent units rather than within units (Brotherton, 1979; Gee and Grosjean, 1984; Rosenfield, 1987; Grosz and Hirshberg, 1992).

A filled pause is a gap in the flow of speech, which is filled with a sound (usually 'uh' or 'um' in English, see Clark–Fox Tree, 2002, schwa or 'mm' in Hungarian, see Horváth, 2010). Filled pause is one of the most frequent disfluencies in spontaneous speech (e.g., Shriberg, 2001). Recently, there has been significant research activity into determining the role and different functions of filled pauses, both in children and adults. Filled pauses provide time for speech planning and self-repair or to mark the speaker's intention to speak (Horváth, 2010). In the early stages of language acquisition, filled pauses play fewer roles than later. Early on, repetitions are produced more often as indicators of uncertainty (DeJoy and Gregory, 1985; Gósy, 2009). Furthermore, both types of pauses allow continuous self-monitoring, and thus contribute to the well-coordinated operations of speech planning and execution. The verbal planning functions of hesitation phenomena were examined in 5-year-old children by MacWhinney and Osser (1977). It was found that filled pauses served three major functions: preplanning of verbalization not yet produced, coplanning of verbalization currently being articulated, and avoidance of superfluous verbalization.

Previous research has demonstrated that language growth and complexity appear to be correlated with disfluency. During the early stages of language development (generally between ages 2 and 3), many children undergo a period of typical disfluency, (Haynes and Hood, 1977; Hall et al., 1993; Ambrose and Yairi, 1999). Kowal et al. (1975a) examined speech disruptions (unfilled and filled pauses, repeats, and parenthetical remarks) in spoken narratives by typically developing children at seven age levels (i.e., from kindergarten to high school seniors). According to their results, younger children tended to produce more silent pauses and longer silent pause durations than older children and that speech rate increased with age. It was suggested that younger children needed more time for planning language production than older age groups. Rispoli and Hadley (2001) argued that as grammatical development proceeded, speech disruptions tended to appear in more-complex sentences, and dysfluent sentences tended to be longer and more complex than fluent ones.

Several studies in the area of children's speech fluency have been primarily motivated by the necessity for interventions for children with language disorders (e.g., Logan and Conture, 1995; Yaruss, 1999; Boscolo et al., 2002; Natke et al., 2006; Guo et al., 2008). The performance of typically developing children was also investigated in order to provide an adequate normative reference. It is determined that language impairment affects the production of fluent speech, for example, children with specific language impairment (SLI) produce more disfluency phenomena, such as hesitations, than their typically developing peers (Hall et al., 1993; Boscolo et al., 2002).

Speech-timing skills have been investigated in children and adults who stutter with the assumption that not only the occurrence of disfluency, but the timing and duration features of fluent speech of stutterers also differ from that of normal speakers (Healey and Adams, 1981; Runyan et al., 1982; Prosek et al., 1982; Winkler and Ramig, 1986). As Winkler and Ramig (1986) revealed, stutterers exhibited more frequent and longer interword pauses than nonstutterers in a story-retelling task. Smith et al. (1996) investigated several temporal characteristics of children's speech longitudinally. They found that individual children might not evidence the same temporal patterns or changes across time than those noted in cross-sectional studies. Singh et al. (2007) examined typically developing (4- to 8-year-old) children's repeated utterances and pointed out a significant reduction in phrase, word and interword pause duration with increasing age. They suggested that the greater lengths of pause duration could be interpreted as evidence for the more complex speech planning required by younger children. They found strong correlations between pause and word duration in the youngest children. This may indicate local planning at word level, whereas these correlations were not noted in the oldest children, who are capable of planning a complete phrase while uttering it.

Compared to adult speech data, less attention has been directed toward Hungarian children's speech fluency and speech-timing skills (e.g., Laczkó, 2009; Vallent, 2010; Menyhárt, 2012; Horváth, 2013). The latter studies mostly have focused on a

certain stage of language acquisition; for instance, Laczkó (2009) and Vallent (2010) investigated the speech of 15- to 18-year-old students. Thus, there is a lack of research-based evidence which enables one to compare the results of children at various ages. This problem can be resolved using cross-sectional analyses.

The aim of this study is to describe some age-specific characteristics associated with the temporal organization of speech. In order to achieve this aim, the rate and duration of pauses, pause-to-pause intervals, as well as the speech tempo of typically developing Hungarian children's spontaneous speech were studied in various phases of language acquisition (between the ages of 6 and 14). The so-called pause-to-pause intervals (also called pause-to-pause region/sections) stretch from one pause to the next (Kajarekar et al., 2003). The main questions of this research were (i) how the pause-to-pause region and pauses are organized in the spontaneous speech of children, and as to (ii) how this organization varies across ages. Three hypotheses were defined: (i) silent and filled pauses would be produced more frequently and with longer duration in younger children's speech; (ii) pause-to-pause intervals would be longer and consist of more words in older children's speech; and (iii) speech tempo would show an increase with age.

2 Subjects, material, and method

2.1 Subjects

Seventy typically developing, Hungarian-speaking monolingual children participated in this project (Table 1). Thirty-three of them were boys and thirty-seven of them were girls (Table 1). Since the number of boys and girls was not equal in all age groups, a statistical analysis was carried out to learn about the possible effect of gender. However, the statistical analysis did not show a significant main effect of gender in any age group, therefore no further analysis was made considering gender. None of the children had any hearing disorders and their intelligence fell within the normal range. The analysis was cross-sectional including five age groups: 7- and 9-year-olds were from lower grades, 11- and 13-year-olds were from the upper grades of elementary school. In addition, 6-year-old preschool children were included in the experiment for further comparisons. There were 14 children in each age group.

Table 1: Age and gender distribution of subjects

Age groups	**Age (year;month)**	**Number of**		
		children	**boys**	**girls**
6-year-olds	6;1–6;11	14	8	6
7-year-olds	7;2–7;7	14	7	7
9-year-olds	9;4–9;10	14	6	8
11-year-olds	11;4–11;10	14	5	9
13-year-olds	13;1–13;9	14	7	7
Total	6;1–13;9	70	33	37

2.2 Material and measurements

Materials consisted of spontaneous speech recordings (digital recordings using a 44.1 kHz sampling rate and a 16-bit resolution). The corpus was recorded in kindergarten and at school in the capital city of the country. Subjects were tested individually. Speaking time was not limited. The total duration of the corpus was 371.2 minutes. The task of the children was to talk about their free-time activities, hobbies and everyday life. Narrative discourse was used; however, not all of the younger children were able to produce sequences of fluent utterances, therefore when they got stuck the interlocutor motivated them by asking a question.

The recordings were labeled using Praat 5.2 software (Boersma and Weenink 2011) in order to analyze the temporal factors of speech. The boundaries of the following units were marked by the author of the present study: pause-to-pause intervals in the children's speech, silent pauses and filled pauses in children's speech, turns in the adult interviewer's speech, and gaps (i.e. silent intervals during turn-taking) (see Figure 1.). While pauses are generally interpreted as within-speaker silences, gaps refer to between-speaker silences (Sacks et al., 1974; Edlund et al., 2009; Lundholm, 2011). Silent intervals longer than or equal to 50 ms were used as acoustic correlates for pausing.

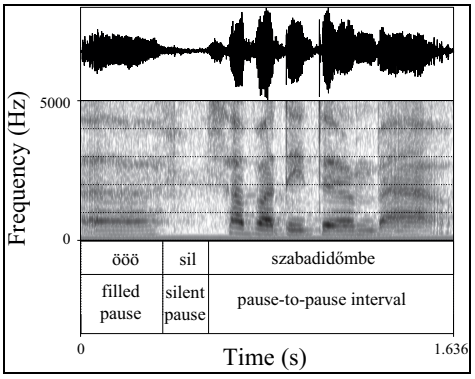

Figure 1. Labeling in Praat

The occurrences, durations and distributions of speech turns, pause-to-pause intervals, silent and filled pauses. In addition the speech tempo was examined. Speech tempo was measured as the rate of words per minute and as the rate of syllables per second. Statistical analysis was conducted using SPSS 13.0 software (Pearson's correlation, One-Way ANOVA, Tukey post hoc test, Kruskal–Wallis test, Mann–Whitney U test). The confidence level was set at the conventional 95%. Besides providing information on the average values for each age group, an emphasis was placed on individual results and outlier values as well.

3 Results

The recordings contain the children's and the interviewer's pause-to-pause intervals, silent and filled pauses, and the turn-taking gaps. The total length of the

recordings (including the speech of interlocutor), as well as the total, average, minimum and maximum durations of the children's speech are presented in Table 2. Mean values for each age group revealed that 7-year-old children produced the shortest speech samples (the average duration was 3.8 minutes in this age group), while the longest speech samples were produced by 9-year-olds (in their case, the average duration was 5.7 minutes). Across age groups, the shortest speech sample was 1.9 minutes long, which was produced by a 6-year-old boy. In contrast, the duration of the longest speech sample was 8.7 minutes, which was produced by a 9-year-old girl. In addition, there was no significant difference between females and males; the average duration of the speech sample in boys was 4.2, while that in girls was 4.7 minutes.

Table 2. Durations of speech material (min)

Age groups	**Total duration of recordings (min)**	**Duration of children's speech samples (min)**		
6-year-olds	75.2	58.6	4.2	1.9–6.9
7-year-olds	64.8	53.2	3.8	2.6–5.9
9-year-olds	98.2	79.4	5.7	3.1–8.7
11-year-olds	67.7	62.5	4.5	2.1–7.2
13-year-olds	65.3	57.2	4.1	2.4–6.6
Total	371.2	311.0	4.4	1.9–8.7

3.1 Speech turns and gaps

We can identify three main parts of the recordings: the speech turns of the children, the speech turns of interlocutor, and the gaps. Turn was defined as a stretch of speech that is not interrupted by the other speaker; a gap is a silent interval during turn-taking.

Children's turns accounted for 83.7% of the entire recording, while the proportion of interlocutor's turns was 8.3%, and the proportion of the gaps was 7.7%. Across age groups, different tendencies were noted (Figure 2). In the case of 6-year-old children, the ratio of the children's turns compared to the total length of the recording was 77.9%, while this ratio was 82% in 7-year-olds, 80.8% in 9-year-olds (the average ratio was 81.4% in lower grade children), 92.3% in 11-year-olds, 87.5% in 13-year-olds (the average ratio was 89.9% in upper grade children). A one-way ANOVA revealed a significant main effect for age group ($F(4, 69) = 3.730$; $p = .009$). By extending the age range (considering three main groups: kindergarten, lower and upper grade children), we noticed a linear increase in the proportion of children's speech. There were significant differences among these three groups according to a one-way ANOVA ($F(2, 69) = 7.220$; $p = .001$). Data of upper grade children differed significantly from that of kindergarten and lower grade children (Tukey post hoc test: $p = .002$ and $p = .030$), but data of the latter two groups did not differ significantly from each other (Tukey post hoc test: $p = .319$).

The ratio of the interlocutor's turns compared to the total length of the recording was 10.7% in kindergarten children, 8.8% in lower grade students, 5.5% in upper grade students. The duration and proportion of the interlocutor's utterances decreased with the age of children, which suggests that older children need fewer motivating questions by the interviewer than younger children. In other words, older children are able to produce longer monologues without the help of the interlocutor. Significant differences were revealed by a one-way ANOVA across the five age groups ($F(4, 69) = 2.857$; $p = .030$) as well as across the three main groups (kindergarten, lower and upper grade children: $F(2, 69) = 4.991$; $p = .010$). The Tukey post hoc test showed significant differences between the data of kindergarten and upper grade children ($p = .008$). Vallent (2010) found that even high school students need motivating questions from the interlocutor during a spontaneous speech task.

The proportion of gaps was the highest in the group of 6-year-old children (11.5% of the total recording duration), it was 9.8% in the lower grade children and 4.6% in the upper grade children. This result reveals a decrease in gap duration with age, which was confirmed by significant between-groups differences according in the one-way ANOVA (considering five age groups: $F(4, 69) = 4{,}261$; $p = .004$) and three main groups: $F(2, 69) = 8.611$; $p < .001$). The ratio of gaps to speech in the oldest children's recordings were significantly lower than that of the two younger groups, as determined by a Tukey post hoc test ($p = .001$ and $p = .008$, respectively).

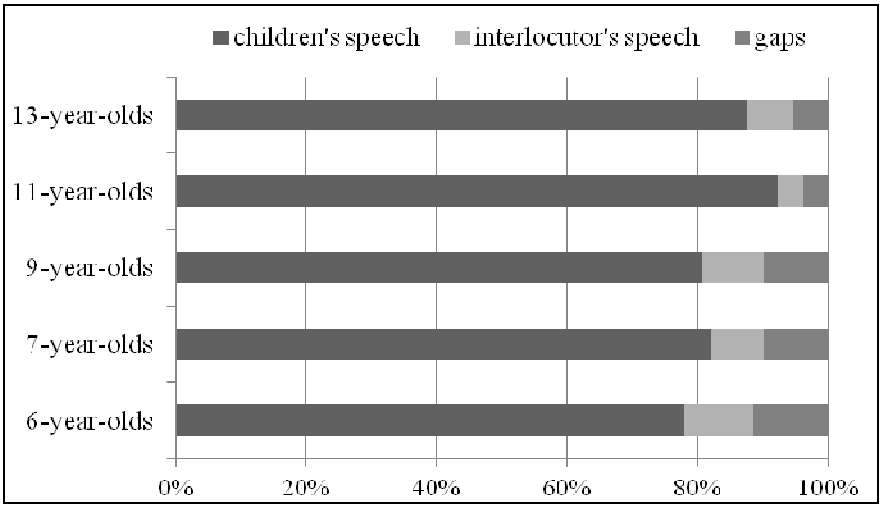

Figure 2. The proportion of speech turns and gaps in the speech samples

3.2 Occurrences of pause-to-pause intervals and pauses

We analyzed the number of pause-to-pause regions, as well as silent and filled pauses in our recordings. The speech samples of all age groups consisted of a total of 7864 pause-to-pause intervals, 6995 silent pauses and 1258 filled pauses, which were distributed non-equally by age group (Table 3). The average number of pause-

to-pause sections in the speech samples across age groups was 112 (SD: 47). The shortest talk, which lasted 1.9 minutes, contains 39, while the longest sample, which took 8.7 minutes, contains 311 pause-to-pause intervals. Children produced an average of 100 silent pauses (SD: 49); 22 silent pauses occurred in the shortest sample and 310 silent pauses occurred in the longest sample. The mean number of filled pauses was 18; however, there were two children (a 6- and a 7-year-old), who did not produce any filled pauses at all. After the age of 9, children produced filled pauses three times more often than younger children.

Table 3. The frequency of occurrence of pause-to-pause intervals and pauses (items)

Age groups	**Number of**		
	pause-to-pause intervals	**silent pauses**	**filled pauses**
6-year-olds	1560	1355	122
7-year-olds	1230	1081	107
9-year-olds	2139	1846	393
11-year-olds	1551	1473	329
13-year-olds	1384	1240	307
Total	7864	6995	1258

A marker of speech fluency is how many pause-to-pause regions and pauses are produced per minute. The more pause-to-pause sections there are, the more pauses there are to interrupt speech. Pearson's correlation analyses on our data revealed a strong positive correlation between these two measurements ($r = .939$; $p < .001$).

We first investigated the number of pause-to-pause intervals per minute in each age group. The following average values were measured: in 6-year-olds 27 (SD: 5.5), in 7-year-olds 23 (SD: 3.3), in 9-year-olds 27 (SD: 3.4), in 11-year-olds 25 (SD: 3.3), in 13-year-olds 24 (SD: 3.3) pause-to-pause sections occurred. These findings suggest that this parameter fluctuates with age. In order to compare these values, a one-way ANOVA and Tukey's post hoc test were conducted on the data. A significant main effect of age was revealed ($F(4, 69) = 2.811$; $p = .032$). A Tukey's post hoc test showed significant differences only between the values of the 6-year-old and 7-year-old children ($p = .047$). In our corpus, 6-year-old subjects produced the most pause-to-pause intervals, while 7-year-olds produced the least.

We then measured the frequency of silent pauses. Six-year-old subjects produced an average of 22.5 silent pauses per minute (SD: 5.4), 7-year-olds: 19.5 (SD: 3.9), 9-year-olds: 22.5 (SD: 5.3), 11-year-olds: 22.9 (3.9), and 13-year-olds: 21.4 (SD: 3.8). In terms of the frequency of silent pauses, the one-way ANOVA did not reveal a significant main effect for age ($p = .354$). The most silent pauses were realized by the 11-year-old children and the fewest number of pauses were produced by 7-year-olds.

Large individual differences in the number of filled pauses were found between and within age groups. The average number of filled pauses per minute were 2.4

(SD: 2.4) in 6-year-olds, 2.0 (SD: 1.4) in 7-year-olds, 5.0 (SD: 3.0) in 9-year-olds, 5.4 (SD: 3.5) in 11-year-olds and 5.3 (SD: 3.2) in 13-year-olds. One-way ANOVA showed that the differences among age groups were significant ($F(4, 69) = 5.052$; $p = .001$). A Tukey's post hoc test for multiple comparisons did not reveal significant differences between the data of 6- and 7-year-olds, but significant differences were found between the means of 7-year-olds and the other three age groups (7-year-olds from 9-, 11-, 13-year-olds: $p = .046$; $p = .015$; $p = .024$, respectively). As Figure 3 illustrates, there was a sharp increase in the frequency of filled pauses between the age of 7 and the age of 9. Filled pauses appeared less frequently in the speech of 6- and 7-year-old subjects than in the older speakers' speech.

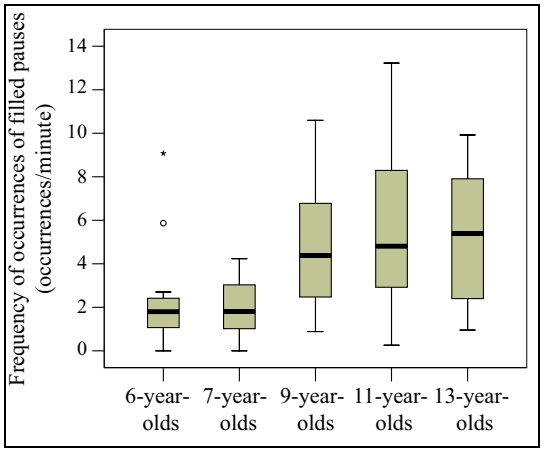

Figure 3. The frequency of occurrence of filled pauses in spontaneous speech

The usage of filled pauses seemed to depend upon the individual. Some of them preferred to produce silent pauses, while others used filled pauses when facing cognitive planning issues. The proportion of filled pauses out of the number of total pauses revealed an individual's preferred strategy, which had an effect on the perception of speech fluency. In our corpus, this ratio ranged between 0 and 42%. The mean ratio of filled pauses was the lowest in the 7-year-old subjects and the highest in 13-year-olds (Table 4). The group means suggested a rising trend of filled pause use with increasing age.

Table 4. The percentage of filled pauses out of the total number of pauses

Age groups	% filled pauses
6-year-olds	9.6%
7-year-olds	8.7%
9-year-olds	17.9%
11-year-olds	18%
13-year-olds	18.2%

3.3 Durations of pause-to-pause intervals and pauses

The mean durations of pause-to-pause region, silent pause, and filled pause are shown in Table 5.

Table 5. Temporal properties of pause-to-pause intervals and pauses

Age groups	Mean duration (ms) and SD of pause-to-pause intervals	silent pauses	filled pauses
6-year-olds	1508±1007	824±835	377±213
7-year-olds	1794±1195	872±915	427±301
9-year-olds	1465±1109	810±730	347±144
11-year-olds	1598±1185	785±764	359±158
13-year-olds	1729±1362	745±639	385±188

The mean duration of pause-to-pause intervals was 1597 ms across all age groups. In other words, the children produced fluent speech for an average of one and a half minutes before interrupting vocalization for a pause. In every age group, the most frequently occurring duration of the pause-to-pause interval was between 500 and 1000 ms. The longest pause-to-pause intervals were measured in the group of 7-year-old children, while the shortest sections were detected in the 9-year-olds. We compared the duration values between groups using a non-parametric Kruskal–Wallis test which revealed significant differences attributable to group ($\chi^2 = 86.099$; $p < 0.001$). A Mann–Whitney *U* test was used to compare the data across age groups. Significant differences were found between all age groups but one (see Table 6).

Table 6. Pairwise comparisons of pause-to-pause durations by age group

	7-year-olds	9-year-olds	11-year-olds	13-year-olds
6-year-olds	$Z = -6.091$; $p < 0.001$	$Z = -3.027$; $p = 0.002$	$Z = -0.490$; $p = 0.624$	$Z = -2.424$; $p = 0.015$
7-year-olds		$Z = -8.861$; $p < 0.001$	$Z = -5.271$; $p < 0.000$	$Z = -3.295$; $p = 0.001$
9-year-olds			$Z = -3.271$; $p = 0.001$	$Z = -5.342$; $p < 0.001$
11-year-olds				$Z = -1.998$; $p = 0.046$

The mean duration of silent pauses across all groups was 806 ms. The shortest mean duration was found in the speech of 13-year-olds and the longest mean duration was produced by the 7-year-olds. The most frequent silent pauses lasted less than 500 ms in all age groups. The Kruskal–Wallis analysis revealed a significant main effect of age on silent pause duration ($\chi^2 = 25.140$; $p < 0.001$). The results of the pairwise comparisons across age groups (using a Mann–Whitney U test) are shown in Table 7.

Table 7. Pairwise comparison of age groups regarding duration of silent pauses

	7-year-olds	**9-year-olds**	**11-year-olds**	**13-year-olds**
6-year-olds	Z = -2.942; p = 0.003	Z = -1.899; p = 0.058	Z = -1.154; p = 0.248	Z = -1.136; p = 0.256
7-year-olds		Z = -1.529; p = 0.126	Z = -3.898; p < 0.001	Z = -3.909; p < 0.001
9-year-olds			Z = -2.943; p = 0.003	Z = -2.921; p = 0.003
11-year-olds				Z = -0.001; p = 0.999

The subjects in the present corpus produced average filled pauses durations of 369 ms on average. While fewer filled pauses were noted in 6- and 7-year-olds' speech, their duration was most frequently around 200–400 ms. The differences among groups approached significance with the Kruskal–Wallis test (p = 0.076). Further testing with the Mann-Whitney U test revealed significant differences in duration of filled pauses between 7- and 9-year-olds (Z = -2,022; p = 0.043) and between 9- and 13-year-olds (Z = -2,452; p = 0.014).

Pauses (both silent and filled pauses) added up to an average of 30% to 35% of the duration of children's speech. This ratio is higher than that of the pauses in adults' spontaneous speech (20% to 30%) found in previous studies (see Duez, 1982; Misono and Kiritani, 1990; Markó, 2005; Bóna, 2007). The range of the children's data (15% to 46%) indicates great inter-speaker variability; a finding also observed for the adults.

The interrelationship between the duration of pause-to-pause intervals and pauses indicates how much perceived speech seems to be fluent. If the speaker produces relatively long pause-to-pause sections and short pauses, his/her speech seems to be more fluent than when producing short pause-to-pause intervals and/or long pauses. Figure 4 presents these relationships for the five age groups. The duration of pause-to-pause intervals is relatively short in 6-year-old children, whereas their pauses were long. In contrast, 13-year-old subjects spoke with relatively long pause-to-pause intervals interrupted by short pauses. Although the duration of the pause-to-pause intervals was long (precisely, the longest) in the speech sample of 7-year-olds, the duration of their pauses was also long.

The mean duration of pause-to-pause intervals and pauses of each subject were also measured, and the correlation of these parameters was determined. We had hypothesized that speakers who produced long pause-to-pause sections would also produce long pauses because he/she needs more time for speech planning. In our study, however, neither a positive nor negative correlation was revealed in this respect (Pearson's correlation analysis: p = .122). This finding suggests that long pause-to-pause intervals are not necessarily accompanied by long pauses (Figure 5).

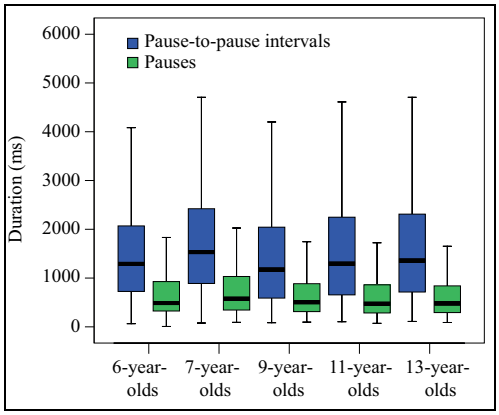

Figure 4. The duration of pause-to-pause intervals compared to both types of pauses across age groups

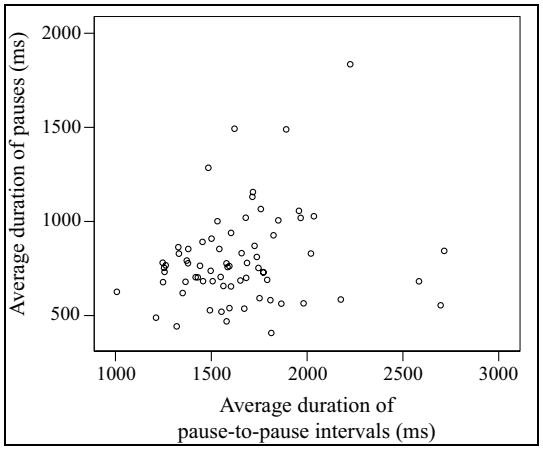

Figure 5. Interrelation between the duration of pause-to-pause intervals and pauses

3.4 The number of words in pause-to-pause interval

The average number of words per pause-to-pause interval for each group was as follows: 3.08 in 6-year-olds, 3.55 in 7-year-olds, 3.36 in 9-year-olds, 3.68 in 11-year-olds and 4.26 in 13-year-olds (Figure 6). A Kruskal–Wallis test revealed significant differences among groups ($\chi^2 = 11.829$; $p = .019$). The average number of words produced per pause-to-pause interval increased with age. In addition, large individual differences were found. The pause-to-pause interval that contained the fewest number of words consisted of 2.3 words (produced by a 6-year-old subject), while the largest number of words per pause-to-pause interval was 7.6 words (produced by a 13-year-old subject). The number of words per pause-to-pause interval showed significant correlation with the interval duration (Pearson's correlation analysis: $r = .783$; $p < .001$).

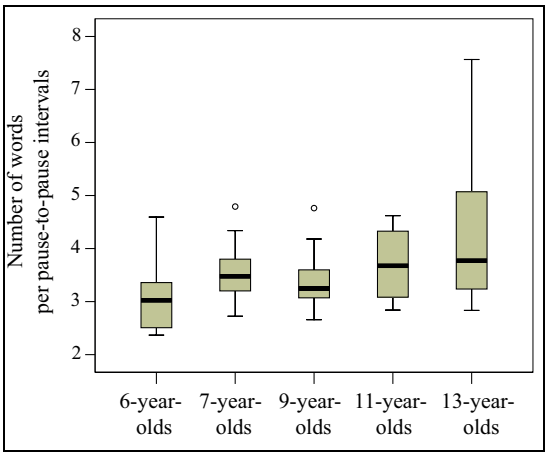

Figure 6. The number of words per pause-to-pause interval

In Loban's research (1976), it was found that the average number of words per communication unit increased from kindergarten through grade twelve both in oral and written language. A communication unit has been defined as 'a group of words which cannot be further divided without the loss of their essential meaning' (Loban 1976: 9), and the average number of words per communication unit in oral language increased from 7 to 12 words across ages.

3.5 Speech tempo of children's narratives

Our results revealed an increasing speech tempo with age (Table 8); this was similar to the findings of previous studies (e.g., Kowal et al., 1975b). A one-way ANOVA revealed significant differences in speech tempo among the five age groups (in the measurement of words per minutes: $F(4, 69) = 2.553$; $p = .047$; and in the measurement of syllables per second: $F(4, 69) = 5.712$; $p = .001$). The Tukey post hoc tests revealed significant differences between the 13-year-olds' speech tempo values and the other age groups, except for 11-year-olds ($p < .022$). The slowest speech tempo was 49.6 words per minute (or 1.75 syllables per second), while the fastest speech tempo was 140.3 words per minute (or 4.63 syllables per second). The largest within-group difference was found in 13-year-old children.

Differences in speech tempo are illustrated with two examples selected from the slowest and the fastest speech. The duration of both of the utterances was 13 s:

(1) *autózni* sil (1290 ms) *és motorozni* sil (1396 ms) *meg* sil (2184 ms) *kártyázni* ('to play with cars [sil 1290 ms] and motors [sil 1396 ms] and cards') (6-year-old boy's utterance).

(2) *olyan két-három órát szoktam tanulni másnapra ha pedig témazáró van hát akkor kicsit tovább hogy* sil (198 ms) *mindenképpen felkészüljek rá* sil (911 ms) *hát nagyon sokat szoktunk elmenni moziba vagy programokra* ('I study for the next day for about 2-3 hours but if there is final test well then a little longer [sil 198 ms] in order to prepare myself for sure [sil 911 ms] well we go to cinema or events very often') (13-year-old girl's utterance).

Table 8. The mean and range of speech tempo across age groups

Age group	Words per minute		Syllables per second	
	Mean	Range	Mean	Range
6-year-old	82.1	49.6–110.6	2.63	1.75–3.28
7-year-old	81.3	60.3–100.8	2.66	2.02–3.41
9-year-old	88.9	66.5–114.4	2.80	2.04–3.74
11-year-old	90.1	66.3–120.6	3.10	2.48–3.94
13-year-old	99.6	61.3–140.3	3.41	2.30–4.63

4 Conclusions

Many factors have impact on speech fluency both in adults and children. These differences can be observed in the temporal patterns used in adults' and children's spontaneous speech. Speakers plan the timing of their spontaneous utterances with pausing consciously and unconsciously. However, there are also unintended pauses due to the disharmonious process of speech planning and execution. The basic assumption of our research was that the temporal organization of the speech stream may provide insight into the covert processes of speech planning and execution, which might be related to different stages of language acquisition.

Everyday experience and observations have indicated that the temporal characteristics (such as the ratio and length of pauses or the speech tempo) change with age (from 6 to 14 years). However, there is limited objective, quantitative data on the temporal characteristics in the fluent speech of Hungarian-speaking children at these ages. The present analysis was carried out using spontaneous speech material gathered from 70 typically developing children. We investigated 7,864 pause-to-pause intervals, 6,995 silent pauses, and 1,258 filled pauses. Our objective data confirmed that younger children's spontaneous speech was less fluent with more pauses than that of older children. This could be related to cognitive development, physical maturation, speech routine and imitation of adult patterns. Longer pause-to-pause intervals, shorter pauses and faster speech tempo indicated that older children produce more fluent speech than the younger ones. This finding can be explained by the higher level of cognitive development in older children, which allows for the quasi-simultaneous operations of speech planning and execution.

Children seem to need more time for speech planning than adults. This assumption is supported by the results that the ratio of pauses in children's spontaneous speech was higher (30 to 35%) than in the adults (20% to 30%) (Duez, 1982; Misono and Kiritani, 1990; Markó, 2005; Bóna, 2007).

It is argued that the duration of silent pauses might reflect different underlying behaviors. For instance, the length of pauses may be connected to microplanning or macroplanning difficulties (in retrieving the phonological form, or in semantic or syntactic planning) (Goldman and Eisler, 1968, 1972; Levelt, 1989; Postma and Kolk, 1993). The findings of the present experiment revealed significant shortening of pause durations between the ages of 6 and 14 years. Furthermore, older children

have gained more experience in the organization of speech, which is not the case in younger children. It is likely that more speech experience involves more attention devoted to the listeners' needs. The length of pauses is associated with the participants' tolerance for silence in conversations; with longer gaps indicating that communication may have broken down (Mushin and Gardner, 2009). The decrease in gap duration might indicate that the tolerance for gaps decreases with age.

The frequency of occurrence of filled pauses was shown to increase rapidly between the ages of seven and nine. My impression is that children at these ages completely acquire and practice this strategy in order to resolve production uncertainties.

The differences in speech tempo among the five age groups were statistically significant. This finding can be explained by the fact that children utilize more routines in speech production, both in the articulatory movements, and in speech planning processes, as they age. As children gain greater awareness in language usage, their speech becomes more fluent, which in turn affects their speech tempo.

By comparing our results to those of previous research, slight differences can be observed. Horváth (2013) investigated the temporal organization of eighteen 9-year-old children. Their average speech tempo was 75 words per minute, while it was 88.9 words per minute for the 9-year-olds in our project. In Horváth (2013) study, average pause-to-pause intervals were 1,241 ms, silent pauses were 944 ms, and filled pauses were 379 ms long, while in the present study, these values were 1,465 ms, 810 ms and 347 ms, respectively. The similar values obtained in these two studies confirm the authenticity of the current results.

Large individual differences were evident for many of the temporal factors. For example, the ratio of pauses ranged between 15 and 46% among children. The speech tempo of 13-year-olds ranged between 61 and 140 words per minute, which shows that some of them spoke as slowly as the six-year-olds while others spoke similarly to adults.

In sum, speech fluency is affected by several different factors, which may occur together. The combined set of long pause-to-pause intervals, few and short pauses, and fast speech tempo collectively provide a fluent impression of spontaneous speech in older children. The results also show that with increasing age, children gradually get closer to the way in which adults control their speech flow.

References

Ambrose, N.G. and E. Yairi 1999. Normative Disfluency data for early childhood stuttering. *Journal of Speech, Language, and Hearing Research*, 42, 895–909.

Bloom, L. 1970. *Language Development: Form and Function in Emerging Grammars.* Cambridge, MA: MIT Press.

Boersma, P. and D. Weenink 2011. Praat: doing phonetics by computer. (Version 5.3.02) [Computer program]. (Retrieved Oct 1, 2011, from http://www.praat.org).

Bóna, J. 2007. *Production and perception of speeded up speech*. Doctoral dissertation, ELTE, Budapest. [in Hungarian]

Bóna, J. 2012. Linguistic-phonetic characteristics of cluttering across different speaking styles: A pilot study from Hungarian. *Poznan Studies in Contemporary Linguistics*, 48/2.

203–222.
Bortfeld, H., S.D. Leon, J.E. Bloom, M.F. Schober, and S.E. Brennan 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, 44/2, 123–147.
Boscolo, B., N. Bernstein Ratner, and L. Rescorla 2002. Fluency of school-aged children with a history of specific expressive language impairment: An exploratory study. *American Journal of Speech-Language Pathology/American Speech-Language-Hearing Association*, 11, 41–49.
Brotherton, P. 1979. Speaking and not speaking: Process for translating ideas into speech. In Siegman, AW. and S. Feldstein (eds.): *Of Time and Speech*. Hillsdale, New Jersey: Lawrence Erlbaum, 179–209.
Brown, R. 1973. *A First Language: The Early Stages.* Cambridge, MA: Harvard University Press.
Carlson, R., J. Hirschberg, and M. Swerts 2005. Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. *Speech Communication*, 46, 326–333.
Clark, H.H. and J.E. Fox Tree 2002. Using *uh* and *um* in spontaneous speaking. *Cognition*, 84, 73–111.
DeJoy, D. A. and H.H. Gregory 1985. The relationship between age and frequency of disfluency in preschool children. *Journal of Fluency Disorders*, 10, 107–122.
Duez, D. 1982. Silent and non-silent pauses in three speech styles. *Language and Speech*, 25, 11–25.
Edlund, J., M. Heldner and J. Hirschberg 2009. Pause and gap length in face-to-face interaction. In *Proceedings of Interspeech 2009*. 2779–2782.
Frazier, L., C.Jr. Clifton, and K. Carlson 2003. Don't break, or do: Prosodic boundary preferences. *Lingua*, 1, 1–25.
Gee, J.P. and F. Grosjean 1984. Empirical evidence for narrative structure. *Cognitive Science*, 8, 59–85.
Goldman-Eisler, F. 1968. *Psycholinguistics: Experiments in spontaneous speech*. New York: Academic Press.
Goldman–Eisler, F. 1972. Pauses, clauses, sentences. *Language and Speech*, 15, 103–113.
Gósy, M. 2009. Self-repair strategies in children's and adult's speech. In Bárdosi V. (ed.): *Quo vadis philologia temporum nostrorum?* Budapest: Tinta Könyvkiadó, 141−150. (in Hungarian)
Grosz, B. and J. Hirschberg 1992. Some intentional characteristics of discourse structure. In *Proceedings of International Conference on Spoken Language Processing*, Banff, 429–432.
Guo, L., J.B. Tomblin and V. Samelson 2008. Speech Disruptions in the Narratives of English-Speaking Children With Specific Language Impairment. *Journal of Speech, Language, and Hearing Research*, 51/3, 722–738.
Hall, N.E., T.S. Yamashita, and D.M. Aram 1993. Relationship between language and fluency in children with language disorders. *Journal of Speech and Hearing Research*, 36, 568–579.
Harley, T. 2008. *The Psychology of Language: From Data to Theory*. Hove: Psychology Press.
Haynes, W.O., and S.B. Hood 1977. Language and disfluency in normal speaking children from discrete chronological age groups. *Journal of Fluency Disorders*, 2, 57–74.
Healey, E.C., and M.R. Adams 1981. Speech timing skills of normally fluent and stuttering children and adults. *Journal of Fluency Disorders*, 6, 233–246.
Horváth, V. 2010. Filled pauses in Hungarian: Their phonetic form and function. *Acta Linguistica Hungarica*, 57/2-3, 288–306.
Horváth, V. 2013. Temporal organization of 9-year-old children's spontaneous speech. *Beszédkutatás*, 2013. 144–159.

Jusczyk, P.W. 1997. *The discovery of spoken language*. Cambridge, MA: The MIT Press.
Kajarekar, S., L. Ferrer, A. Venkataraman, K. Sonmez, E. Shriberg, A. Stolcke, H. Bratt, and V.R.R. Gadde 2003. Speaker recognition using prosodic and lexical features. In *Proceedings of the IEEE Speech Recognition and Understanding Workshop*, St. Thomas, U.S. Virgin Islands, 19–24.
Klatt, D.H. 1975. Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, 3, 129–140.
Kohler, K.J. 1983. Prosodic boundary signals in German. *Phonetica*, 40, 89–134.
Kowal, S., D. O'Connell, and E.J. Sabin 1975a. Development of temporal patterning and vocal hesitations in spontaneous narratives. *Journal of Psycholinguistic Research*, 4/3, 195–207.
Kowal, S., D. O'connell, E.A. O'brien, and E.T. Bryant 1975b. Temporal aspects of reading aloud and speaking: Three experiments. *American Journal of Psychology*, 88, 549–569.
Kreiman, J. 1982. Perception of sentence and paragraph boundaries in natural conversation. *Journal of Phonetics*, 10/2, 163–175.
Laczkó, M. 2009. The phonetic and stylistic analysis of the speech of teenagers. *Anyanyelv-pedagógia*, 2009/1. http://www.anyanyelv-pedagogia.hu/cikkek.php?id=151
Lehiste, I. 1979. Perception of sentence and paragraph boundaries. In B. Lindblom, B., and S. Öhman (eds.). *Frontiers of speech communication research*. London – New York – San Francisco: Academic Press, 191–201.
Lehiste, I., J.P. Olive, and L.A. Streeter 1976. The role of duration in disambiguating syntactically ambiguous sentences. *Journal of the Acoustic Society of America*, 60, 1199–1202.
Levelt, W.J.M. 1989. *Speaking. From intention to articulation*. Cambridge, MA, London, England: The MIT Press.
Loban, W. 1976. Language Development: Kindergarten through Grade Twelve. *NCTE Committee on Research Report* No. 18. http://files.eric.ed.gov/fulltext/ED128818.pdf
Logan, K.J., and E.G. Conture 1995. Length, grammatical complexity, and rate differences in stuttered and fluent conversational utterances of children who stutter. *Journal of Fluency Disorders*, 20/1, 35–61.
Lundholm, F.K. 2011. Pause length variations within and between speakers over time. In *Proceedings of the 15th workshop on Semantics and Pragmatics of Dialogue*, 198–199.
MacWhinney, B., and H. Osser 1977. Verbal Planning Functions in Children's Speech. Department of Psychology. Paper 200. http://repository.cmu.edu/psychology/200
Markó, A. 2005. *On some suprasegmental characteristics of spontaneous speech. Comparison of monologes and dialoges, and the analysis of humming.* Doctoral dissertation, ELTE, Budapest. [in Hungarian]
Menyhárt, K. 2012. Temporal characteristics of children's speech 60 years ago. *Beszédkutatás*, 2012, 246–259. [in Hungarian]
Misono, Y. and S. Kiritani 1990. The distribution pattern of pauses in lecture-style speech. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics*, 24, 101–111.
Mushin, I., and R. Gardner 2009. Silence is talk: Conversational silence in Australian Aboriginal talk-in-interaction. *Journal of Pragmatics*, 41/10, 2033–2052.
Natke, U., P. Sandrieser, R. Pietrowsky, and KT. Kalveram 2006. Disfluency data of German preschool children who stutter and comparison children. *Journal of Fluency Disorders*, 31, 165–176.
Nelson, K. 1973. Structure and strategy in learning to talk. *Monographs of the Society for Research in Child Development*, 39/1-2.
Ninio, A., and C.E. Snow 1996. *Pragmatic development. Essays in developmental science*. Boulder, CO, US: Westview Press.
Postma, A., and H. Kolk 1993. The covert repair hypothesis: Prearticulatory repair processes in normal and stuttered disfluencies. *Journal of Speech and Hearing Research*, 36, 472–487.

Prosek, R.A., and C.M. Runyan 1982. Temporal characteristics related to the discrimination of stutterers' and nonstutterers' speech samples. *Journal of Speech and Hearing Research*, 25, 29–33.
Rispoli, M., and P. Hadley 2001. The leading-edge: the significance of sentence disruptions in the development of grammar. *Journal of Speech, Language and Hearing Research*, 44/5, 1131–1143.
Rosenfield, I.B. 1987. *Pauses in oral and written narratives*. Boston: Boston University.
Runyan, C.M., P.E. Hames, and R.A. Prosek 1982. A perceptual comparison between paired stimulus and single stimulus methods of presentation of the fluent utterances of stutterers. *Journal of Fluency Disorders*, 7, 71–77.
Sacks, H., E.A. Schegloff, and G. Jefferson 1974. A simplest systematics for the organization of turn-taking for conversation. *Language,* 50/4, 696–735.
Shriberg, E. 2001. To 'errrr' is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31, 153–169.
Singh, L., P. Shantisudha, and N.C. Singh 2007. Developmental patterns of speech production in children. *Applied Acoustics*, 68, 260–269.
Smith, B.L., M.K. Kenney, and S. Hussain 1996. A longitudinal investigation of duration and temporal variability in children's speech production. *Journal of the Acoustical Society of America*, 99, 23–44.
Strangert, E. 2003. Emphasis by pausing. *Proceedings of the 15th International of Phonetic Sciences*, Barcelona, 2477–2480.
Streeter, L.A. 1978. Acoustic determinants of phrase boundary perception. *Journal of the Acoustical Society of America*, 64, 1582–1592.
Trainor, L.J., and B. Adams 2006. Infants' and adults' use of duration and intensity cues in the segmentation of tone patterns. *Perception and Psychophysics*, 62, 333–340.
Vallent, B. 2010. Characteristics of spontaneous narratives in high school students. *Beszédkutatás*, 2010, 199–210. (in Hungarian)
Vihman, M.M. 1996. *Phonological development: The origins of language in the child. Applied language studies*. Malden: Blackwell Publishing.
Winkler, L.E., and P. Ramig 1986. Temporal characteristics in the fluent speech of child stutterers and nonstutterers. *Journal of Fluency Disorders*, 11, 217–229.
Yaruss, J.S. 1999. Utterance length, syntactic complexity, and childhood stuttering. *Journal of Speech, Language, Hearing Research*, 42, 329–344.
Zellner, B. 1994. Pauses and the temporal structure of speech. In E. Keller (ed.), *Fundamentals of speech synthesis and speech recognition,* Chichester: John Wiley, 41–62.

DIMENSIONS STYLISTIQUE ET PHONÉTIQUE DE LA DISPARITION DE NE EN FRANÇAIS

Pierre Larrivée et Denis Ramasse
Université de Caen Basse-Normandie et CRISCO (EA4255)
pierre.larrivee@unicaen.fr, denis.ramasse@unicaen.fr

Abstract

The issue investigated in this paper is that of the competition between different factors in driving linguistic change. Such competition is given detailed consideration through the classical case of contemporary French preverbal negative clitic *ne*, the loss of which is generally explained with reference to phonetic, syntactic and stylistic dimensions. In order to establish whether phonetics or register are preponderant determiners of the change, we look at (non) realizations of the negator in a corpus of television interviews. If phonetics is primary, the non-realizations of *ne* should be promoted by phonetic environments, in particular where several reduced clitics could yield ill-formed sequences of three consonants that the omission of *ne* would repair. If primacy is to be found in register, social variables such as gender, age and education should correlate to rates of use. Statistical analysis show that such a correlation does exist, involving not gender surprisingly, but mostly age and professional occupation. Perspectives for further research are suggested in the study of factor competition in other corpora and on other comparable questions.

1 Introduction

Une des raisons pour lesquelles la variation et le changement linguistiques sont difficiles à établir dans leurs formes et leurs causes est l'enchevêtrement des facteurs divers qui s'y associent. Tel changement peut être promu à la fois par des vecteurs phonétiques, des paramètres syntaxiques, et des dimensions stylistiques. C'est le cas d'un des exemples les mieux reconnus de changement linguistique en français contemporain, la disparition du clitique préverbal de négation *ne*. Le fait d'envisager les choses sous un seul angle comme le font maintes études a le défaut d'obscurcir l'éventuelle convergence des différents rapports sous lesquels existe et évolue une forme. La négation *ne* en français actuel est un marqueur de style ; il appartient à une zone clitique en dehors de laquelle il ne se retrouve pas ; son autonomie phonétique est de même réduite à titre de pronom atone contenant un schwa susceptible de réduction. C'est le rapport entre le phonétique et le social comme déterminant de la disparition de *ne* en français contemporain que nous explorons dans ce travail. Notre but est d'établir lequel de ces deux paramètres a un rôle prépondérant pour un phénomène de changement bien documenté. En utilisant un

corpus d'entretiens télévisuels où devrait se manifester une production abondante de *ne* selon le style de la prise de parole, nous sommes à même d'établir si les non-emplois de *ne* sont corrélés à des facteurs primordialement sociaux, ou à des facteurs essentiellement phonétiques. Le travail est présenté selon le plan suivant. La partie initiale rappelle les travaux sur les facteurs responsables du déclin de *ne* en français contemporain, et les débats dans lesquels ils s'inscrivent. La partie suivante cherche à contribuer à ces débats en présentant les données recueillies sur le (non-)emploi de *ne* dans le corpus télévisuel utilisé et le poids respectif des facteurs sociaux et phonétiques. Les résultats et leur portée sont envisagés dans la section finale.

2 Facteurs sociaux et facteurs phonétiques

Le changement diachronique dans l'expression de la négation de proposition est un phénomène rendu notoire par sa régularité dans de nombreuses langues. Connu sous le nom de cycle de Jespersen, le phénomène consiste à voir un marqueur préverbal de négation se doubler d'une négation postverbale qui subsiste à la disparition de la négation initiale (Larrivée et Ingham, 2011 ; van Gelderen, 2011 ; Breitbarth, Lucas et Willis, 2010). Illustrée par le cas du français qui passe de *ne* seul à *ne … pas* puis à *pas*, ce changement implique la disparition aujourd'hui pratiquement achevée de *ne* dans les styles vernaculaires. La disparition de *ne* est l'objet d'un nombre important de travaux qu'animent quatre questions fondamentales : la disponibilité ou non du marqueur dans la grammaire des locuteurs, la nature de variable stylistique stable ou en déclin de *ne*, la prise en charge à côté de la valeur stylistique d'une dimension pragmatique, et les causes de la disparition.

En tant que variable particulièrement saillante du français actuel, *ne* voit ses taux d'emplois abondamment documentés pour le français d'Europe (Armstrong, 2002 ; Ashby, 2001 ; 1981, 2001 ; Coveney, 1996 ; Fonseca-Greber, 2007 ; Gadet, 1997 ; Hansen et Malderez, 2004 ; Moreau, 1986) et d'Amérique du Nord (van Compernolle, 2010 ; Poplack et St-Amand, 2007 ; Sankoff et Vincent, 1977), comme ils le sont pour le français langue seconde autour de la question de la maîtrise du sociolinguistique par les non-natifs d'une langue (par exemple Coveney, 1998 ; Dewaele et Regan, 2002 ; Rehner et Mougeon, 1999 ; van Compernolle et Williams, 2009). Les chiffres fournis pour les pratiques vernaculaires sont de 5 % de rétention en français européen (avec des résultats plus élevés pour les études plus anciennes, voir le tableau de Armstrong et Smith, 2002 : 28 ; de van Compernolle, 2009), avec un taux dix fois inférieur en français québécois. C'est pourquoi on pourrait envisager l'emploi de *ne* comme dans ces styles une insertion plutôt que de voir son absence comme une élision. Conceptuellement, parler d'élision pour 95% des cas de figure apparaît problématique comme le souligne Fonseca-Greber (2007). Si l'emploi de *ne* représente une insertion, cela pourrait signifier que la grammaire du français vernaculaire *ne* prévoit plus de position syntaxique pour ce marqueur lorsqu'il n'est pas employé. C'est la conclusion à laquelle Claire Blanche-Benveniste s'oppose ; qu'il soit réalisé ou non, *ne* a toujours une place prévue dans

la syntaxe (avis partagé par Martineau, 2011). Elle en veut pour preuve le fait que tout locuteur est susceptible de produire des ne à des taux élevés dans les contextes d'interlocution appropriés, même les enfants ; c'est l'expérience des Dames snobs où des jeunes filles jouent le rôle de dames dans un restaurant en adoptant le registre qui convient (Blanche-Benveniste et Jeanjean, 1987). Si les mêmes locuteurs produisent aussi facilement des *ne*, il est difficile d'imaginer qu'ils passeraient pour ce faire d'une syntaxe à une autre, l'enjeu étant de savoir si le français vernaculaire a une grammaire différente de la pratique normée. C'est donc à l'intérieur de la même grammaire que *ne* marquerait le style normé, ce sur quoi s'accordent tous les auteurs. Certains soulignent que le style communiqué par un tel marqueur *ne* caractérise pas nécessairement tout un échange, ainsi que le montrent les cas de micro-changement de style selon la nature du sujet par exemple, où un *ne* est utilisé pour marquer la formalité d'une intervention dans un contexte informel (Fonseca-Greber, 2007 inter alia). Ce statut stylistique serait stable selon Blanche-Benveniste (1995), soutenue en cela par Dufter et Stark (2008), s'appuyant dans les deux cas sur le Journal d'Héroard. Les notes minutieuses prises sur la vie et la langue du futur Louis XIII par son médecin nous montreraient que l'omission de *ne* est un trait du style vernaculaire depuis le XVIIe siècle, et probablement avant, ce qu'on pourrait montrer si on avait les sources pour le faire. On établit en outre que la chute de ne dans Héroard est soumise aux mêmes conditions que dans la langue contemporaine, et en particulier la nature du sujet, le sujet clitique amenant un taux important d'absence. S'il est vrai que l'omission de *ne* se retrouve déjà à date ancienne (Ingham 2011 et les références qu'il cite), c'est surtout à partir du XVIIIe siècle qu'est perceptible la chute de *ne* (Martineau, 2011 ; Martineau et Mougeon, 2003). Ce déclin est particulièrement bien illustré par l'étude d'Ashby (2001), qui contraste la pratique de locuteurs tourangeaux dans un corpus d'entretiens réalisés en 1995 répliquant des entretiens de 1976 dans la même région. L'analyse montre que l'ensemble des locuteurs produit moins fréquemment la négation préverbale qu'il y a vingt ans. La comparaison des taux actuels de 5 % de rétention attestés par les études réalisés sur des données postérieures à 1985 et de ceux beaucoup plus élevés pour les données antérieures suggère de même que la variable stylistique approche la fin d'un processus de changement historique. Ce changement en cours et les faibles taux d'emploi suggèrent à différents auteurs que les occurrences de *ne* dans les styles vernaculaires se justifient par une valeur pragmatique d'emphase (Fonseca-Greber, 2007 ; van Compernolle, 2009 ; Williams, 2009). Les exemples d'échanges informels entre amis en français suisse étudiés par Fonseca-Greber montreraient une association entre l'emploi de *ne* et des facteurs comme l'accent d'insistance et des intensifieurs. Ces facteurs peuvent favoriser l'emploi de *ne*, qui n'a pas de valeur pragmatique de façon catégorique (Larrivée, 2010). Ce déclin, quelles en sont les causes ? Si on ne peut valider la réanalyse comme marqueur pragmatique d'emphase, qu'est-ce qui amène le marqueur stylistique à être moins employé dans les registres vernaculaires ? La question n'est pas résolue en disant que cela est dû à la valeur stylistique de *ne*, puisque cette valeur n'empêchait pas l'usage de *ne* de par

le passé. On trouve là la tension bien documentée entre la réanalyse catégorique et la disparition graduelle d'un marqueur. On observe que les omissions sont associées à des contextes caractérisés, les expressions très fréquente et plus ou moins figées *il y a, c'est* et *je sais* (Moreau, 1986 ; Gadet, 1997 ; Coveney, 1996), et surtout la nature du sujet, la chute de *ne* étant nettement plus fréquente avec les sujets clitiques (voir la synthèse et les nouvelles données de Meisner et Pomino, 2012). Cela attire l'attention sur le statut de clitique de *ne* lui-même. Les clitiques ont un comportement syntaxique particulier, qui explique leur acquisition relativement tardive en langue maternelle (en particulier Meisel, 2008). En outre, ils sont justiciables de réductions diverses particulièrement bien repérées pour la diachronie (Larrivée, 2012 ; Wanner, 1999 ; Posner, 1985). Ces réductions s'expliquent vraisemblablement par le fait que les clitiques sont composés de schwas finaux souvent élidés, ce qui peut créer des groupes consonantiques complexes qu'éviteraient les élisions. La disparition de *ne* serait-elle liée aux réductions qui affectent le groupe de clitiques préverbaux du français ? Si ces réductions ont une dimension phonétique au sens large, l'environnement phonétique a-t-il un impact sur l'emploi de *ne* et son omission ? C'est à répondre à cette question à notre connaissance nouvelle que nous voulons contribuer dans ce qui suit.

3 Les données

Le but de travail est d'établir ce qui du registre ou du phonétique est le facteur prédominant pour la disparition de *ne* en français. Si le phonétique est déterminant, des interventions même dans des contextes normalement marqués par le registre normé, devraient donner des signes de chute de *ne* suivant l'environnement phonique. On songe en particulier aux suites de consonnes que l'élision de *ne* permettrait d'éviter. Si au contraire le registre reste la dimension dominante, ce seront surtout les interventions se donnant comme appartenant au style vernaculaire (échanges spontanés non-surveillés) qui amèneront la chute de *ne*, quelle que soit la qualité phonétique du contexte. Pour savoir laquelle de ces prédictions est soutenue par les faits, il nous a semblé opportun de choisir un corpus de parole publique, où les réalisations devraient être plus abondantes et où on s'attendrait à voir prédominer les facteurs stylistiques.

3.1 Description du corpus

Le corpus est constitué d'interviews attestant de la variété normée du français de France ne présentant pas de marques régionales ou étrangères dans la syntaxe, le lexique ou la prononciation[1]. La majorité de ces interviews a été diffusée à la télévision, dans des journaux comme *le journal de 20 heures* (TF1) ou *le 19/20* (France 3) ou dans des émissions littéraires (*La Grande Librairie*), des talk-shows (*On n'est pas couché*) ou des émissions politiques (*Ripostes*). D'autres interviews

[1] Même si l'accent de Didier Deschamps est perçu comme "chantant", ce n'est qu'une légère nuance d'un français considéré néanmoins comme normé.

ont été trouvées sur des sites web (*Allociné*, *MusiqueMag*, *etc.*). Les enregistrements (vidéos, dont les bandes son ont été extraites) s'inscrivent dans une période de 4 ans, de septembre 2008 à août 2012.

Table 1. Liste des extraits analysés avec la référence des locuteurs et des interviews. Pour faciliter leur identification dans la description, un code a été attribué à chacun.

code	prénom	nom	date de naissance	lieu de naissance	activité	origine de l'interview	date de l'interview
f01CB	Chimène	BADI	30/10/1982	77 Melun		télé-loisirs.fr	22/11/2011
f02JB	Josiane	BALASKO	15/04/1960	Paris		On n'est pas couché (France 2)	05/05/2012
f03LB	Leïla	BEKHTI	06/03/1984	Issy-les-Moulineaux		Le journal de 20h (TF1)	08/08/2012
f04CC	Cécile	COULON	13/06/1990	Clermont-Ferrand	écrivain	La Grande Librairie (France 5)	16/02/2012
f06BG	Brigitte	GIRAUD	13/05/1905	Sidi Bel Abbès	écrivain	La Grande Librairie (France 5)	13/10/2011
f07NL	Nolwenn	LEROY	28/09/1982	29 Saint-Renan		MusiqueMag	25/07/2011
f09LM	Laure	MANAUDOU	09/10/1986	69 Villeurbanne		Le journal de 20h (TF1)	25/06/2012
f10SM	Sophie	MARCEAU	17/11/1966	Paris		Le journal de 20h (TF1)	27/03/2012
f12VP	Vanessa	PARADIS	22/12/1972	94 Saint- Maur-des Fossés		laissez-vous tenter (RTL.fr)	02/03/2012
f13NP	Natacha	POLONY	18/04/1975	95 Notre-Dame-De-Pury		On n'est pas couché (France 2)	16/06/2012
f14AP	Audrey	PULVAR	21/02/1972	Fort-de-France		On n'est pas couché (France 2)	16/06/2012
f15AS	Anna	SAM	1979	35		Ripostes (France 5)	28/09/2008
f19AT	Audrey	TAUTOU	09/08/1976	63 Beaumont		Le journal de 20h (TF1)	26/05/2012
m01DA	Daniel	AUTEUIL	24/01/1950	Alger		19/20 (France 3)	19/02/2012
m02CB	Christophe	BARBIER	25/01/1967	74 Sallanches	journaliste	myboox.fr	15/06/2012
m03PB	Patrick	BRUEL	14/05/1959	Tlemcen (Algérie)		allocine	24/06/2012
m04EC	Émmanuel	CARRÈRE	09/12/1957	Paris		La Grande Librairie (France 5)	15/09/2011
m05AC	Alain	CHAMFORD	03/02/1949	Paris		On n'est pas couché (France 2)	16/06/2012
m06GC	Gérard	COLLARD	21/05/1952	94 Perreux	libraire	On n'est pas couché (France 2)	05/05/2012
m07RD	Raymond	DEPARDON	06/07/1942	69 Villefranche-sur-Saône		La Grande Librairie (France 5)	07/10/2010
m08DD	Didier	DESCHAMPS	15/10/1968	64 Bayonne		Le journal de 20h (TF1)	09/08/2012
m09AD	Alou	DIARRA	15/08/1981	93 Villepinte	€ équipe de France de football	Le journal de 20h (TF1)	11/06/2012
m10JPF	Jean-Pierre	FOUCAULT	23/11/1947	Marseille		On n'est pas couché (France 2)	16/06/2012
m12JH	Johnny	HALLIDAY	15/06/1943	Paris		Le journal de 20h (TF1)	14/05/2012
m18S	(Julien)	SOAN (Decroix)	06/04/1981	Annemasse		On n'est pas couché (France 2)	16/06/2013
m19RZ	Roschdy	ZEM	28/09/1965	Gennevilliers		Le journal de 20h (TF1)	08/08/2012

Les interviewés sont au nombre de 26 (13 femmes, 13 hommes) ; ces locuteurs avaient entre 21 ans (Cécile COULON, venue présenter son 4e roman) et 68 ans (Raymond DEPARDON) lors de leur enregistrement. Leur activité est révélatrice du choix des médias ; ce sont des écrivains, des journalistes, des sportifs (football, natation), des acteurs, des chanteurs ; il y a aussi un libraire et un photographe (Raymond DEPARDON)[2]. Tous les locuteurs sont français ; même si certains sont nés à l'étranger, ils ont vécu en France depuis l'enfance, et ils vivent actuellement en France ; c'est le cas, en particulier, de Leila Bekhti. Selon ces critères, plusieurs interviews comme celle de Tony Parker (il a passé son enfance en Belgique et il vit le plus souvent aux États-Unis) ont été éliminées de ce corpus. La liste des locuteurs du corpus (avec la référence des enregistrements) figure dans le tableau 1. Leur activité n'a été précisée que pour quelques-uns, un peu moins connus. Pour faciliter leur identification au cours de l'analyse de leurs énoncés, un code a été attribué à chacun (avec, bien sûr, respectivement *f* et *m* pour les locuteurs féminins et masculins ; la discontinuité dans la numérotation s'explique par le fait que ce corpus est extrait d'un corpus plus important ─ 38 locuteurs).

2 Il n'y a pas d'interview de personnalités politiques car ce corpus est destiné aussi à l'enseignement et il était préférable de ménager les sensibilités politiques ; trois interviews ont d'ailleurs fait l'objet d'exercices de transcription, et des analyses acoustiques, en particulier prosodiques, ont été faites sur plusieurs autres.

3.2 Analyse du corpus

Les phrases négatives de ces 26 locuteurs ont exhaustivement été étudiées quant à la présence ou l'absence de *ne*. Pour chaque phrase, l'absence ou la présence du clitique négatif a été notée et, quand il était présent, sa forme élidée ou non. Le relevé a été fait avec une étude du contexte immédiat que le proclitique ait été réalisé ou non : nature et fonction de l'élément précédent et suivant. Pour l'élément suivant, il a été noté s'il était à initiale consonantique ou vocalique.

Sur 382 phrases négatives, 257 (soit 67%, 2/3) présentent une omission du proclitique négatif alors qu'il y a emploi de *ne* dans 125 (33%, 1/3).

3.2.1 Analyse phonétique ; les différentes réalisations de *ne*

3.2.1.1 Réalisation acoustique de ne. La réalisation de [n] se fait par ce qui est appelé un "murmure nasal" et se manifeste selon Fry (1979) par deux formants : un premier formant bas (formant nasal résultant d'une amplification par le nez, le résonateur nasal) et un formant au niveau du F_3 de la voyelle voisine. Ce genre de manifestation de cette consonne nasale a été confirmé récemment par Angélique Amelot (2004) qui, se fondant en particulier sur l'étude de Fujimura (1962), rappelle que [n] comporte théoriquement 4 formants mais que les 2^e et 3^e formants nasals disparaissent à cause d'un problème de couplage entre le résonateur pharyngo-buccal et le résonateur nasal, ce qui crée des atténuations pouvant aller jusqu'à l'élimination d'amplifications dans une partie du spectre, atténuations appelées "antirésonances" :

> Pour /n/, le "cluster" est constitué du deuxième et du troisième formant, et des anti-formants. Dans ce cas, le premier et le quatrième formant sont assez stables. (Amelot, 2004 : 28)

Les formants du [ə] français ont la fréquence suivante en moyenne selon Bürki et collaborateurs (2008) dans une étude menée sur 10 locuteurs :

F3 = 2880 Hz

F2 = 1760 Hz

F1 = 390Hz

Cependant, une récente étude faite sur les voyelles du français à partir de 40 locutrices par Georgeton et collaborateurs (2012) permet de trouver des valeurs différentes. Bien que le schwa n'y soit pas décrit, une comparaison avec les voyelles les plus proches permet d'obtenir des valeurs approximatives de chaque formant. Le [ə] français est une centrale arrondie d'aperture intermédiaire entre mi-ouvert et mi-fermé. D'après le tableau des valeurs formantiques qu'ils donnent et qui comprend leurs valeurs et celles de deux autres descriptions (Calliope, 1989 ; Gendrot et Adda-Decker, 2005), on peut déduire pour [ə] le F_1 d'après les valeurs de [a] qui est central, le F_2 d'après les valeurs des mi-fermées et des mi-ouvertes et le F_3 d'après celles de [ə], qui est la voyelle arrondie la plus proche ; ce qui donne :

$2580 \text{ Hz} < F_3 < 2700 \text{ Hz}$

$1430 \text{ Hz} < F_2 < 1677 \text{ Hz}$

400 Hz < F_1 < 600 Hz

3.2.1.2. Réalisation avec [ə]. Parmi les 125 emplois de *ne*, 34 clitiques négatifs sont réalisés avec un schwa, tous suivis d'un mot à initiale consonantique, à une exception près dans une hésitation : "et **ne** et ne pas…" (m07RD). Un exemple de réalisation (m02CB) avec [ə] est donné dans le spectrogramme de la figure 1. Conformément à la description de Fry (1979), le [n], présente un premier formant bas (350 Hz) et un deuxième formant (2507 Hz) au niveau du F_3 de la voyelle voisine, en l'occurrence le [ə] ; la valeur des 3 formants est la suivante :

F_3 = 2509 Hz

F_2 = 1540 Hz

F_1 = 430 Hz

A part une valeur un peu plus faible du F_3, on voit qu'il y a pour [ə] une correspondance avec les valeurs de la deuxième étude citée, et, conformément à la description de Fry, une grande proximité entre le formant haut de [n] et le F_3 de la voyelle (seulement 2 Hz d'écart, mesures effectuées par *Speech Analyzer* 3.1).

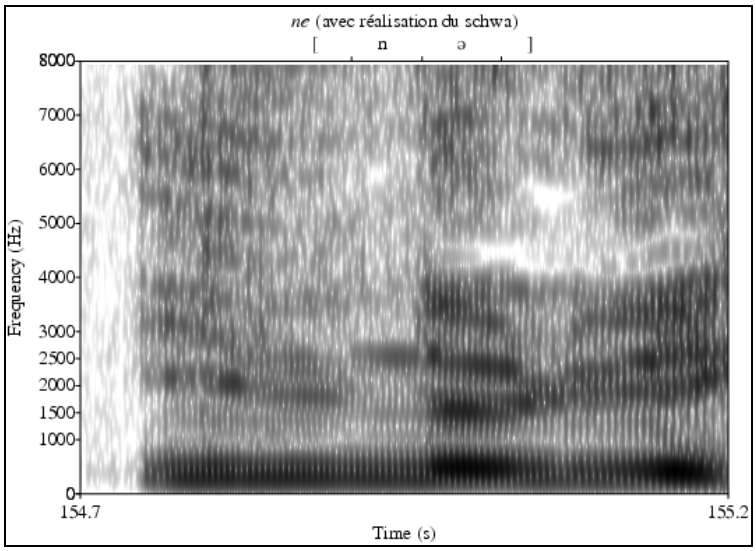

Figure 1. Exemple de réalisation (extrait m02CB) de *ne* avec schwa sur un spectrogramme obtenu avec le logiciel *Praat* (version 5.3.16) ;
on remarque la présence de 2 formants, un bas à 350 Hz et un haut à 2507 Hz ; les 3 premiers formants du [ə] ont les valeurs suivantes : F1 = 390Hz, F_2 = 1540 Hz, F_3 = 2509 Hz (mesures effectuées grâce au logiciel *Speech Analyzer* 3.1).

3.2.1.3. Réalisation avec élision. Sur les 91 réalisations sans [ə], 65 relèvent d'une élision, c'est-à-dire de la chute de la voyelle finale d'un mot devant un autre mot commençant par une voyelle[3].

Dans certains cas d'élision, une difficulté s'est présentée lors de l'analyse des phrases. Il s'agit de la distinction entre un [n] de négation et un [n] de liaison. Ceci survient avec le pronom *on* précédant un mot commençant par une voyelle. 12 cas peuvent être dénombrés : 1 devant *était*, 1 devant *avait*, 5 devant *est* et 5 devant *a*. Comme dans la phrase [ɔ̃napasy] ; est-ce que le [n] est un [n] de liaison, ce qui donnerait *On a pas su* ou un [n] de négation, ce qui s'écrirait *On n'a pas su*. Pour résoudre ce problème, des analyses acoustiques ont été faites en comparant le [n] pour lequel la question se posait avec un [n] d'une phrase affirmative du même locuteur (en général, très proche dans l'énoncé), pour lequel on était sûr que c'était une liaison.

Un examen de spectrogrammes était décisif, un [n] de négation étant réalisé de façon beaucoup plus nette qu'un [n] de liaison. L'exemple (figure 2) d'une comparaison de deux [n] prononcés par Alou DIARRA entre une voyelle [ɔ̃] et une voyelle [a] permet de l'illustrer. Grâce à ces spectrogrammes, on a une mise en regard de la réalisation acoustique des deux sortes de [n]. On voit que le [n] de liaison, à gauche, n'est pas très net ni très stable ; en revanche, le formant haut du [n] de négation à droite est très net et très stable jusqu'à la fin de la réalisation de la consonne.

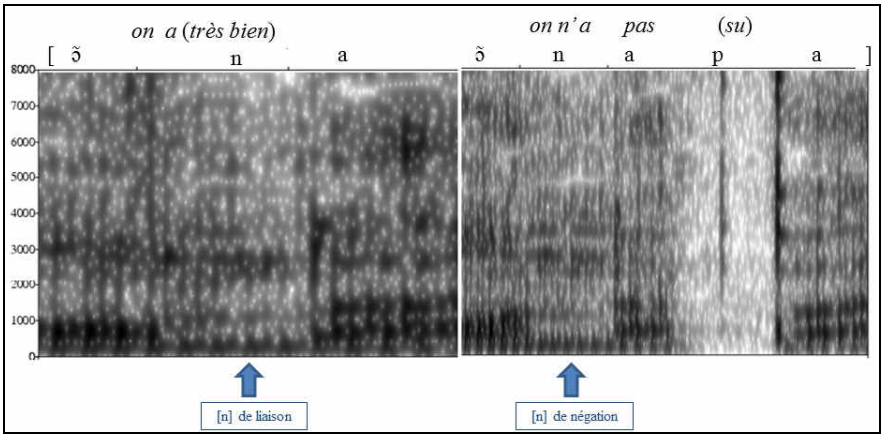

Figure 2. Exemple de réalisation (extrait m09AD) de *ne* avec élision (à droite) ; la comparaison avec un [n] de liaison (à gauche) permet de mettre en évidence la spécificité du [n] de négation.

[3] Bürki et collaborateurs (2008) ne font pas de distinction entre chute de schwa devant voyelle (cas correspondant à ce qui est appelé l'élision de [ə]) et chute de schwa devant consonne, en français chute du "e caduc". L'élision de [ə] sera donc ici distinguée de la chute de e caduc.

3.2.1.4. Réalisation avec chute de l' "e caduc". Il y 26 réalisations avec chute de l'e caduc, cet *e susceptible de tomber* et qui *se prononce seulement lorsqu'il est nécessaire pour éviter la rencontre de trois consonnes* selon Grammont (1914), qui ne donne que des exemples de chute devant consonne.

On remarque que le [n] de négation est réalisé très nettement devant consonne (*cf.* un exemple extrait de f04CC, fig. 3), même si cette consonne est non voisée et on s'aperçoit qu'il n'y a pas d'assimilation régressive du caractère non voisé de cette consonne comme l'illustrent la fig. 4 (f01CB) devant [p] et la fig. 5 (m03PB) devant [s]. Dans ce corpus, quand il apparaît avec chute de schwa, il n'est jamais précédé par une consonne non voisée. Il ne peut donc pas y avoir ici d'assimilation progressive du caractère non voisé d'une consonne sur le [n] de négation dans ce contexte.

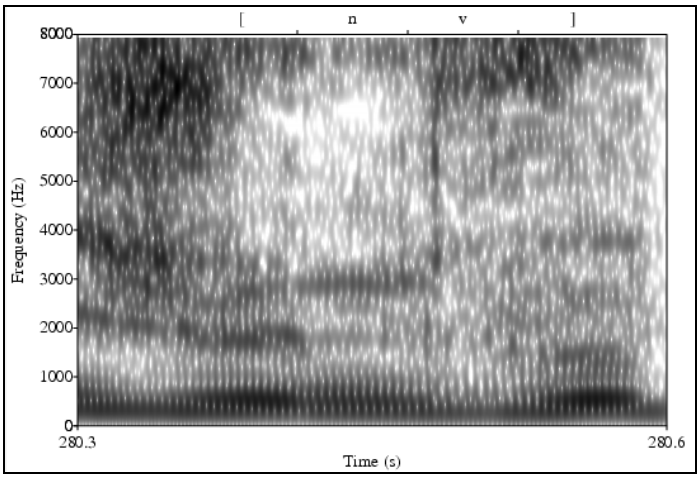

Figure 3. Exemple de réalisation (f04CC) de ne avec chute de schwa ;
la chute s'est faite devant la fricative voisée [v].

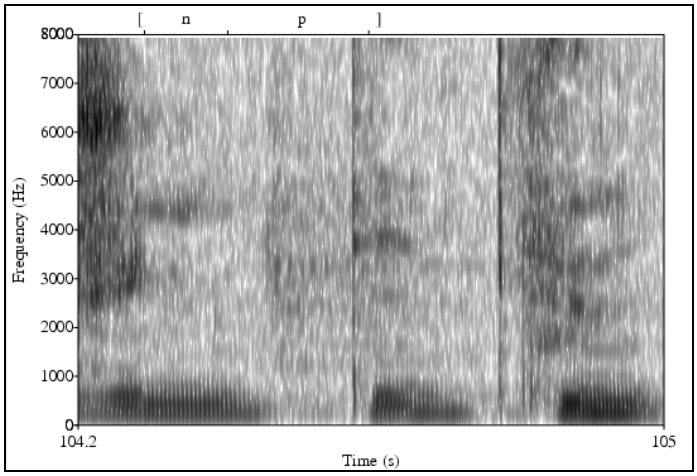

Figure 4. Exemple de réalisation (f01CB) de *ne* avec chute de schwa ; la chute s'est faite devant la plosive non voisée [p].

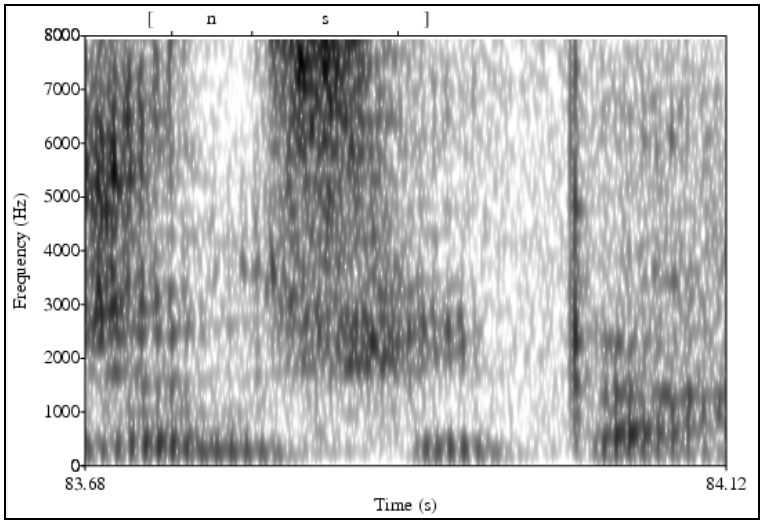

Figure 5. Exemple de réalisation (m03PB) de *ne* avec chute de schwa ; la chute s'est faite devant la fricative non voisée [s].

3.2.1.4 Cas particulier de réalisation avec contact de 3 consonnes. Le corpus contient 5 cas de réalisation de *ne* avec une chute de e caduc à la source d'une séquence de trois consonnes, ce qui tendrait à montrer que le contexte phonétique n'est pas un élément déterminant pour l'emploi ou l'omission du clitique négatif. Dans 4 cas, il y a la semi-voyelle labiale-palatale [ɥ] comme troisième consonne (en troisième position), ce qui donne les groupes :

[nfɥ] dans [ʒə nfɥi] "je ne fuis"

[nsɥ] dans [ʒə nsɥi] "je ne suis"

[npɥ] dans [ɔ nfɥi] [$\tilde{ɔ}$ npɥis] "on ne puisse"

[nlɥ] dans [sa nfɥi] [sa nlɥi] "ça ne lui"

Mais dans un 5^{e} cas apparaît le groupe [ʁnv], dans la phrase "ceux qui ont assisté au premier concer**t n'v**oulaient pas sortir…". La représentation spectrographique de ce groupe de trois consonnes est donnée en figure 6 ; il s'agit d'une phrase de Jean-Pierre Foucault (m10JPF).On s'aperçoit que le [n] est bien réalisé mais que la coarticulation avec les deux autres consonnes tend à affaiblir l'antirésonance qui d'habitude occulte un formant intermédiaire entre les deux formants "habituels" ; c'est pourquoi on voit se manifester le formant aux environs de 1000 Hz dont parle Fant (1960 : 147) dans sa description du [n] ; ce formant se manifeste ici à 1119 Hz. Donc malgré ce formant supplémentaire, cette consonne manifestée ici reste bien un [n].

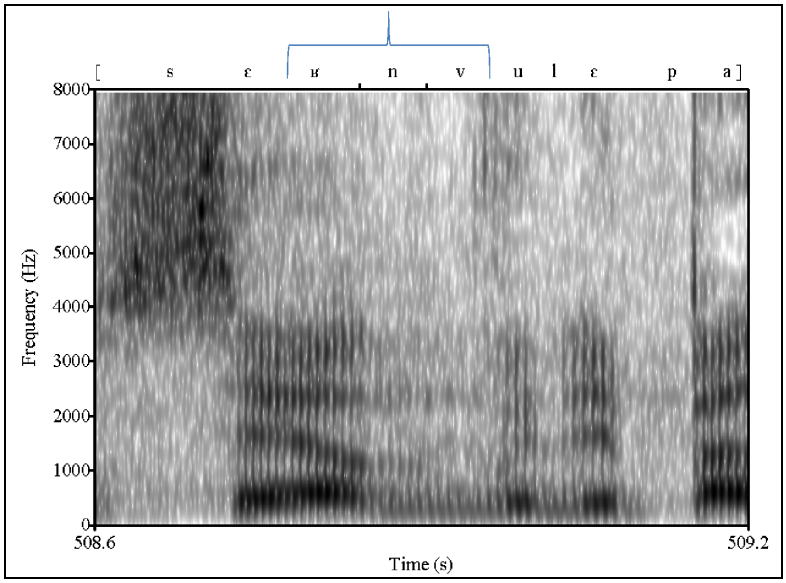

Figure 6. Groupe de 3 consonnes [ʁnv] provoqué par l'emploi d'un clitique négatif avec chute de schwa dans la phrase (m10JPF) :
"ceux qui ont assisté au premier conce**rt n'v**oulaient pas sortir…"
On remarque l'apparition d'un formant supplémentaire aux environs de 1000 Hz provoquée par la coarticulation.

3.2.1.5 Résumé de l'analyse phonétique. La figure 7 rappelle les conclusions de cette partie ; le [nə] avec consonne et voyelle n'apparaît que dans 34 phrases ; la voyelle est alors réalisée avec un F_1 de 500 Hz, un F_2 de 2500 Hz et un F_3 de 2600 Hz environ. Le [n] se manifeste par un « murmure nasal » avec deux formants, un bas (350 Hz), et un haut au niveau du F_3 de la voyelle suivante. Il y a chute du [ə] devant voyelle (élision) dans 65 phrases et chute devant consonne (chute de e caduc) dans 26 phrases.

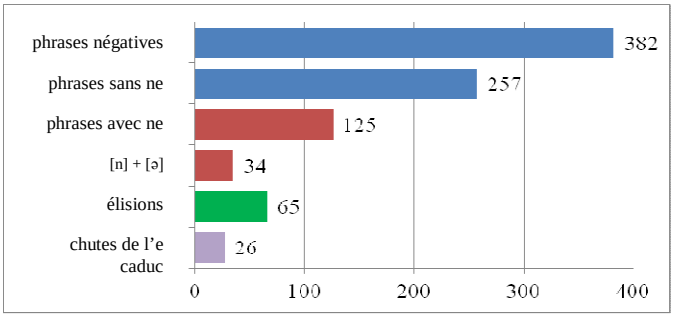

Figure 7. Différentes réalisations de la négation dans le corpus étudié
Sur 382 phrases négatives, seulement 125 comportent un proclitique negative dont 34 avec la voyelle [ə], 65 avec élision (chute du [ə] devant voyelle) et 26 avec chute de l'e caduc (chute de [ə] devant consonne)

3.2.2 Facteurs sociaux ; emploi de *ne* en fonction des locuteurs

3.2.2.1 Établissement de groupes. Après une analyse plus détaillée, les locuteurs ont été regroupés selon plusieurs critères et les étiquettes des groupes dont la définition est décrite ci-dessous sont indiquées en italiques ; elles figureront dans les différents tableaux ou graphes d'analyse statistique. En plus de la répartition selon le sexe F ou M, les critères retenus sont l'âge, le niveau d'éducation et la profession. La dimension de l'âge a été répartie en trois groupes à peu près homogènes après une étude de distribution par histogrammes :

groupe 1 : 20 ans à 35 ans ; *âge 20-35* (8 locuteurs)

groupe 2 : 35 ans à 50 ans ; *âge 35-50* (8 locuteurs)

groupe 3 : 50 ans à 68 ans ; *âge 50-68* (10 locuteurs)

Le niveau d'études est de même appréhendé en trois groupes :

groupe 1 : niveau inférieur au bac ; *< bac* (13 locuteurs)

groupe 2 : niveau bac ou équivalent ; *bac* (5 locuteurs)
Dans ce groupe entrent des locuteurs ayant fait une terminale, même sans avoir obtenu le bac (c'est le cas de Patrick BRUEL qui a été en terminale à Henri IV) et de trois locutrices qui ont été en première année d'université, l'une en ayant suivi une formation sur l'art-thérapie (Leïla BEKHTI), les autres ayant fait une première année en droit anglo-américain (Nolwenn LEROY) ou en lettres modernes (Audrey TAUTOU).

groupe 3 : *> bac +3* (8 locuteurs)
Appartiennent à ce groupe des diplômés de l'enseignement supérieur, par exemple, une agrégée (Natacha POLONY), un ingénieur (Emmanuel CARRERE) ou même une étudiante n'ayant pas encore obtenu son diplôme final (Cécile COULON, qui a fait hypokhâgne, khâgne et une année de faculté).

C'est en 5 catégories qu'ont été regroupés les locuteurs sous le rapport de leur activité ; l'activité prise en compte est le plus souvent l'activité professionnelle, mais parfois aussi celle qui justifie l'interview, c'est le cas pour Cécile COULON, bien qu'à l'heure actuelle elle ne soit pas écrivain. Les locuteurs ont donc été répartis en 5 groupes selon leur activité :

groupe 1 : *artiste* (12 locuteurs)
pour désigner les chanteurs et les acteurs réunis puisque certains comme Vanessa PARADIS et Patrick BRUEL sont à la fois acteurs et chanteurs

groupe 2 : *écrivain* (5 locuteurs)

groupe 3 : *sportif* (3 locuteurs)

groupe 4 : *journaliste* (4 locuteurs)

groupe 5 : *autre* (2 locuteurs, un libraire et un photographe)

3.2.2.2 Analyse statistique. Pour chaque type de répartition, quand les données s'y prêtaient, une Analyse Factorielle des Correspondances (AFC) a été pratiquée (Bachelet, 2010). Il y avait toujours, bien sûr, 2 modalités des données en colonnes :

l'emploi ou l'omission de *ne*, mais il a paru intéressant d'en ajouter une troisième : la régularité pour chaque locuteur dans l'emploi ou l'omission de *ne*, évaluée par le calcul de la valeur absolue de la différence entre le nombre de *ne* employés et le nombre de *ne* omis.

- le sexe

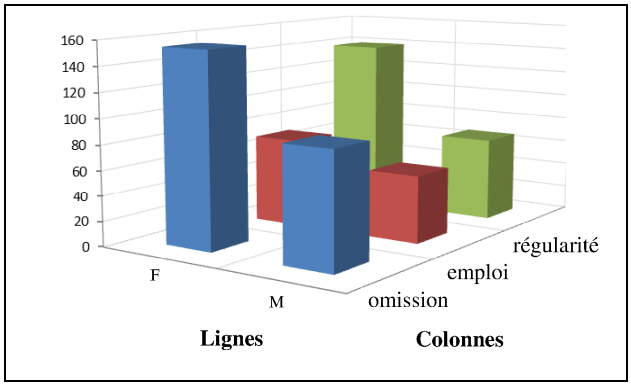

Figure 8. Vue 3D du tableau de contingence avec le sexe comme modalité en lignes

Le tableau de contingence est représenté figure 8, selon un test de χ^2 il n'y a pas de différence statistiquement significative dépendant du sexe du locuteur dans l'emploi ou l'omission de *ne*.

- l'âge

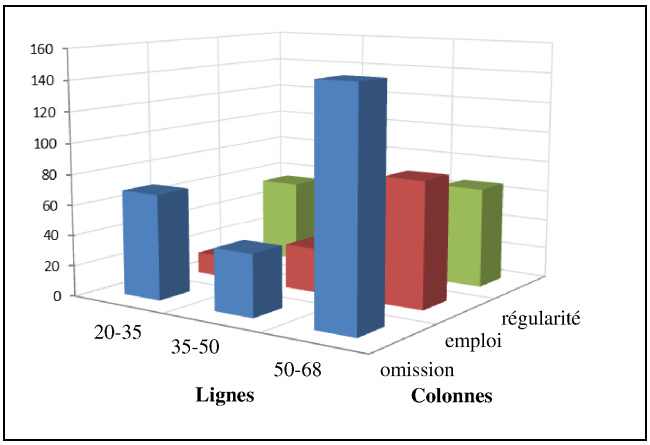

Figure 9. Tableau de contingence avec l'âge comme modalité

Contrairement au cas précédent, le χ^2 est significatif (p <.01) et une AFC a été pratiquée. Le tableau de contingence et le *mapping* (représentation graphique des résultats) de l' AFC avec l'âge comme modalité sont représentés respectivement figure 9 et figure 10. Le mapping révèle une différence de répartition très nette selon l'axe F_1 entre les moins de 35 ans (à droite) pour qui l'omission de *ne* est une pratique régulière et les plus de 35 ans (à gauche) qui en font un emploi qui, sans

être vraiment régulier, est plus important. L'axe F_2 montre qu'il faut faire une distinction entre la classe d'âge 50-68 ans qui fait un emploi régulier du *ne*, et la classe des 35-50 ans qui en fait un emploi moyen.

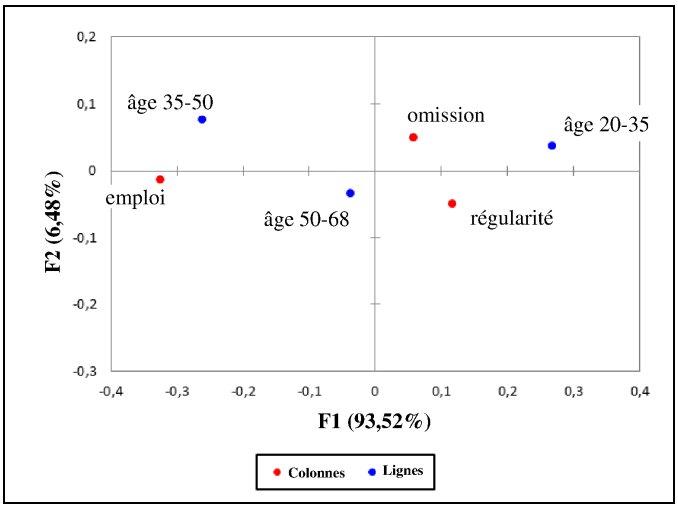

Figure 10. Mapping de l'AFC avec l'âge comme modalité

- le niveau d'études

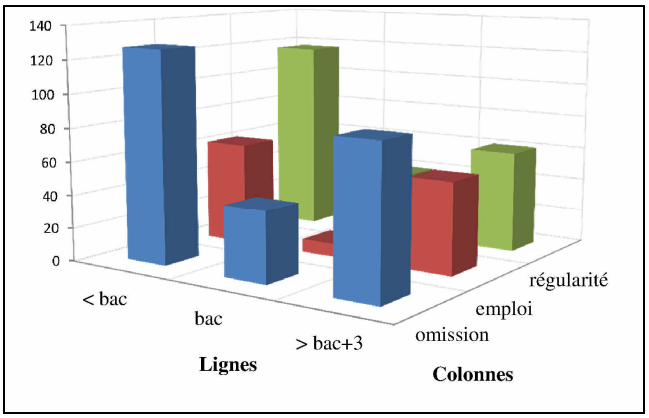

Figure 11. Vue 3D du tableau de contingence avec le niveau d'études comme modalité

Là aussi, en ce qui concerne le niveau d'études des locuteurs, le χ^2 est significatif ($p < .01$) ; le tableau de contingence est présenté figure 11 (l'indice de régularité pour les locuteurs du niveau bac étant de 40), et le mapping de cette AFC, figure 12. L'axe F_1 représente 96 % de la variance, on trouve, à gauche, les locuteurs d'un niveau bac + 3 ou supérieur qui emploient le *ne*, et à droite les locuteurs d'un niveau égal ou inférieur au bac qui omettent le *ne*. L'axe F_2 met en évidence une plus grande régularité dans l'omission par les locuteurs d'un niveau d'études inférieur au

bac, qui figurent en bas, par rapport aux locuteurs du niveau bac, en haut, qui ne présentent pas cette régularité. Pour le niveau d'études, on peut conclure que les locuteurs d'un niveau supérieur ou égal à bac +3 emploient *ne* alors que les locuteurs d'un niveau inférieur au bac l'omettent régulièrement, cette régularité dans l'omission étant beaucoup plus faible en ce qui concerne les locuteurs d'un niveau équivalent au bac.

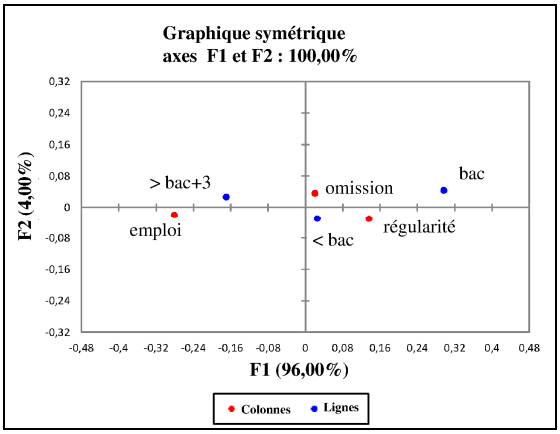

Figure 12. Mapping de l'AFC avec le niveau d'études comme modalité

- l'activité

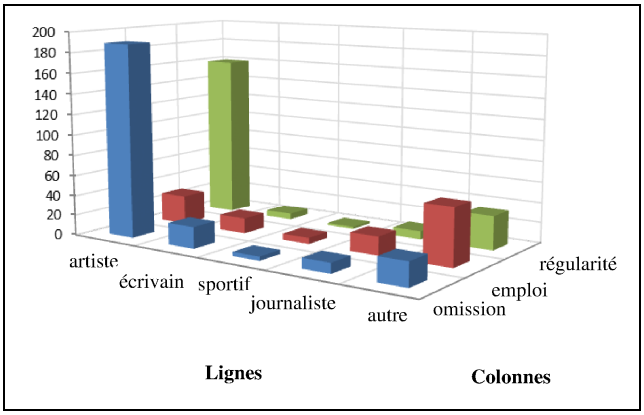

Figure 13. Vue 3D du tableau de contingence avec l'activité comme modalité en lignes

Là-aussi, le degré de signification du χ^2 en ce qui concerne l'activité des locuteurs est inférieur à .01 (le tableau de contingence est présenté figure 13). Le mapping de cette AFC est présenté Figure 14. Selon ce mapping, il y a dans ce cas également une répartition très nette selon l'axe F_1, qui représente 75.19 % de la variance, mais entre la profession d'artiste (à gauche) au sein de laquelle les locuteurs pratiquent régulièrement l'omission de *ne*, profession d'artiste que l'on

peut opposer à toutes les autres activités. Selon l'axe F_2, on trouve, en bas, les écrivains et les sportifs qui ont une grande régularité dans leur emploi de *ne*, contrairement aux journalistes et autres professions (photographe et libraire) qui sont très irréguliers dans leur emploi de *ne*.

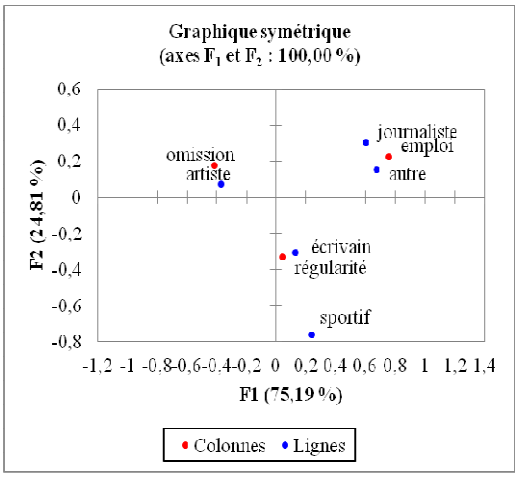

Figure 14. Mapping de l'AFC avec l'activité comme modalité

3.2.2.3 Synthèse et résumé de l'analyse des facteurs sociaux. Les modalités ont jusqu'ici été étudiées séparément.

Pour savoir comment elles se regroupent pour établir des classes, un algorithme de classification automatique, Classification Ascendante Hiérarchique (CAH), a été utilisé.

Le résultat est représenté dans le tableau 2, où se trouve le détail de la composition des classes, la figure 15 avec le dendrogramme illustrant leur répartition et la figure 16 détaillant leur profil.

Tableau 2. Résultat obtenu par le test de classification ascendante hiérarchique (CAH) ; la classe 1 correspond à une omission régulière, tandis que les trois autres correspondent à un emploi de *ne*, important et régulier dans la classe 3, puis d'importance et de régularité décroissantes, progressivement la classe 2 et la classe4.

Classe	**1**	**2**	**3**	**4**
	artiste âge20-35	autre âge50-68 >bac+3	journaliste écrivain sportif âge35-50 bac	<bac

Le dendrogramme de la figure 15 montre que les données se structurent selon l'omission de *ne* à gauche et l'emploi de *ne* à droite.

A gauche, une seule classe :

- la classe 1 : ces locuteurs qui omettent le clitique négatif ont une profession d'artiste et/ou un âge inférieur à 35 ans.

La branche de droite permet de distinguer trois classes :

une classe qui pratique un emploi régulier du clitique :

- la classe 3 : elle comprend des locuteurs, de la plupart des autres professions, de 35 à 50 ans, et d'un niveau bac.

Deux autres classes qui pratiquent moins régulièrement l'emploi du clitique :

- la classe 2 : elle comprend des locuteurs avec une profession de photographe ou de libraire, les locuteurs les plus âgés et les locuteurs d'un niveau d'études supérieur
- la classe 4 : elle comprend des locuteurs d'un niveau d'études inférieur au bac.

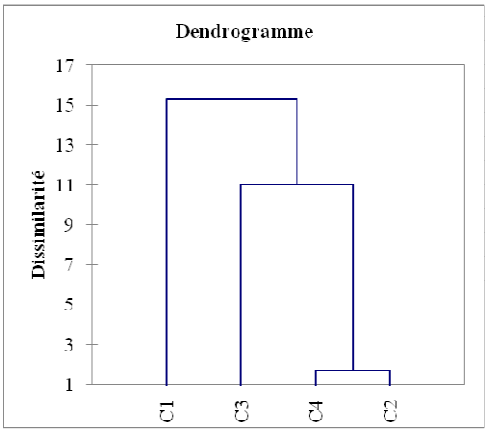

Figure 15. Dendrogramme illustrant la répartition des données en ce qui concerne l'emploi de *ne*, avec la branche de gauche correspondant à une omission et celle de droite à un emploi. Différents degrés d'emploi peuvent être distingués dans la branche de droite, un degré important avec la classe 3, et un degré intermédiaire avec les classes 4 et 2.

Le profil des classes illustré figure 16 montre sur la double échelle d'une part de l'omission de *ne*, d'autre part de la régularité de cette omission, l'existence de deux classes totalement opposées, la classe 1, celle de l'omission, et la classe 3, celle d'un emploi régulier, entre ces deux extrêmes, les deux classes intermédiaires 2 et 4 manifestent un emploi moins important et moins régulier.

Cette classification montre de façon très nette que le niveau d'études n'est pas déterminant en ce qui concerne l'emploi ou l'omission de *ne* puisque tous les niveaux d'études envisagés apparaissent dans la branche principale de droite (celle de l'emploi). Il n'intervient donc que pour déterminer le degré d'emploi ; paradoxalement, ce degré n'est pas corrélé au niveau d'études, mais les locuteurs d'un niveau bac utilisent beaucoup le clitique négatif contrairement à ceux d'un niveau inférieur ou supérieur.

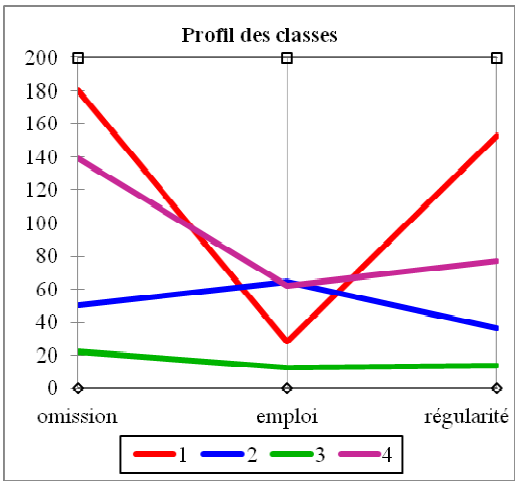

Figure 16. Description du profil des classes obtenues par la classification ascendante hiérarchique. La hiérarchie dans l'omission de *ne* est la suivante :
artiste, âge 20-35
niveau d'études <bac
autre profession (libraire, photographe), âge 50-68, bac+3
journaliste, écrivain, sportif, âge 35-50, niveau bac (emploi le plus régulier)

On peut ainsi ranger les quatre classes distinguées selon une échelle omission ↔ emploi de *ne* :

omission

↕ artiste, âge 20-35
niveau d'études < bac
autre profession (libraire, photographe), âge 50-68, bac + 3
journaliste, écrivain, sportif, âge 35-50, niveau bac

emploi

4 Discussion et conclusion

Le but de cet article est d'établir le poids relatif de l'environnement phonétique et des dimensions sociales comme déterminant primordial de la (non) réalisation du clitique négatif *ne* en français actuel. Pour ce faire, un corpus d'entretiens médiatique a été retenu. De tels entretiens permettaient non seulement de recueillir des données plus abondantes dans la mesure où un marqueur stylistique comme *ne* y est normalement mieux maintenu, mais également de faire peser les facteurs sociaux défavorisant a priori une interprétation phonétique. Et en effet, exactement un tiers (33%) des phrases négatives est réalisé avec un *ne* dans le corpus, niveau de réalisation élevé par rapport aux autres études s'expliquant par la situation publique médiatique des interactions. La méthode appliquée consiste à identifier les emplois de *ne* dans le corpus, et de les coder selon des variables sociales et phonétique.

L'identification des emplois apporte une contribution nouvelle par l'étude des formants révélant la distinction entre un [n]de liaison et la réalisation de *ne*. Les données statistiques rendues possible par le codage montrent que dans ce corpus, la réalisation du clitique négatif n'est pas corrélée à des facteurs phonétiques. Bien entendu, l'élision du schwa est catégorique devant un mot à initiale vocalique, qui se retrouve aussi devant les consonnes. Les suites de consonnes avec *ne* se limitent à deux, bien que des suites de trois « consonnes » se retrouvent avec la semi-consonne [ɥ]. Cette observation pose question aux présupposés de la règle des trois consonnes, et appelle des réévaluations, en termes de hiérarchie de sonorité possiblement. La présence ou l'absence de *ne* est corrélée à des facteurs sociaux, et en particulier à certains critères :

- le genre du locuteur n'a aucun impact significatif sur la rétention de *ne* ;
- le niveau d'études intervient de façon très limitée ;
- en revanche, l'âge du locuteur est important ; en simplifiant, plus on est jeune, moins on utilise le clitique négatif. Cela soutient l'idée d'un changement en cours, selon lequel *ne* disparaît effectivement, et sera moins utilisé par les plus jeunes, y compris dans les contextes médiatique ;
- l'activité du locuteur est un élément prépondérant : les professionnels de la parole utilisent beaucoup le *ne*. Il faut noter que cette tendance est plus importante pour les professionnels de l'oral (journalistes à la radio ou à la télévision) que ceux de l'écrit (écrivains). Il semble que les marqueurs stylistiques soient en rapport avec la présentation de soi et l'image de la profession qui est la base de l'interaction ; cela met en exergue le rôle de la relation au public, laquelle définit le stylistique pour Bell (1984). On remarque que les professionnels du spectacle, appelés *artistes* dans cette étude, les chanteurs et les acteurs, sont ceux qui utilisent le moins le clitique négatif.

Il est surprenant de remarquer que, chez les locuteurs qui font un très grand usage du *ne*, la disparition du clitique négatif est plus fréquente et devient presque systématique en fin d'interview. Ceci a été remarqué chez Jean-Pierre FOUCAULT, mais aussi pour une locutrice sur laquelle certains détails manquaient (âge, niveau d'études…) et qui n'a pu être retenue dans le corpus (son activité aurait été classée en *autre*, ce qui confirme pour cette modalité les résultats obtenus). Elle faisait un emploi quasi systématique de *ne* pendant presque toute la durée de l'interview et ce n'est qu'à la fin que des omissions ont pu être constatées. Le marquage d'un positionnement stylistique demande une attention qui se reporte vers le contenu du propos à mesure qu'il s'engage.

Ce que cette étude montre est qu'un marqueur stylistique est justiciable de variations même dans le contexte où il est le plus attendu. Avec un taux d'emploi d'un tiers, il est suggéré qu'un marqueur stylistique en cours de disparition relève de la présentation de soi dans l'interaction publique médiatique et de la représentation qu'on se fait et qu'on attribue aux auditeurs de sa profession. Comment ces

représentations sont rattachées au style linguistique est une question complexe dont on peut espérer qu'elle trouve une réponse dans de futurs travaux.

Réferences

Amelot, A. 2004. *Etude aérodynamique, fibroscopique, acoustique et perceptive des voyelles nasales du français.* Thèse, Paris III.

Armstrong, N. 2002. Variable deletion of French ne : A cross-stylistic perspective. *Language Sciences* 24, 2, 153-173.

Armstrong, N. et A. Smith. 2002b. The influence of linguistic and social factors on the recent decline of French ne. *Journal of French Language Studies* 12,2, 23-41.

Ashby, W.J. 2001. Un nouveau regard sur la chute du ne en français parlé tourangeau : s'agit-il d'un changement en cours? *Journal of French Language Studies* 11,1, 1-22.

Ashby, W.J. 1991. When does variation indicate linguistic change in progress? *Journal of French Language Studies* 1, 1-19.

Ashby, W.J. 1981. The loss of the negative particle ne in French: A syntactic change in progress. *Language* 57,3, 674-687.

Ashby, W.J. 1976. The Loss of the negative particle ne in Parisian French. *Lingua* 39, 119-137.

Bachelet, R. 2010. L'analyse factorielle des correspondances. http://rb.ec-lille.fr

Bell, A. 1984. Language Style as Audience Design. *Language in Society* 13, 145-204.

Blanche-Benveniste, C. 1995. De quelques débats sur le rôle de la langue parlée dans les évolutions diachroniques. *Langue française* 107, 25-35.

Blanche-Benveniste, C. et C. Jeanjean. 1987. *Le français parlé.* Paris : Didier.

Breitbarth, A., C. Lucas et D. Willis (dirs.). 2010. *The Development of negation: the languages of Europe and the Mediterranean.* 2 volumes. Oxford: Oxford University Press.

Bürki, A., I. Racine, H-N. Andreassen, C. Fougeron et U. Frauenfelder 2008. Timbre du schwa en français et variation régionale : une étude comparative. *XXVIIe Journées d'Études sur la Parole*, Avignon

Calliope 1989. *La Parole et son traitement automatique.* Paris : Masson,.

Coveney, A. 1998. Awareness of linguistic constraints on variable ne omission. *Journal of French Language Studies* 8,2, 159-187.

Coveney, A. 1996. *Variability in spoken French. A Sociolinguistic study of interrogation and negation.* Exeter: Elm Bank.

Dewaele, J.-M. et V. Regan. 2002. Maîtriser la norme sociolinguistique en interlangue française : le cas de l'omission variable de 'ne'. *Journal of French Language Studies* 12,2, 123-148.

Dufter, A. et E. Stark. 2008. La linguistique variationnelle et les changements linguistiques 'mal compris' : Le cas de la 'disparition' du ne de négation. B. Combettes et C. Marchello-Nizia (dirs). *Études sur le changement linguistique en français.* Nancy : Presses Universitaires de Nancy, 115-128.

Fant, G. 1960. *Acoustic Theory of Speech Production.* Mouton, La Haye.

Fry, D.B.1979. *The Physics of Speech.* Cambridge University Press.

Fonseca-Greber, BB. 2007. The emergence of emphatic ne in conversational Swiss French. *Journal of French Language Studies* 17,3, 249-275.

Gadet, F. 2000. Des corpus pour ne ... pas. M. Bilger (dir.). *Corpus, méthodologie et applications linguistiques.* Paris: Champion. 156-167.

Gadet, F. 1997. *Le français ordinaire.* Paris: Armand Colin.

Gelderen, E. v. 2011. *The linguistic cycle. Language change and the language faculty.* Oxford: Oxford University Press.

Gendrot, C. et M. Adda-Decker 2005. Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German.

Proceedings of Interspeech. 2453-2456.
Georgeton, L., N. Paillereau, S. Landron, J. Gao et T. Kamiyama 2012. Analyse formantique des voyelles orales du français en contexte isolé : à la recherche d'une référence pour les apprenants de FLE. *19ème conférence sur le Traitement Automatique des langues naturelles*, Grenoble, 145-148.
Grammont, M. 1914. *Traité pratique de prononciation française*. Paris : Delagrave.
Hansen, A.B. et I. Malderez. 2004. Le "ne" de négation en région parisienne: une étude en temps réel. *Langage et Société* 107, 5-30.
Ingham, R. 2011. Ne-drop and indefinites in Anglo-Norman and Middle English. P. Larrivée et R. Ingham (dirs). *The evolution of negation: Beyond the Jespersen cycle.* Berlin: Mouton de Gruyter. 157-176.
Larrivée, P. 2012. *Reanalysis of negatives as polarity markers? The last 400 years of decline of the French preverbal negative clitic.* MS, Université de Caen.
Larrivée, P. 2010. The Pragmatic motifs of the Jespersen Cycle. Default, activation and the history of negation in French. *Lingua* 120,9, 2240-2258
Larrivée, P. et R. Ingham (dirs). 2011. *The evolution of negation: Beyond the Jespersen cycle.* Berlin: Mouton de Gruyter.
Martineau, F. 2011. Ne-absence in declarative and yes/no interrogative contexts: Some patterns of change. P. Larrivée et R. Ingham (dirs). *The evolution of negation: Beyond the Jespersen cycle.* Berlin: Mouton de Gruyter. 191-224.
Martineau, F. et R. Mougeon. 2003. Sociolinguistic research on the origins of ne deletion in European and Quebec French. *Language* 79,1, 118-152.
Meisel, J.M. (2008) Child second language acquisition or successive first language acquisition ? B. Haznedar et E. Gavruseva (dirs.). *Current Trends in Child Second Language Acquisition: A Generative Perspective.* Amsterdam et Philadelphie : Benjamins, 55–82.
Meisner, C. et N. Pomino. 2012. *Synchronic variation in the expression of French negation.* Document d'accompagnement, Université de Zurich, 16 pages.
Moreau, M.-L. 1986. Les séquences préformées : entre les combinaisons libres et les idiomatismes. Le cas de la négation avec ou sans ne. *Le Français moderne* 54, 137-160.
Poplack, S. et A. St-Amand 2007. A real-time window on 19th century vernacular French: The Récits du français québécois d'autrefois. *Language in Society* 36: 707-734.
Posner, R. 1985. Post-verbal negation in non-standard French: A historical and comparative view. *Romance Philology* 39, 170–197.
Rehner, K. et R. Mougeon. 1999. Variation in the spoken French of immersion students: To ne or not to ne, That is the sociolinguistic question. *Revue canadienne des langues vivantes* 56,1, 124-154.
Sankoff, G. et D. Vincent 1977. L'emploi productif de ne dans le français parlé à Montréal. *Le Français Moderne* 45, 243-256.
Compernolle, R.A. v. 2010. The (slightly more) productive use of ne in Montreal French chat. *Language Sciences*, 32,4, 447-463.
Compernolle, RA. v. 2009. Emphatic ne in informal spoken French and implications for foreign language pedagogy. *International Journal of Applied Linguistics*, 19,1, 47-65.
Compernolle, R.A. v. et L. Williams 2009. Variable omission of ne in real-time French chat: A corpus-driven comparison of educational and non-educational contexts. *Canadian Modern Language Review*, 65,3, 413-440.
Wanner, D. 1999. Clitic clusters in Romance : A modest account. J. Franco et al. (dir.). *Grammatical analyses in Basque and Romance linguistics.* Amsterdam et Philadelphie : Benjamins. 257-277.
Williams, L. 2009. Sociolinguistic variation in French computer-mediated communication. A variable rule analysis of the negative particle ne. *International Journal of Corpus Linguistics* 14,4, 467-491.

BOOK REVIEWS

Ndinga-Koumba-Binza, Hugues Steve (2012)
A phonetic and phonologic account of the Civili vowel duration,
Newcastle upon Tyne: Cambridge Scholars Publishing.(xii + 309 pp.,
ISBN 1-4438-3609-5, $67.99)

Reviewed by: Christopher R. Green
University of Maryland - Center for Advanced Study of Language (CASL)
e-mail: greencr@umd.edu

A phonetic and phonological account of the Civili vowel duration is among the first substantive works focused on the phonology of a Gabonese language. Civili [iso:vif] is spoken by approximately four thousand people in Gabon; nearly another seven thousand speakers of the language can be found in nearby nations. The overall objective of this work is to establish, from acoustic and perceptual points of view, the status of "short-sounding" vs. "long-sounding" vowels (hereafter short and long) in the language. The phonemic status of short vs. long vowels in Civili is a matter of longstanding discrepancy discussed in the earlier work of the author and several of his contemporaries.

The book is divided into seven chapters that cover a variety of topics related to Civili and the phonetic and phonological behavior of its vowels. Despite its 309 pages, the prose of the monograph is quite short. Chapters include 1) Introduction (pp. 1-12) - an introduction to Civili and underlying motivations for the study; 2) Background and Literature Review (pp. 13-20) - an overview of the literature on Civili; 3) Overview of the Vowel Duration Issue (pp. 21-41) - further discussion of the seemingly problematic behavior of Civili vowels; 4) Acoustic Aspects and Physical Nature (pp. 42-72) - an acoustic and statistical study of Civili vowels; 5) Perception and Vowel Duration (pp. 73-97) - a perception study of Civili vowels that includes three tasks; 6) Phonological Implications (pp. 98-126) - discussion of phonological implications related to the findings; and 7) Conclusion (pp. 127-135). The majority of the remaining 152 pages are comprised of eight appendices. These include Appendix A: Corpus for Vowel Duration (pp. 136-144); Appendix B: Minimal Pairs Based on Vowel Duration (pp. 145-147); Appendix C: Samples of Text Grids for Test Words (pp. 148-154); Appendix D: Duration Measurements (pp. 155-165); Appendix E: Statistical Results of Vowel Duration Measurements (pp. 166-209); Appendix F: List of Words Recorded for Stimulation (pp. 210-212); Appendix G: Pages, Stimuli, Responses, and Answers (pp. 213-219); and Appendix H: Perception Experiment Results (pp. 220-288). While the materials in these

appendices may be pertinent to the phenomena under study, they are beyond what is necessary to include in the book to support the author's findings.

The author provides a thorough explanation of extant work on Gabonese languages, with a focus on the Civili literature. Upon beginning the theoretical and analytical portion of the book, the author describes the sociolinguistic situation in Gabon, Civili dialectology, and the importance of resolving the issue of Civili vowel duration leading to the creation of a standard orthography for the language. There are unfortunately two immediately apparent aspects of the book that distract the reader from the author's focus on the phenomena under study. The first of these is a perpetual restatement of overall "aims", "intents", and "objectives" of the study and its components in each chapter, which are often restated in subsections of a chapter. This causes the monograph to read less like a cohesive exploration into a particular topic and more like separate, disjointed papers. A second is a frequent use of parenthetical and anecdotal sidebars which offer definitions of even the most common linguistic terms and concepts (often quoting and citing dictionaries and introductory textbooks), as well as arguments explaining the merits and goals of particular types of linguistic research. These then give the monograph a textbook feel, rather than a sophisticated contribution to the experimental phonetics and phonology literature. The book also suffers from a substantial number of formatting and typographical errors.

Besides these stylistic criticisms, the argumentation in the book is fairly difficult to follow. That is, in an attempt to be theoretically and analytically neutral, the author obscures his own underlying assumptions about Civili, which leaves the reader without a grounded idea of the true challenge that Civili vowel duration presents. To the author's credit, the reader eventually learns in Chapters 6 and 7 how the author chooses to analyze his findings. The ultimate finding is that although Civili has minimal pairs containing vowels that are both acoustically and perceptually identifiable as short vs. long, the length contrast is neutralized in the direction of long vowels in predictable phonological environments (e.g. before nasal+consonant clusters and after glides), while it is analogously neutralized in the direction of short vowels in certain syntactic environments. Instead of laying out this anomaly in an early chapter, the author focuses on issues of autosegmental formalizations and orthography to introduce readers to the phenomena.

The outcomes of the author's perception experiments are most telling, wherein Civili speakers do not reliably identify short vs. long vowels; however they readily discriminate between them in both same/different (A vs. B) and icon (ABX) discrimination tasks. The outcome and discussion of the acoustic study are less convincing, as they contain data from four speakers, one of whom is a clear outlier, compared to the other three speakers. I shall set aside a variety of other concerns that distract less significantly from the author's findings overall.

While this book may be lacking somewhat in technical and analytical sophistication, the author should nonetheless be commended for his contribution to the study of African languages within an experimental framework. The book

provides interested students and scholars with insights into the challenges that some languages present and some possible methods and techniques by which to investigate them.

Anna Łubowicz (2012)
The Phonology of Contrast
Bristol: Equinox. (134 pp.
ISBN 9781845534165. Hardcover. Price: £60.00/$95.00)

Reviewed by: **Noam Faust**
The Hebrew University, Jerusalem, Israel.
e-mail: faustista@yahoo.com

General Introduction. *The Phonology of Contrast* by Anna Łubowicz appears in the series *Advances in Optimality Theory*. As its title suggests, it is devoted to contrast, and specifically to contrast in the phonological signal. However, it is not a book about whether contrast is an active principle in grammar, or even why it is an active principle in the grammar; rather, it explores how contrast, assuming it is such an active principle, can be used to account for certain problems within the popular framework of Optimality Theory. In other words, if one is interested in principled arguments for or against contrast - this is not where they should look. If one is interested in a discussion about the cognitive workings of contrast in phonology – again, they will be disappointed in the book. But if you are interested in how contrast can be put in constraint form and shown to interact with other, more straightforwardly phonological requirements - Łubowicz's book was written for you. In it, you will find the latest methods in such formalization, along with the adequate comparisons to previous, competing analyses.

All that said, the main claim of the book *is* that contrast is an active principle of linguistic knowledge, at least in Phonology. The evidence in favor of the claim can be summarized as follows: some phenomena can be explained by factoring in contrast requirements. The analyses that emerge make better predictions or cover more grounds than analyses that do not factor in such requirements. This is a legitimate claim. However, that contrast is useful in language analysis does not prove that it exists as a principle in language…

Moreover, the manner in which contrast is used in the book left me, for one, unimpressed. In the different analyses, the general principle of contrast is broken down into "a family of constraints" – Contrast-in-vowel, Contrast-in-stress, Constrast-in-features – and these constraints are ranked above or below a phonological requirement. As often happens to me with work in Optimality Theory, I ended up suspending my disbelief for far too long, only to find that there was nothing to relieve it in the end, besides the "success" of the analysis. As a result, I

was in fact unconvinced that the analysis proposed fared significantly better than previous analyses of the same phenomena.

The book is structured in the following manner. Chapter 1 (pp. 1-8) is an introduction. Chapter 2 (pp. 9-58) introduces the theory and discusses in depth the case of chain shifts in Finnish. Chapters 3 (pp. 59-88) and 4 (pp. 89-112) apply contrast-based analyses to stress-epenthesis interaction in Arabic and allomorph-selection in Polish. A short conclusion is found on p. 113, and two technical appendices follow (p.115, 117). There is an 11-page bibliography (p.121), and an index (p.133).

Each of the three analytic chapters is a unit in its own right, and can be read as such. In what follows, I will briefly summarize the analyses in these chapters, highlighting the central ideas advanced.

Chapter 2 of the book presents the framework, which is called "PC theory" (for Phonological Contrast). The main problem of this chapter is chain shift: phonological scenarios where A=>B, B=>C etc. The data come from Finnish, where long /aa/ shortens before a suffix [i], but /a/ becomes round [o]. Long /oo/ also shortens, but /o/ remains unaltered:

(1) Finnish chain shift

Singular nominative	Plural essive	
a) *maa*	*ma-ina*	'earth'
b) *kissa*	*kisso-ina*	'cat'
c) *talo*	*talo-ina*	'house'

The main claim of PC theory is that there are constraints that argue in favor of preserving contrast. For the Finnish data, it is claimed that such a constraint will prevent [aa] and [a] from neutralizing by rounding the latter, but not the former, to [o]. To illustrate what is ruled out by such a constraint, Łubowicz compares the actual state-of-affairs to other imaginable situations, for example </aai/=>[ai], /ai/=>[ai]/>, where neutralization only concerns length. Such possibilities are called "scenarios".

The set of possible scenarios is the set of candidates that enter the evaluation process. Scenarios are evaluated by a "family" of ranked contrast constraints. These constraints can be specified to apply to the input or to the output, and may be sensitive to particular aspects of the input or output, such as length or rounding.

Let us consider the analysis. First, one must assume that all PC constraints are ranked below a constraint against preserving the underlying length, say $*\sigma_{\mu\mu\mu}$, banning three morae in the same syllable (presumably, /aai/ must syllabify under one syllable, because of higher ranked constrained). The scenario that neutralizes /aa/ and /a/ violates a contrast preservation constraint on the output, which requires that length distinctions be preserved. The scenario that rounds the underlyingly short /a/ does not violate this constraint, because the originally short vowel is rounded. This scenario does neutralize short /o/ and /a/, but the requirement that there be contrast in rounding is less important, i.e. ranked lower. The ranking, I suspect, is

established in order to get the right answer: “For the actual scenario to win, it must be more important to improve on the distribution of length neutralization in a scenario […] than to avoid merging rounding (p.36).” I leave it to the reader to decide whether this type of reasoning is acceptable.

OT accounts that do not integrate contrast as a principle use faithfulness constraints. These are always violated whenever there is a change. In contrast, PC constraints allow for contrast transformation, that is, they are satisfied when a given underlying contrast is transformed into a different surface contrast. The author claims that this is an advantage of PC “theory” over the more classic use of faithfulness constrains to preserve contrast.

The rest of the chapter is devoted to exploring the predictions that the PC theory makes. The factorial typology is explored, and it is claimed that integrating contrast as a formal requirement predicts certain attested chain shifts and excludes unattested ones.

Chapter 3 applies the same theory to the problem of stress in different Arabic dialects and its interaction with epenthesis. The problem is the following. Arabic stress is weight-based, so non-final closed or long syllables receive stress, and so do final heavy syllables. In the dialects considered here, if no heavy syllable exists, then the antepenultimate is stressed.

All dialects have a process of epenthesis. In some dialects (Syrian, 2a), the epenthetic vowel is always ignored by stress; in others, it is treated like a non-epenthetic vowel (Omani, 2b); a third group of dialects exhibit “hybrid behavior”, with epenthetic vowels taken into account only word-medially (Iraqi,2c).

(2) Stress in Arabic dialects

	Without epenthesis		**With epenthesis**		**Comment**
a.	*darábna*	‘we hit’	*ʔákəlton*	‘their meat’	[ə] epenthetic, ignored
	fátaħ/e/t	‘she opened’	*fatáħət*	‘I opened’	[ə] epentheti c, ignored
b.	*waládhum*	‘their child’	*ʔakílhum*	‘their food’	í epenthetic, yet counted
	mánzila	‘status’	*saráq laħin*	‘stole a tune’	í epenthetic, yet counted
c.	*sallátha*	‘her basket’	*ʔíbínha*	‘her son’	[í] epenthetic, yet stressed
d.	*ʃárika*	‘firm, company’	*kitábit*	‘I wrote’	[i] epenthetic , ignored

The chapter treats this inter-dialectal distribution from the point of view of the contrast between vowels which are included in the underlying representation and ones that aren’t, viz. epenthetic vowels. Recall that PC constraints admit cases where a given contrast in the underlying representation is expressed in another fashion on the surface. In this chapter, it is claimed that some Arabic dialects

maintain the underlying difference between the two types of vowels by treating them differently in the calculation of stress.

As is well-known, stress-assignment in Arabic is based on syllable weight. This process is summarized in one constraint, WEIGHTBYPOSITION (WBP), which assigns morae to codas. This constraint is ranked with respect to a constraint demanding that "inputs that are distinct in the presence/absence of a vowel […] need to remain distinct in the output". In interaction with other stress-related constraints, the two possible rankings of these two constraints yield the completely neutralizing and the contrast-preserving dialects. The third, hybrid type of dialect is accounted for using a constraint MORACONTINUITY. This constraint requires that in any CVCCV, the second C must be moraic. Its effect is to cancel the effect of the contrast, but only in word-medial positions. Once again, the choice of constraints is dictated by the data, and for every piece of data that is not covered by the analysis, a constraint is matched to it and accounts for it.

Finally, it is argued that this analysis makes a correct prediction with respect to the interaction of contrast and the placement - rather than the appearance - of the epenthetic vowel. For the same underlying /VCCCV/ clusters, some Arabic dialects epenthesize to create a new coda [VC**ə**CC], whereas others create a new onset [CCəC]. Those that create an onset are not expected, according to the account, to distinguish between epenthetic and non-epenthetic vowels for the purpose of stress assignment. This is apparently true.

As in the case of chain shifts, the analysis here left me unimpressed. As is so common in OT analyses, the use and formalization of constraints is extremely unconstrained, and thus harmful to the validity of the analysis. In other words, it is not clear whether there is anything that this type of analysis can**not** explain… For instance, apparently there is no dialect with contrast preserved only in non-final positions. Is such a dialect predicted to be impossible for the theory? Or is it the case that if such a dialect existed, some constraint could easily be recruited to be ranked above the contrast constraint and produce the desired effect? Moreover, the prediction that the analysis claims to make does not seem unique to it: since Arabic relates stress to weight, and since onsets are not moraic, it is clear that newly-formed onsets will not affect stress, regardless of the specific analysis. Finally, cognitively, I find it hard to believe that there is a requirement to distinguish between real and epenthetic vowel. The fact that if there were such a requirement, it could be used to explain the different behavior of the two types of vowels - this fact seems to me insufficient as evidence for its existence.

Chapter 4 applies the same tools to the selection of allomorphs. In Polish, there is a process of coronal palatalization before front vowels. For instance, when *list* 'letter.SG.NOM' is followed by the locative singular suffix [e], it is pronounced lis[če]. The resulting palatal is also a phoneme of the language, e.g. *liščˇ* 'leaf.SG.NOM'. One would expect the locative form of 'leaf' to be liš[če], exactly like the word for 'letter'. Instead, however, the vowel of the suffix is [u], and the locative form is liš[ču].

The chapter analyzes this set of data in the following manner: there are two possible allomorphs for the singular locative: [e] and [u]. If [e] were chosen for both *list* and *lišč*, the underlying contrast between /t/ and /č/ would disappear. But if [u] is used with underlying /č/, then the underlying difference will be detectible through the choice of that allomorph, and contrast will not be lost.

For this to work, something has to assure that the allomorph [u] is not selected for both nouns, in which case the original contrast will be preserved: /list+u/ = [lis**t**u], and /lisč+u/ = [lis**č**u]. The allomorph [e] has to be given independent priority over the allomorph [u], so that when both are possible, [e] will be preferred. In order to achieve this, the author concocts a constraint *ALVEOLAR/U, banning [u] after alveolar consonants. This constraint is ranked below the contrast constraint, so that it is active only if the contrast is preserved, but at the expense of maintaining the [e]~[u] allomorphy.

Allomorph-selection is commonly treated as phonological optimization (at least when such a treatment makes more or less intuitive sense). To that logic, PC "theory" adds one other aspect of being optimal, which is more morphological than phonological: maintaining contrast. Setting aside the ad-hoc feel of the solution in the chapter, the argument is clear.

That said, a possible objection comes from the phenomenon of syncretism, that is, morphological distinctions that are not reflected in the form of the item. For instance, the 1st and 3nd person suffixes in Yiddish are distinct in the singular, but identical in the plural. For every such case of syncretism – and there are quite a few – one would have to assume that creating contrast is more costly than violating PC.

Summary of the review. This book argues for the usefulness of admitting contrast as an active principle in one's theory of grammar. The test-cases come from three areas of phonological research: chain shifts, stress-epenthesis interaction and allomorph selection. The arguments are presented in a clear fashion, although the theoretical moves are at times ad-hoc. The advantages of the contrast-based approach on other approaches are not always so obvious, further weakening the ability of the proposal to really convince the reader. That said, the book constitutes a positive, sincere attempt to formalize contrast in OT.

A general question that emerges from this formalization attempt is whether it is legitimate: if contrast is important, why are there so many neutralizations? And why so much syncretism? A principled answer is not provided in this book.

Anne Cutler (2012)
Native Listening: Language Experience and the Recognition of Spoken Words

Cambridge, Mass: MIT Press (xvii + 555 pp., Phonetic Appendix, Notes, References, Name Index, Subject Index.
ISBN: 9780262017565, Price: $50.-)

Reviewed by: Judith Rosenhouse
Swantech Ltd., Haifa 326842 Israel
e-mail: swantech@013.net.il

Professor Anne Cutler, Emerita Director of Max-Planck-Institut für Psycholinguistik, Nijmegen, The Netherlands, and now full-time Research Professor at The MARCS Institute, University of Western Sydney, Australia, is a world renowned psycholinguistics researcher. This masterpiece of a book sums up much of the work of the last four decades – work done by Prof. Cutler (as noted in the Preface), as well as by numerous other researchers.

The book studies spoken language – a communication system of auditory signals – and gradually creates a coherent picture of human language and its functioning. All this is included in 12 chapters dealing with (1) first language (L1) listening, and (2) bilingual listening (L1 + another language). The chapters are followed by several technical lists: a Phonetic Appendix presenting the categories of consonants and vowels, as well as their phonetic symbols (pp. 451-454), Notes for of all the chapters (pp. 455-458), an impressive list of References (pp. 459-532), a Name Index (pp.533-548), and a Subject Index (pp.549-555). As in many text books, important points are clarified and demonstrated by figures, tables and framed blocks. Thus, this book is suitable for students and scholars of at least phonetics, psycholinguistics and speech synthesis.

Listening is more than hearing sounds and speech sounds: it involves analysis and understanding. It is therefore clear that the physiological and acoustical features of speech come together in the brain which analyzes, integrates and understands (categorizes) the incoming sounds. Chapter 1 “Listening and native language” (pp. 1-32) is therefore concerned with listening to a native language, and how universal this process is. Other issues that are discussed in this chapter (based on psychoacoustic experiments) are, e.g., word recognition and reconstructing, phonetic effects of vowel or consonant contexts, language specificity, and answering the question “what would life be like if we only had one language?” (pp. 30-32).

Chapter 2 “What is spoken language like?” (pp. 33-72) describes general language features, considering that language is fast, continuous, variable and non-unique. The sections deal with ambiguous onsets, within-word and cross-word embeddings, lexical statistics of stress, and other features of the lexicon, e.g., word categories and implications of different consonants and vowels. Also included are

sections on lexical entries, morphological structure, closed and open classes and word frequency effects on lexical processing. The last section in this chapter concludes by claiming that vocabularies guide the operation of spoken word recognition.

Chapter 3 "Words: How they are recognized" (pp. 73-116) focuses on word recognition processes. Various experimental research methods are reviewed (e.g. cross modal priming, eye tracking), and the findings are modeled related to concurrent recognition alternatives and the competition between them, phonological and conceptual representation of lexical items and the differences between such items under different presentation contexts (separately, within a sentence, with/without contrastive accent, etc.). A separate section deals with lexical tone (of Cantonese, for example) and durational structure in lexical activation. Morphology is again addressed in gender marking contexts (in cases such as *le bouton* 'the button' vs. *la bouteille* 'the bottle' vs. *les boutons / bouteilles* 'the buttons / bottles'). Many experiments have produced the models of lexical recognition and decision.

Chapter 4 "Words: How they are extracted from speech" (pp. 117-153) continues with word recognition, but now focus on their use within sentences. Word boundaries can be recognized and used more efficiently in order to understand the words. Speech rhythm (expressed also by stress, e.g., in English) is widely used for this goal, along with phonotactic likelihood, to determine word boundaries across languages. This aspect is considered for several languages, with different stress systems, including stress-less systems, such as French. Along with stress, syllable structures also differ between languages, and they are discussed in this context. The author suggests then a rhythmic segmentation hypothesis which has also been examined, as well as phonotactic cues (language-specific limitations on phoneme adjacency occurrences). This chapter then moves on to artificial language learning, by humans and machines, and the problems and results they involve. Thus, phonological/phonetic segmentation has been considered as a basic element in listening for language cues; and boundary cues, which differ from language to language, also make speech segmentation a language specific process.

Chapter 5 "Words: How impossible ones are ruled out?" (pp. 155-189) refers to the opposite analysis in listening: ruling out optional words, or Possible Word Context (PWC). The experiments examined word contexts in which certain words were embedded. Differences were found between the responses of human participants and the computer program model. The next question was whether PWC was universal, since it is based on language specific constraints. Some additional experiments, with an African language for example, yielded the same results as for English, which suggests that PWC may indeed be universal. Some languages however have one-consonant words (v 'in,' 'k' 'to' in Slovak) or devoiced vowels (Japanese) which yield quite complex findings for PWC experiments. In such cases, morphological or syntactic considerations get involved in the PWC processes. But the constraint that PWC involves is broken down only in languages which have vowel-less words, such as the Slovak example. The PWC is also a useful tool in

infants' language acquisition, as discussed in the chapter making PWC almost universal.

Chapter 6 and Chapter 7 are perhaps the most relevant chapters to phonetics-minded readers. Chapter 6 "What is spoken language like? Part 2: The fine structure of speech" (pp. 191-226) describes phonetic systems by features, such as predictable/unpredictable variation, assimilation and regressive assimilation, liaison between words due to personal differences or language specific features, segment deletion, fine phonetic features, such as VOT, sub-phonemic variations at the beginning or end of words, etc. All of these concurrent variations exist in speech and participate in word recognition. The chapter ends with commenting on spontaneous speech as being the most natural manner for listeners (compared to read, rehearsed or synthesized speech), in spite of the many ambiguities it involves.

Chapter 7 "Prosody" (pp. 227-258) again starts with stress structure and pitch contours. The discussion continues with inter-language differences based on prosodic contexts and cues (short vs. long sentences and their differences as demonstrated by expected prosodic cues). The cues on which listeners rely are duration, pitch and intensity, which vary the segments with their position in the utterance. But it has been found that not all cues are used by listeners: some cues are apparently more informative than others (if they indicate word boundaries, for example) and some are driven by language-specific phonetic features. The next question again considers whether prosody is universal or not, but the chapter concludes by observing that there are still many "untrodden paths" and much to study in the study of prosody and its universal aspects.

From the next chapter on, the book delves into listening issues related to languages beyond the mother tongue. Chapter 8. "Where does language specificity begin?" (pp. 259-301) considers the details of how infants begin accumulating their knowledge of their L1 in the first year and a half of their lives. This chapter describes research methods developed for infants – even from their fetal stage – which results in infants' skills of discrimination, preference and recognition in their L1. These methods use high-amplitude sucking, visual fixation, head turn and looking responses, for example. While still in the uterus, infants seem to amass prosodic information of rhythm and intonation of their L1 because of low-pass filtered sounds. After the first half year of life, babies begin to become language-specific listeners by discriminating L1 elements from other languages. Incoming input of course helps enhance this process, and research describes the various communication situations an infant of 6-9 months can receive each day, including speech directed to the infant or to other people in her/his environment. Importantly, a child hardly hears single words: less than 9% of all the sequences collected per day which the child hears (in van de Weijer, 1999) include isolated words, or 13.3% of the utterances specifically directed to the child. The chapter goes on to review phonemic cues that help infants in acquiring the language: vowels and consonants that are enhanced in infant-directed speech more than in adult-directed speech (e.g., /i, u, a, s, ʃ/, certain segments occurring within or beyond word boundaries, etc). Yet

speech to somewhat older children abounds in phonotactic processes, such as elision and other reductions. Speaking the first words requires additional processes of discriminating between familiar (L1) and unfamiliar (from another language) words. These studies further explore determinants of segmentation, the statistics of segmentation as a universal cue, and differences between open and closed lexical classes as universal cues. The chapter then moves to the perception and form of the first word, its relevant features, and what happens in bilingual input. In the end, Cutler writes that languages train the listeners: speech has universal features, but languages have specific features which human users have to learn.

Chapter 9 “Second language listening: Sounds to words” (pp. 303-335) analyzes some of the differences between first and second language listening. Adult listeners use basically the same learning system as that for L1, which requires distinguishing minimal inter-word contrasts, activation of words from memory, segmentation of continuous speech into its component words and constructing sentences from the resulting words. But the L1 specific system is not well adapted to an L2 due to different language-specific features. Six possible unexpected difficulties which relate to the above requirements are described (p. 304). Problems include familiar phonetic contrasts in unfamiliar positions, effects of category goodness differences, and pseudo-homophones in L2 lexical activation and competition. The reported research exposes the differences which make it harder to listen to a new language. These difficulties are expressed in prolonged ambiguity judgments in lexical activation and competition, as shown by lexical statistics. Examples are taken from experiments in, e.g., French, Japanese and Dutch listeners tested in English L2. The conclusion emphasizes that the ability to perceive an L2 contrast (of phonemes, for example) does not necessarily mean that this ability will be correctly deployed to discriminate words, and inability to perceive a contrast does not necessarily rule out accurate encoding of the contrast in lexical entries. This asymmetry can be hard to get rid of, which is one of the properties that make L2 listening harder than L1 listening.

Chapter 10 “Second language listening: Words in their speech contexts” (pp. 337 374). Since speech is not limited to segments, difficulties in segment recognition in L2 continuous speech carry over to larger speech units, such as phrases and sentences. This chapter analyzes such problems in areas of perceived speech rate (which is harder in L2 than in L1), L2 rhythm and L2 segmentation, phonotactic processes (e.g. French liaison), casual speech processes (e.g. British English final added /r/ in words such as ‘idea’), idiom processing and prosody processing in L2. Idioms are analyzed with special difficulty in L2, because literal understanding (shown by translation) often reveals that prosodic cues that help L1 listeners may pass unnoticed by L2 listeners. Other higher level processing difficulties for L2 listeners involve fast speech, speech in noise and voice recognition in L2 vs. L1, to which considerable attention is paid in this chapter. Yet some tests show that certain L2 features can be acquired and used by L2 listeners. But some of these features require L2 listeners who are very highly proficient at their L2; and curiously, such

proficient L2 listeners may end up (in other tests) performing in their L1 like an L2 listener. The case of early bilinguals is also discussed in this context; even early bilingual listeners reveal asymmetric use of their two languages, as found in some lexical tests. The chapter ends with the conclusion that both universal and language-specific features are to be found in L1 and L2 listeners, as well as early bilinguals, and that there are infinite gradations in the continuum of language use. The main question that remains at the end of this chapter is how much language experience is "enough" for learning to avoid an inefficient procedure (of L1, mainly) when using an L2.

Chapter 11, "The plasticity of adult speech perception" (pp. 375-409), deals with the dynamic aspects of language skills. Several such factors exist. Human listeners adapt to the speech of others in spite of differences in language, phonetic context, speech rate, dialect, accent, vocal features, etc. Human adaptation to such factors requires some time and effort, but the desired communication is usually possible. This involves perceptual learning, the author writes, which is personal and at times long lasting, i.e., generalizes beyond the learning situation (though such adaptation can also be un-learned). But all of these processes are better performed in the L1 than in any later learned language. Language change processes over time, even in a speaker's lifetime, have shown differences in vowel pronunciation (the author mentions changes in Queen Elizabeth's vowels); vowel durations that are ignored by listeners of certain dialects vs. other dialects (e.g., speakers of Standard French vs. Swiss French); and flapping used more than palatalization (by British long-time residents of the USA). Contrasts that often involve mismatching affect word recognition between different dialects (e.g., 'look, luck, Luke' have three different vowels in most of Standard British English, but only two vowels in the Yorkshire dialect.) Such differences affect L1 listeners in foreign language perception more than in different dialect perception, as speech-with-noise experiments demonstrate. Dialect differences also appear in suprasegmental features, as well as in foreign-accented L1. Such features, which involve phonemic categories and their categorical perception, show "flexibility," as Repp and Liberman (1987) described them. Adaptation to a new pronunciation may be induced by learning of lexical items, even in artificial experiments of learning (such as artificial words with /f/ instead of /s/), but this learning "wears off" under various conditions, including time since learning. All of these processes have their limits, however, which depend on the cues used by the listener. More plasticity or flexibility has been shown in the language skills of bilinguals than in monolinguals at all ages, even for infants. Bilinguals, or people who acquired the other language later in life, maintain various cognitive activities which are lost in monolinguals, and aged bilinguals reveal delayed cognitive aging symptoms compared to monolinguals. Thus, early exposure to more than one language, even for sign language acquisition, is important. At the end of this chapter we read that adaptability and generalization are key elements of human cognition, and speech processing benefits from both of them. Thus both L1

and L2 have their specific features, but L1 is characterized by greater robustness and apparently unlimited flexibility.

Chapter 12 “Conclusion: The architecture of a native listening system” (pp. 411-449) is not just a conclusion: it gives a general look at listening and its features. In spite of differences of inter-language phonemic systems, all listeners begin at the same point. Thus, universal and language-specific aspects intertwine. In addition, abstract representations of pre-lexical, phoneme sequence probabilities and lexical speech processing intertwine with episodic or incidental details. Other summarized processes are word forms and meanings and phonological representations. An example is the case of a Japanese mora which presents a rhythmic category and its role in speech processing. Identifying a word with a mora is facilitated if the part that is used in the tested non-word is a mora part rather than when it is a whole mora; but other tests showed that, as in other languages, Japanese word recognition utilizes phonemes and not morae. The author concludes this section by saying that rhythmic units do not operate as intermediate levels of representation in listening – just as assumed by cascade models which hypothesize that “the relation between the prelexical and lexical processing levels involves no irretrievable commitment to categorical decision but rather it is probabilistic, with the weighting of probabilities going up and down as the input alters” (p. 426). Another possible model is the Merge model (Norris et al., 2000), which merges both pre-lexical and lexical processes, according to task requirements. The relatively long concluding section of the chapter sums up the changes in the development of psycholinguistics since the 1960s-1970s. The main difference is that from concentration on language universals only, language specificity has been realized as yielding more fruitful results. The author also lists several topics for future research, based on the current state of the art, because there are many gaps in current knowledge. But in sum, “it is clear that spoken word recognition research has come far in the few decades of its existence.” (p. 449).

This book is a great contribution to the field of psycholinguistics, as it presents the basics as well as more in-depth developments in methodology and theories, in a readable and clear style for both students and professionals. As it involves experiments of phonetic elements tested in the laboratory and many considerations of live speech and listening, it is an important addition to our book shelf.

References

Norris, DM., JM. McQueen, and A. Cutler 2000. Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23: 299-325.

Repp, BH., and AM. Liberman 1987. Phonetic category boundaries are flexible. In SR. Harnad (ed.): *Categorical Perception*, Cambridge: Cambridge University Press, 89-112.

Weijer, J. van de 1999. *Language Input for Word Discovery*. PhD dissertation, MPI Series in Psycholinguistics 9, University of Nijmegen

CALL FOR PAPERS

The *Phonetician* will publish peer-reviewed papers and short articles in all areas of speech science including articulatory and acoustic phonetics, speech production and perception, speech synthesis, speech technology, applied phonetics, psycholinguistics, sociophonetics, history of phonetics, etc. Contributions should primarily focus on experimental work but theoretical and methodological papers will also be considered. Papers should be original works that have not been published and are not being considered for publication elsewhere.

Authors should follow the *Journal of Phonetics* guidelines for the preparation of their manuscripts. Manuscripts will be reviewed anonymously by two experts in phonetics. The title page should include the authors' names and affiliations, address, e-mail, telephone, and fax numbers. Manuscripts should include an abstract of no more than 150 words and up to four keywords. The final version of the manuscript should be sent both in .doc and in .pdf files. It is the authors' responsibility to obtain written permission to reproduce copyright material.

All kinds of manuscripts should be sent in electronic form (.doc and .pdf) to the Editor. We encourage our colleagues to send manuscripts for our newly released section entitled MA research, which is a summary of the student's phonetics research describing their motivation, topic, goal, and results (no more than 1,200 words).

INSTRUCTIONS FOR BOOK REVIEWERS

Reviews in the *Phonetician* are dedicated to books related to phonetics and phonology. Usually the editor contacts prospective reviewers. Readers who wish to review a book mentioned in the list of "Publications Received" or any other book, should address the editor about it.

A review should begin with the author's surname and name, publication date, the book title and subtitle, publication place, publishers, ISBN numbers, price, page numbers, and other relevant information such as number of indexes, tables, or figures. The reviewer's name, surname, and address should follow "Reviewed by" in a new line.

The review should be factual and descriptive rather than interpretive, unless reviewers can relate a theory or other information to the book which could benefit our readers. Review length usually ranges between 700 and 2500 words. All reviews should be sent in electronic form to Prof. Judith Rosenhouse (e-mail: swantech@013.net).

ISPhS MEMBERSHIP APPLICATION FORM

Please mail the completed form to:

Treasurer:
Prof. Dr. Ruth Huntley Bahr, Ph.D.
Treasurer's Office:
Dept. of Communication Sciences and Disorders
4202 E. Fowler Ave. PCD 1017
University of South Florida
Tampa, FL 33620 USA

I wish to become a member of the International Society of Phonetic Sciences

Title: _____ Last Name: _____ First Name: _____
Company/Institution: _____
Full mailing address: _____
_____
Phone: _____ Fax: _____
E-mail: _____
Education degrees: _____
Area(s) of interest: _____

The Membership Fee Schedule (check one):

1. Members (Officers, Fellows, Regular)	$ 30.00 per year
2. Student Members	$ 10.000 per year
3. Emeritus Members	NO CHARGE
4. Affiliate (Corporate) Members	$ 60.000 per year
5. Libraries (plus overseas airmail postage)	$ 32.000 per year
6. Sustaining Members	$ 75.000 per year
7. Sponsors	$ 150.000 per year
8. Patrons	$ 300.000 per year
9. Institutional/Instructional Members	$ 750.000 per year

Go online at www.isphs.org and pay your dues via PayPal using your credit card.

☐ I have enclosed a cheque (in US $ only), made payable to ISPhS.

Date _____ Full Signature _____

Students should provide a copy of their student card

NEWS ON DUES

Your dues should be paid as soon as it convenient for you to do so. Please send them directly to the Treasurer:

Prof. Ruth Huntley Bahr, Ph.D.
Dept. of Communication Sciences & Disorders
4202 E. Fowler Ave., PCD 1017
University of South Florida
Tampa, FL 33620-8200 USA
Tel.: +1.813.974.3182, Fax: +1.813.974.0822
e-mail: rbahr@ usf.edu

VISA and MASTERCARD: You now have the option to pay your ISPhS membership dues by VISA or MASTERCARD using PayPal. Please visit our website, www.isphs.org, and click on the Membership tab and look under Dues for "paid online via PayPal." Click on this phrase and you will be directed to PayPal.

The Fee Schedule:

1. Members (Officers, Fellows, Regular)	$ 30.00 per year
2. Student Members	$ 10.00 per year
3. Emeritus Members	NO CHARGE
4. Affiliate (Corporate) Members	$ 60.00 per year
5. Libraries (plus overseas airmail postage)	$ 32.00 per year
6. Sustaining Members	$ 75.00 per year
7. Sponsors	$ 150.00 per year
8. Patrons	$ 300.00 per year
9. Institutional/Instructional Members	$ 750.00 per year

Special members (categories 6–9) will receive certificates; Patrons and Institutional members will receive plaques, and Affiliate members will be permitted to appoint/elect members to the Council of Representatives (two each national groups; one each for other organizations).

Libraries: Please encourage your library to subscribe to *The Phonetician*. Library subscriptions are quite modest – and they aid us in funding our mailings to phoneticians in Third World Countries.

Life members: Based on the request of several members, the Board of Directors has approved the following rates for **Life Membership** in ISPhS:

Age 60 or older:	$ 150.00
Age 50–60:	$ 250.00
Younger than 50 years:	$ 450.00