On Speech and Language

Studies for Sieb G. Nooteboom

On Speech and Language

Studies for Sieb G. Nooteboom

edited by

Hugo Quené and Vincent van Heuven

Netherlands Graduate School of Linguistics
Utrecht

Published by:

Netherlands Graduate School of Linguistics (LOT)
Trans 10, 3512 JK Utrecht, The Netherlands
E: lot@let.uu.nl
W: http://wwwlot.let.uu.nl
T: +31 30 253 6006

ISBN: 90-76864-53-5
NUR 632

Cover illustration: *Niña Sonriente*, detail.
Glass statue after an original design by Juan Ripollès (Spain, 1932),
created by the Berengo Fine Arts studio in Murano (Italy).
Photo by Wieke Eefting.

Table of Contents

List of Contributors

Joan Baart
SIL International
(joan_baart@sil.org)

Gerrit Bloothooft
Utrecht institute of Linguistics OTS, Utrecht University
Trans 10, 3512 JK Utrecht, The Netherlands
(gerrit.bloothooft@let.uu.nl)

Johanneke Caspers
Universiteit Leiden Centre for Linguistics, Universiteit Leiden
PO Box 9515, 2300 RA Leiden, The Netherlands
(V.J.J.P.van.Heuven@Let.LeidenUniv.nl)

Anne Cutler
Max-Planck-Institut für Psycholinguistik
PO Box 310, 6500 AH Nijmegen, The Netherlands
(Anne.Cutler@mpi.nl)

Gary S. Dell
Department of Psychology, University of Illinois at Urbana-Champaign
603 E Daniel Street, Champaign, IL 61820, U.S.A.
(gdell@cyrus.psych.uiuc.edu)

Arthur Dirksen
Fluency
Prins Hendrikkade 159a, 1011 TB Amsterdam, The Netherlands
(info@fluency.nl)

Carlos Gussenhoven
Center for Language Studies, Radboud University Nijmegen
PO Box 9103, 6500 HD Nijmegen, The Netherlands
(c.gussenhoven@let.ru.nl)

Willemijn Heeren
Utrecht institute of Linguistics OTS, Utrecht University
Trans 10, 3512 JK Utrecht, The Netherlands
(willemijn.heeren@let.uu.nl)

Caroline G. Henton

University of California,
1156 High Street, Santa Cruz, CA 95064, U.S.A.

Vincent J. van Heuven

Universiteit Leiden Centre for Linguistics, Universiteit Leiden
PO Box 9515, 2300 RA Leiden, The Netherlands
(V.J.J.P.van.Heuven@Let.LeidenUniv.nl)

Heleen Hoekstra

Utrecht institute of Linguistics OTS, Utrecht University
Trans 10, 3512 JK Utrecht, The Netherlands
(heleen.hoekstra@let.uu.nl)

Esther Janse

Utrecht institute of Linguistics OTS, Utrecht University
Trans 10, 3512 JK Utrecht, The Netherlands
(esther.janse@let.uu.nl)

Joop Kerkhoff

Center for Language Studies, Radboud University Nijmegen
PO Box 9103, 6500 HD Nijmegen, The Netherlands
(j.kerkhoff@let.ru.nl)

Truus Kruyt

Institute for Dutch Lexicology INL,
PO Box 9515, 2300 RA Leiden, The Netherlands
(Kruyt@inl.nl)

Bob Ladd

Department of Linguistics, University of Edinburgh
40 George Square, Edinburgh EH8 9LL, United Kingdom
(bob@ling.ed.ac.uk)

John Ohala

Department of Linguistics, University of California
1203 Dwinelle Hall, Berkeley, CA 94720-2650, U.S.A.
(ohala@socrates.berkeley.edu)

Peter Pabon

Utrecht institute of Linguistics OTS, Utrecht University
Trans 10, 3512 JK Utrecht, The Netherlands
(peter.pabon@let.uu.nl)

Louis Pols

Amsterdam Center for Language and Communication, University of Amsterdam
Herengracht 338, 1016 CG Amsterdam, The Netherlands
(L.C.W.Pols@uva.nl)

Hugo Quené

Utrecht institute of Linguistics OTS, Utrecht University
Trans 10, 3512 JK Utrecht, The Netherlands
(hugo.quene@let.uu.nl)

Toni Rietveld

Center for Language Studies, Radboud University Nijmegen
PO Box 9103, 6500 HD Nijmegen, The Netherlands
(a.rietveld@let.ru.nl)

Jill A. Warker

Department of Psychology, University of Illinois at Urbana-Champaign
603 E Daniel Street, Champaign, IL 61820, U.S.A.
(warker@express.cites.uiuc.edu)

Frank Wijnen

Utrecht institute of Linguistics OTS, Utrecht University
Trans 10, 3512 JK Utrecht, The Netherlands
(frank.wijnen@let.uu.nl)

Introduction

Vincent J. van Heuven
Universiteit Leiden
and
Hugo Quené
Utrecht University

Siebout Govaert Nooteboom (Sieb, [sip]) was born 65 years ago (19 April, 1939) in Makassar on the isle of Celebes (now Sulawesi), in the former Dutch East Indies, which is now part of the Republik Indonesia. Nooteboom obtained obtained a BA degree (kandidaats-examen, in 1963) in Dutch Language and Literature at Leiden University, and then specialised in General Linguistics at the same university (MA degree, doctoraal-examen, in 1966). In 1972 Nooteboom defended his doctoral dissertation entitled *Production and Perception of Vowel Duration in Dutch: A study of durational properties of vowels in Dutch*. The dissertation was defended cum laude at Utrecht University, with Antonie Cohen as the primary supervisor; the research leading up to the dissertation was done at the Institute for Perception Research (IPO) at Eindhoven. This dissertation, and the articles based on it which appeared in the international journals, presented ground-breaking work and established Nooteboom's name as a player in the international field of the discipline.

From 1966 until 1988 Nooteboom was a researcher at IPO, but he held a part-time lectureship in Phonetics at Leiden University from 1968 until 1981, when it was upgraded to a professorship. From 1986 until 1988 Nooteboom was also a part-time professor of Experimental Linguistics at Eindhoven University. In 1988 Nooteboom gave up both part-time professorships, as well as his position in the Institute for Perception research (co-leader of the hearing and speech group from 1977 until 1988), in order to succeed Antonie Cohen as the professor of Phonetics at Utrecht University, which position he held until this 65th birthday. Since then Sieb continues to be employed by Utrecht University, albeit in a part-time position. Nooteboom is the only phonetician ever to have held three professorships at three different universities in the Netherlands. In the course of these professorships, Nooteboom has been the formal promotor to 27 young phoneticians in Leiden (7), Eindhoven (8) Utrecht (11) and even Groningen (1). In the Dutch academic system only full professors can be promotors. There were quite a few more doctoral candidates who were supervised by Nooteboom in the earlier days of his career, but Sieb does not list these as his own.

The present volume commemorates Sieb Nooteboom's career as a leading phonetician with a collection of short articles by his former and current PhD candidates, his co-workers at the Department of Phonetics at Utrecht University, by other professors of Phonetics (or equivalent) in the Netherlands and by a few colleagues from abroad who in one way or another interacted significantly with Nooteboom. Within the large group of PhD graduates we approached only those who are still active in phonetics. (A significant number left the discipline after they obtained the doctorate, and found jobs in the electronics and computer industry, in governmental agencies, in hospitals and audiology clinics, or in engineering departments at technical universities). Regrettably, a few prospective authors were unable to contribute to this volume, due to the strict time constraints imposed by the editors.

Pervasive in Nooteboom's approach to the discipline is his conviction that there is more to phonetics than just the description of vowels and consonants. Speech consists of vowel and consonant segments, but these are enriched with a temporal and melodic organisation that reflects higher-order linguistic processes. In other words, Nooteboom believes that speech can be fruitfully studied only in close relationship with the linguistic code of which it is a manifestation. Nooteboom has worked on several areas of study, always following this general approach. The authors in this volume further explore five of these areas, along the lines set out by Nooteboom's research in each area.

The first area concerns the issue of planning and programming during speech production. This includes the study of speech errors (Cutler & Henton, Dell & Warker) and the amount of planning, anticipation and look-ahead that becomes apparent from the melodic organisation of speech (Van Heuven, Hoekstra).

A second, related area is the study of intonation and focus. This includes the relation between linguistic tone and musical pitch (Baart), the discourse functions of certain pitch movements (Caspers), and the temporal alignment of pitch movements in speech (Ladd, Rietveld & Kerkhoff).

The third area is that of Nooteboom's dissertation: the temporal organisation of speech. This includes the study of factors affecting (not produced but) perceived vowel duration (Gussenhoven), as well as an investigation into the effect of "silent prosody" during silent reading (Wijnen).

Although perceptual processes are relevant for all areas, we distinguish speech perception as a fourth area of investigation, at least for the purpose of this introduction. Studies in this area are concerned with speech perception per se: lower-level psychophysical processes (Quené), higher-order linguistic influences on phonetic processing (Heeren), and problems in spoken-word perception (Janse & Quené).

The fifth area is that of technological applications: for speech synthesis (Dirksen, Kruyt) and for studying the singing voice (Bloothooft & Pabon).

Finally, covering all these areas are two more general reflections on the relation between phonetics and its related disciplines (Ohala, Pols).

Sieb was — and still is, we assume — a firm believer in the indispensable relationship between fundamental research and practical application. He has always argued that technological applications are a challenge to fundamental research: they are the acid test to what we think we know. If we really know how speech is produced by the speaker and processed by the human listener, then we should be able to build machines that simulate these processes, in such a way that the result is indistinguishable from the human process. In so far as practical applications fail, we learn where our fundamental knowledge is insufficient and where further research is needed. (This view is reflected, of course, in the fifth area identified above). It should come as no surprise, then, that in 1985 Nooteboom was among the founding fathers of the Netherlands Speech Technology Foundation. This foundation was a consortium of the Phonetics research groups in the Netherlands. It secured a large grant from the Netherlands Ministry of Economic Affairs. The research programme 'Analysis and Synthesis of Speech' involved virtually all the tenured faculty of the phonetics research groups at the universities of Amsterdam, Utrecht, Leiden, Nijmegen and Eindhoven, as well as some 20 full-time employed paid PhD graduates and postdocs. The coordinator of the national programme was Antonie Cohen, and after his retirement in 1987 this task was taken over by Nooteboom, who successfully completed the program by 1991.

It is important to realise that Nooteboom is a linguist by training, and believes in strong integration between phonetics and linguistics, between speech and language. Both in Leiden and in Utrecht the phonetics research groups are part of the department of linguistics. As a corollary of this view, speech and language technology should never be separated. It comes as no surprise, then, to see that Nooteboom also took an active role in the organisation of language technology. At present Nooteboom is the chairperson of both the Speech Technology Foundation and of the Language Technology Foundation.

Sieb Nooteboom has always taken a keen interest in the teaching of phonetics and linguistics. A major effort concerned the setting up of a teaching program for the training of phoneticians at the graduate level. At each university within the Netherlands, the number of PhD graduates in phonetics is too small to warrant an independent teaching programme. By pooling students and teaching resources on a national scale, however, an internationally competitive PhD programme could be developed. Nooteboom has played a key role here as the first director of the Netherlands Graduate School in linguistics (and as a board member of the graduate school ever since).

A preliminary version of this Festschrift was presented to Sieb Nooteboom on 23 April 2004, during a festive celebration of his 65th birthday. The scientific programme of that event is included at the end of this book. Apart from the individual contributions, the book also provides a list of publications by Nooteboom, and a list of his (former and current) PhD candidates.

Nooteboom will remain in office as professor of Phonetics at Utrecht University for several more years, albeit on a part-time basis. This will allow him to have the best of both worlds: to enjoy life with his wife Maaike, and to remain active in our discipline.

Tone and song in Kalam Kohistani (Pakistan) *

Joan L.G. Baart

SIL International

Abstract

Like many other languages in the north-western corner of the South-Asian subcontinent, Kalam Kohistani, spoken in two mountain valleys in northern Pakistan, has contrastive lexical tone. This paper explores how the tonal distinctions of Kalam Kohistani are reconciled with the musical use of pitch in sung recitations of traditional poetry.

1 Introduction

One of the better-kept secrets in linguistics — known to the immediate specialists but hardly at all to the wider linguistics community — is the widespread occurrence of tone (defined here as the use of pitch variations to contrast word meanings) in the languages of the north-western corner of the Indo-Aryan language territory. People are often aware that Punjabi is a tonal language, but usually think of it as an island in a sea of non-tonal languages. The following quote from a recent textbook is rather typical. Speaking about the South-Asian subcontinent the author says, '[E]ven here we find the occasional tonal language, such as Punjabi' (Yip, 2002:171). Stronger yet is Bhatia (1993:xxv), who states that 'Punjabi is the only modern Indo-Aryan language which has developed tonal contrasts.'

In actual fact, Punjabi is not the only tone language in the region, and neither is it an 'occasional' tone language. Rather, it is part of a much larger area, covering north-western India, northern Pakistan, and possibly also bordering regions in Afghanistan, that is remarkably rich in tone languages. These include language varieties such as Hindko and Pahari-Pothwari, which are closely related to Punjabi, and also more distant ones such as many of the so-called 'Dardic' languages of the mountains of northern Pakistan, several languages belonging to the Rajasthani and Western Pahari subgroups of Indo-Aryan, and the non-Indo-Aryan language Burushaski. A survey of tone languages of northern Pakistan can be found in Baart (2003).

Not only is this area rich in number of tone languages, it is also rich in terms of the complexity of tonal phenomena displayed. While many of these languages have not been studied in depth, recent years are witnessing the appearance of a number of important contributions to the study of tone in these languages, notably Schmidt & Kohistani (1998)

* I would like to thank my friend and colleague Muhammad Zaman of Shahoo, Kalam. Neither this nor any of my work on Kalam Kohistani would exist without his help. Lal Badshah, Muhammad Nabi, Maulana Abdul Haq and other poets and musicians of Kalam have made their songs and music available to me, sometimes in written form, sometimes in audio-recorded form, sometimes in both, and sometimes in the shape of a live performance. I benefited from the ideas of several colleagues, including Cal Stevens, Todd and Mary Beth Saurman, and Tom Avery. My research in Pakistan is carried out under the auspices of an agreement of cooperation between SIL and the National Institute of Pakistan Studies (NIPS) in Islamabad. The support over many years of Ghulam Hyder Sindhi, the director of NIPS, has been of tremendous value. Sieb Nooteboom, my Ph.D. supervisor, has had a pervasive influence on my academic work. Even so, not he but I alone am to blame for any shortcomings in this paper.

and Radloff (1999) for Shina, Zoller (forthcoming) for Indus Kohistani, Losey (2002) for Gujari, and my own work (Baart, 1997, 1999, 2004) for Kalam Kohistani.

Many of the languages in this area have rich traditions of storytelling, poetry recitation, music, and song. As the linguistic features of the tone systems of these languages are beginning to be unraveled, curiosity leads us to ask further questions, and one of these concerns the correspondence (or lack thereof) of phonological tone and the melody of song: Is there a systematic relation between the pitches of a song and the phonological tones of the words of the song? To my knowledge, this question has hardly been addressed, if at all, for any of the Indo-Aryan languages. An exception is Vedic chant, to which I will return near the end of this paper.

In the remainder of this contribution I present some results of a study of traditional Kalam Kohistani poetry and its sung recitation. I look at the syllable structure and meter of poetic lines in this particular genre, and at the relationship between linguistic tone and sung pitch. Interestingly, it turns out to be the case that there is no straightforward correlation of tones and sung pitches. This result leads to yet another question, namely, to what extent is it common in the tone languages of the world for tone to be ignored in singing? I am not in a position to give an answer to this question, but will present a few preliminary observations. First, however, I start with a brief introduction to the Kalam Kohistani language.

2 The Kalam Kohistani language

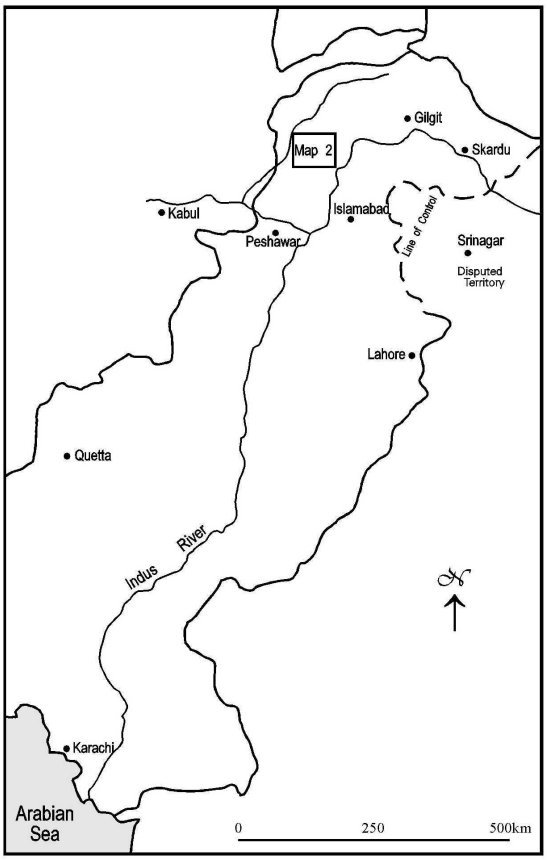
2.1 Language, speakers, and classification

Kalam Kohistani (also called Gawri) is one of about thirty languages that are spoken in the mountain areas of northern Pakistan. *Kohistan* is a Persian word that means ‘land of mountains’ and *Kohistani* can be translated as ‘mountain language’. As a matter of fact, there are several distinct languages in the area that are all popularly called Kohistani. The language under study in this paper is spoken in the upper parts of the valley of the river Swat, in the North-West Frontier Province of Pakistan (see Maps 1 and 2). The name of the principal village of this area is Kalam, and hence the area is known as Kalam Kohistan.

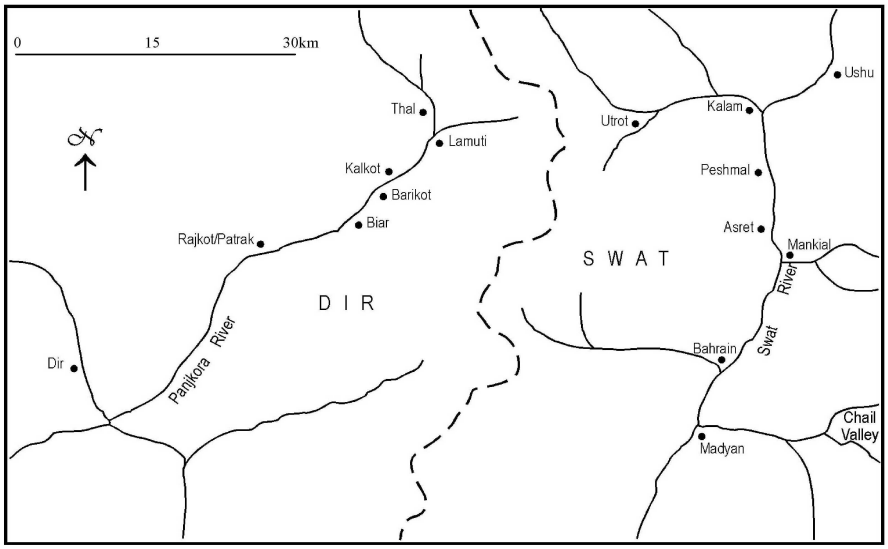
In the older linguistic literature, the language of Kalam Kohistan is referred to as Bashkarik (Morgenstierne, 1940), or as Garwi or Gawri (Grierson, 1919; Barth & Morgenstierne, 1958). These names are hardly, if at all, known to the speakers of the language themselves, who normally just call their language Kohistani. However, very recently a number of intellectuals belonging to a local cultural society have started to call their language Gawri, a name that has old historical roots (see Baart, 1997:4-5).

The same language is also spoken across the mountains to the West of Kalam Kohistan, in the upper reaches of the Panjkora valley of district Upper Dir. When added together, the two Kalam-Kohistani-speaking communities comprise approximately 100,000 people.

According to its genealogical classification (Strand, 1973:302 and 2004), Kalam Kohistani belongs to the Kohistani subgroup of the north-western zone of Indo-Aryan languages, along with several closely related languages in its geographical vicinity: Torwali (in the Swat valley south of Kalam), Indus Kohistani, Bateri, Chilisso, and Gawro (the latter four east of Kalam in Indus Kohistan). Together with a range of other north-western Indo-Aryan mountain languages, these languages are sometimes collectively referred to as ‘Dardic’ languages.



Map 1: Pakistan, showing inset for Map 2



Map 2: Upper Panjkora and Swat valleys

2.2 Phoneme inventory

Vowels

The Kalam Kohistani vowel system consists of six basic vowel qualities to which a distinction of length is applied, see Table 1. In agreement with standard 'Orientalist' practice, long vowels are represented in this paper by writing a macron over the vowel symbol, as in *ā*.

Probably all vowels can have nasalized counterparts. In this paper these are written with a tilde following the vowel symbol, as in *mā~* 'my'.

Table 1. Kalam Kohistani oral vowels.

	front		back	
	short	long	short	long
close	i	ī	u	ū
mid	e	ē	o	ō
open	ä	ǟ	a	ā

Consonants

Table 2 presents the inventory of Kalam Kohistani consonants. Note the voiceless lateral fricative, which is written *ɬ*. It contrasts with the voiced lateral *l* as in *lām* 'village' vs. *ɬām* 'cedar wood'.

Table 2. Kalam Kohistani consonants.

		labial	dental	retroflex	palatal	velar	post-velar
plosive	aspirated	ph	th	ṭh		kh	
	voiceless	p	t	ṭ		k	q
	voiced	b	d	ḍ		g	
affricate	aspirated		tsh	ch	čh		
	voiceless		ts	c	č		
	voiced				ǰ		
fricative	voiceless	f	s	ṣ	š	x	h
	voiced		z			ɣ	
nasal		m	n	ṇ		ŋ	
glide		w			y		
lateral	voiced		l				
	voiceless		ɬ				
flap			r	ṛ			

2.3 Tone

As mentioned above, many of the languages spoken in the north-western corner of the South-Asian subcontinent are tonal. Baart (2003) presents a preliminary overview of tone languages in northern Pakistan. In that paper, these tone languages are grouped into three classes. In Shina-type languages (to which belong Shina itself, Burushaski, Palula, Indus Kohistani, and probably a range of other languages), a contrast between a rising and a falling tone is possible on long vowels (in Indus Kohistani on bimoraic syllables in general), as in the Shina words *kâm* 'relative' vs. *kǎm* 'a vegetable'.

In Punjabi-type tone languages (which include Punjabi itself and also Hindko, Pahari-Pothwari, Gujari, and possibly others) there is a three-member contrast between what is commonly described as a mid or level tone, a high or high-falling tone and a low or low-rising tone. A classical example of the contrast in Punjabi is *kōṛā* 'whip' vs. *kôṛā* 'leprosy patient' vs. *kǒṛā* 'horse'. In Punjabi, there is a clear correlation between the loss of breathy-

voiced consonants and the emergence of tonal contrasts (see Masica, 1991:118-9, and references cited there).

Finally there are a number of languages in the area that have more than three contrastive tones on the surface, and Kalam Kohistani is one of them (along with Torwali and Khalkoti). The five contrastive tones of Kalam Kohistani are illustrated in the following examples. With the examples I give a description of the phonetic pitch that is observed when these words occur in non-final position in the sentence. (In final position, intonation normally adds a low tone, which may then cause a modification of the lexical pitch of the last word.)

(1)	bőr	(high level pitch)	'lion' (singular)
(2)	bô̋r	(high-to-low falling pitch)	'lions' (plural)
(3)	bô̋r̀	(delayed high-to-low falling pitch)	'deaf'
(4)	bőr	(low level pitch)	'Pathan'
(5)	gǒr	(low-to-high rising pitch)	'horse'

There is a distinction between the high-to-low falling pitch and the *delayed* high-to-low falling pitch, in that the delayed falling pattern typically falls, so to speak, from the last syllable of a word onto the first syllable of the next word, while the regular falling pattern is fully executed within one and the same word. The five tones of Kalam Kohistani are 'word melodies': the pitch patterns described above belong to whole words, rather than to single syllables, as in the word *bùbãy* 'apple', where the rise spreads over two syllables. For further information on Kalam Kohistani tone the reader is referred to Baart (1999, 2004).

3 Kalam Kohistani poetry

3.1 Introduction

There are a number of different styles of poetry and song in Kalam Kohistani. In this paper, I focus mainly on what can be called the classical form of Kalam Kohistani poetry, for which the local name is *rō*. This is no doubt also the most popular form of poetry in the language, and audiocassette recordings of the most famous poet-singers of this genre are for sale in the bazaars of the area. The theme of this poetry is usually romantic love, in particular the pain of separation from the beloved, or the sorrow caused by the death of one of the lovers. When reciting, some poets begin every verse with *alā!*, which is the exclamation that is also used in real life for expressing pain or woe.

Kalam Kohistani *rō* is sung to the accompaniment of a local variant of the Chitrali sitar (called *särōd* by the Kalam Kohistanis) and a little drum called *mangey*. Akbar (2000) describes the Chitrali sitar as a long-necked lute made of mulberry-wood. It consists of a guitar-like neck, an oval-shaped, hollow base, six metal playing strings (including two main strings) and five sympathetic strings. The Chitrali sitar is a popular instrument all over the North-West Frontier Province, including not only Chitral but also areas such as Swat, Dir, and the central plains inhabited by the Yousafzai tribe. Most of the sitars in use in Kalam Kohistan are locally made, and may deviate from Akbar's description in several ways. The Kalam Kohistani variant usually has seven playing strings (two of which are used as melody strings) and no sympathetic strings. For the neck of the sitar, Kalam Kohistanis often use the wood of the Himalayan blue pine, while the hollow base may indeed be made of mulberry-wood, or, alternatively, of the wood of one or two other locally available trees, including walnut.

In daily life, the *mangey* is an earthenware water pot. It has a large round belly and a much narrower short cylindrical neck. For use as a musical instrument, the opening of the pot may be covered with a round piece of rubber. Players beat this rubber membrane with one hand

and the round belly of the pot with the other. They may wear a ring on one finger to produce a clicking sound. (Further information on these musical instruments can be found in Akbar 2000:787-9.)

The musicians and the singer take turns performing. The musicians pause when the singer begins to sing the first line of a verse. The sitar plays two or three chords at the end of the first line, and then the singer starts the second line of the verse. During the last few notes of the song, the sitar joins in again, followed by the *mangey*. The two instruments perform together for half a minute or a minute or so, and then the singer begins the next verse.

3.2 Melody and meter

The melody used in the sung recitation of *rō* is traditional and follows a rather rigid pattern, within which only minor variations are possible. The lyrics themselves also conform to rather specific rules of composition. Each individual *rō* consists of only two lines. Together, these two lines express a complete thought, and can stand alone as an independent poem. The lines of a *rō* consist of seventeen syllables each. Main accents are assigned to the fourth, sixth, tenth, twelfth and sixteenth syllables in a line. In sung recitation, these accents are primarily realized by stretching out the length (duration) of the syllables, while downward glissandos may also be executed on these syllables (except for the last main accent, which starts and remains on the bottom of the pitch range). Minor accents are assigned to the other even-numbered syllables in a line.

Syllables that occur in positions of main accent almost always contain a phonologically long vowel (in my corpus of Kalam Kohistani poetry, about 95 percent of the vowels occurring under a main accent are phonologically long). Syllables that occur in positions of minor accent or in unaccented positions usually contain short vowels (only 19 percent of the vowels occurring under a minor accent are phonologically long, and only 16 percent of the vowels occurring in unaccented positions are phonologically long). In all positions, both open and closed syllables are possible. Figure 1 presents an example of a Kalam Kohistani *rō*.

Not indicated in Figure 1 is the fact that the two poetic lines of a *rō* are padded on both sides with some non-changing additional syllables. Lal Badshah, the poet-singer on whose work this paper is mainly based, usually precedes the first line with *alā!* (an exclamation of pain, as explained above). He always starts the second line with the vowel *ē*, and this same sound is also appended at the end of both the first and the second line. *ē* (sometimes *ō)* is also the preferred sound for 'stuffing', that is, for filling holes in a poetic line that has fewer than the required seventeen syllables. An example is seen in the poetic line in (6), glosses and translation in (7), where the tenth syllable is a stuffed *ē*.

(6)

îz-	rā-	ĩl	sāb	kă	tē-	đi	kă-	rá~š	ē	nî-	mă-rù-	ṣă~	khăn	rā-	kă
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

(7) Izraeel master what haste was.making of.Nimarush mountain on!
'With what haste did (the angel) Izraeel strike on the mount of Nimarush!'

The second line in Figure 1 shows an occurrence of a stuffing syllable (the vowel *ō* between syllables 12 and 13) that is less easy to explain, as without it the line already has the required seventeen syllables. My conjecture is that in this case the presence of the stuffing syllable slows down the pace and in this way serves to focus the attention (like a drum roll) on what immediately follows; the following part of the verse indeed contains a surprising turn (it refers to a heap of earth, that is, a grave, and from this we infer the tragic fact that the beloved is dead).

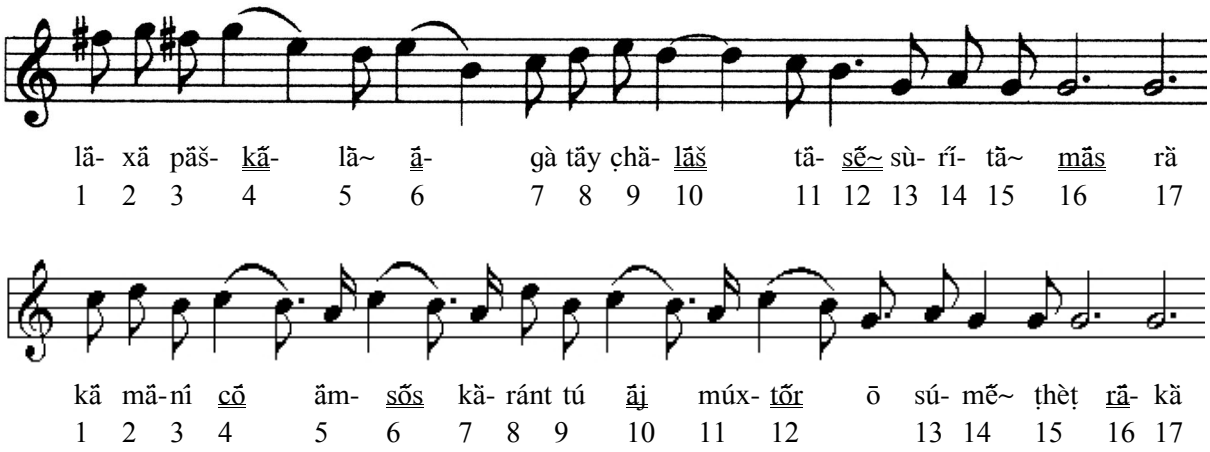


Figure 1. Melody and syllable structure of a Kalam Kohistani *rō* (double underlines indicate the main accented syllables). Glosses and translation: (a) like of.monsoon rain you had.sprayed her of.body flesh on; 'As a monsoon shower you have sprayed (your cruelties) on the flesh of her body'; (b) what having.said having.gone regret you.are.doing, you today offender of.earth mound on. 'Why are you sorry now? You are the guilty one, (sitting in mourning) on a mound of earth.'

3.3 Phonological tone and sung pitch

As Kalam Kohistani is a tone language with lexically contrastive high, low, rising and falling tones (see Section 2.3 above), one might expect a syllable with phonological high tone to be sung to a relatively high pitch, and a syllable with low tone to be sung to a relatively low pitch. However, this expectation is not always borne out. A counter-example can be seen in the first line of Figure 1. The word *ṣhālā̆š* 'sprayed it' (syllables 9 and 10) has a low tone on the first syllable and a high tone on the second syllable (the word as a whole, in other words, has a low-high or rising pattern). In the sung recitation of this verse, however, the pitch of the first syllable is higher than the pitch of the second syllable, thus giving a falling pattern to the word as a whole. Another example of a mismatch of phonological tone and sung pitch is seen in the second line of Figure 2. The word *thō̆s* 'head' (syllable 12) has phonological low tone, yet it is sung to a high pitch (which glides down, but that is normal for all syllables under a main accent, except for the last main accent in a line).

In order to assess the degree to which linguistic tones are ignored in the sung recitation of Kalam Kohistani *rō*, I looked more closely at a sample of fourteen poems, randomly selected from my corpus. Taking a cue from a study of Cantonese song by Wong & Diehl (2002), I compared for each pair of consecutive syllables in the sample the direction of pitch change in the melody of the song with the direction of pitch change that would occur over those same two syllables in normal speech.

The direction of pitch change between two notes in a song is *rising* if the second note is higher than the first note, *falling* if the second note is lower, and *level* if the two notes are the same.[1]

[1] The second one of two tied (i.e. prolonged) notes in the song was disregarded for the purpose of this comparison.

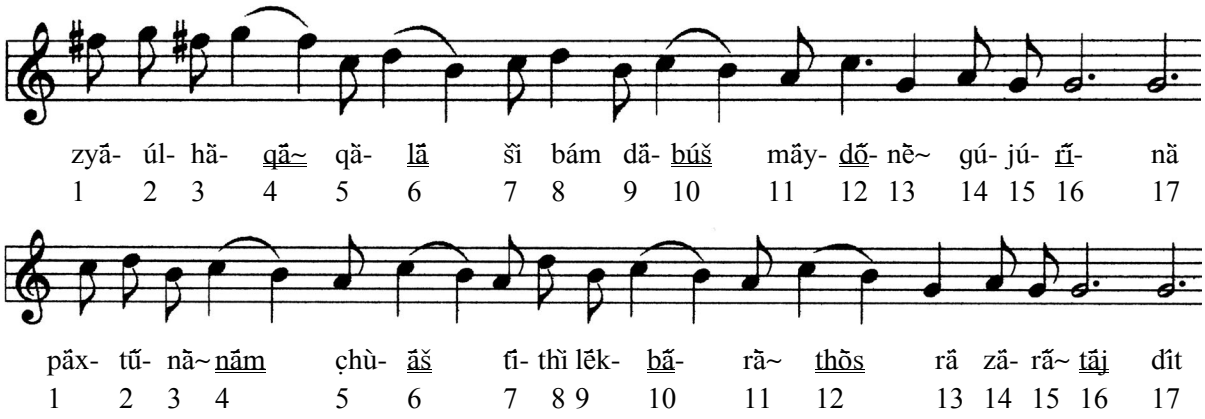


Figure 2. Melody and syllable structure of a Kalam Kohistani *rō*. Glosses and translation: (a) of.Zia.ul.Haq fort in bomb had.buried of.Maidon Gujar.woman; 'A bomb was planted in the fort of Zia-ul-Haq by that Gujar woman of Maidon;' (b) of.Pakhtun name had.betrayed, hence of.youth head on of.gold crown has.given. 'The name 'Pakhtun' had been betrayed and hence a golden crown was placed on the head of a child.'

The linguistic pitch of syllables was classified (on the basis of surface realization) as either low, high, or extra-high. On the basis of this classification, the direction of pitch change (as it would occur in normal speech) for each pair of consecutive syllables was determined, using once more the categories *rising*, *falling*, and *level*. The results of the comparison are shown in Table 3.[2]

Table 3. Comparison of pitch direction over pairs of consecutive syllables in sung recitation (columns) and in normal speech (rows). For each combination, the number of occurrences is listed, with percentage figures in parentheses. The total number of syllable pairs in the sample was 434. The cells on the main diagonal sum to 47.9%.

		Pitch direction in song		
		Rising	Falling	Level
Pitch direction in normal speech	Rising	99 (22.8)	34 (7.8)	25 (5.8)
	Falling	13 (3.0)	89 (20.5)	25 (5.8)
	Level	61 (14.1)	68 (15.7)	20 (4.6)

As is seen in the table, the direction of sung pitch matches the direction of phonological pitch in less than 50 percent of the syllable pairs in the sample. In other words, alignment of linguistic pitch and musical pitch does not seem to be a high-ranked constraint for Kalam Kohistani *rō*. However, when we restrict our examination to rising and falling pitch sequences (the shaded area in Table 3), ignoring level sequences for the time being, there does seem to be a strong tendency to preserve phonological tonal relations in song. A rising sequence of tones corresponds to a non-matching, falling sequence of sung pitches in 34 instances, but to a matching rising sequence of sung pitches in 99 instances (74%). A falling

[2] The initial part of a falling or delayed-falling contour is usually realized as extra-high surface pitch; this is indicated with a double acute accent in Figures 1 and 2. The effect of declination and other perturbations of pitch were disregarded. There were relatively few syllables in the sample bearing a rising or falling contour tone; in such cases, the classification of the syllable was based on the initial pitch of the contour.

sequence of tones corresponds to a non-matching rising sequence of sung pitches in 13 instances, but to a matching falling sequence in 89 instances (87%). In the shaded submatrix in Table 3, then, we observe a much tighter alignment constraint than in the table as a whole.

This observation is to some extent explained by the fact that in the sample, sequences of identical phonological tones are as frequent as rising or falling sequences. However, in the melody used for sung recitation of *rō* (transcribed in Figures 1 and 2), there are no sequences of identical musical notes, except for two or three such sequences at the end of a line. Consequently, it is difficult to accommodate level sequences of tones in the sung melody.

Even if we restrict ourselves to the shaded submatrix, it is still the case that in twenty percent of those instances the direction of pitch change between two consecutive sung syllables does not agree with the direction of change between the phonological tones of those syllables. So, while distinctions between rising and falling tones are preserved in song more often than not, there is a considerable minority of instances where these distinctions are ignored.

4 Tone and song in some other languages

Chao (1956, cited in Wong & Diehl 2002) investigated the relationship between sung pitch and linguistic tone in Chinese songs of various styles. In Chinese 'Singsong' (a style that is intermediate between speaking and singing), each tone is sung with a consistent pitch pattern, making it relatively easy for listeners to identify linguistic tones and, hence, word meanings. On the other hand, in contemporary Mandarin songs, composers mostly ignore linguistic tones in their compositions, according to Chao's findings. Yung (1983, cited in Wong & Diehl 2002) looked at Cantonese opera and found a relatively consistent relation between melody and tone, comparable to Chao's finding for Singsongs. Wong & Diehl (2002) analyzed four contemporary Cantonese songs. They looked at direction of pitch change over pairs of consecutive syllables and found an overall correspondence of 92% between musical and tonal sequences (definitely much higher than my result of 48% for Kalam Kohistani *rō* in Table 3 above).

Saurman (1999) analyzed a number of Thai songs of different styles, looking at direction of pitch change over consecutive syllables. The degree of correspondence between tones and sung pitches in classical and traditional songs was around 90 percent. For contemporary popular songs (that borrow elements of western music) the number of matching correspondences was between 60% and 70%. In a western hymn, translated into Thai, tones and sung pitches matched in only 42% of the cases. For the Thai national anthem, the degree of matching she found was 32%.

Saurman suggests that the degree to which tones and sung pitches match with one another in Thai is related to the degree to which the melody of a song is 'imposed' on the lyrics. A melody is imposed when it is a pre-existing traditional melody to which new lyrics have been set (as is the case with the Thai national anthem), or when the melody, or certain elements of it, have been borrowed from outside.

Howard (2000:240-241) reports on the relation of tones and sung pitches in Vedic recitation. In traditional terminology, syllables in Vedic texts can bear a raised accent (*udātta*), an unraised accent (*anudātta*) that normally precedes the *udātta*, a sounded (transitional) accent (*svarita*) that marks a shift down from a raised accent, or no accent at all. One might expect the raised accent to be sung to a high pitch, and conversely, the unraised accent to be sung to a non-high pitch, but Howard cites research that shows that this is not consistently the case and that the accents are in fact not distinguished by pitch.

In March 2004, I posted a question on a linguistics discussion list about tone languages and song. I obtained the following information about what happens to the tones of various languages when people sing:

Piraha (Brazil): The non-normal channels of speech (hum speech, yell speech, musical speech, whistle speech) all crucially preserve the tones (D.L. Everett).

Cheyenne (USA): In general, Cheyennes align the notes of music so that phonemic high tone has a higher pitch in music (W. Leman).

Mamainde (Brazil): Tonal distinctions are partially preserved in only some types of songs (D. Eberhard).

Kabiye (Togo): In church music at least, the musical melody does not always align with spoken tonal patterns (D. Roberts).

Nawuri (Ghana): It is easy to find mismatches between phonological tones and sung pitches (R. Casali).

Chumburung (Ghana): The tones of the language are ignored in song (K. Snider).

Akoose (Cameroon): There seems to be no correlation between the lexical and grammatical tones of the language and the melody of songs (R. Hedinger).

Mazatec (Mexico): From an inspection of four popular songs, it appears that the tones of the spoken text do not coincide with the melody of the sung text (S. Marlett).

Isthmus Zapotec (Mexico): Tonal behavior of speech does not correspond to the musical pattern of songs (S. Marlett).

5 Summary and conclusion

In view of its rich tonal structure, one would expect tonal distinctions in Kalam Kohistani to be preserved in song. However, my study so far of *rō*, the most popular style of song and poetry in Kalam Kohistani, does not show a simple, straightforward relation between phonological tones and sung pitches. Even though a rising sequence of tones more often than not corresponds to an ascending sequence of musical notes in a song, and a falling sequence of tones more often than not corresponds to a descending sequence of musical notes, yet the number of instances where tones and sung pitches do not match is considerable.

A preliminary survey of song in some other tone languages of the world shows that there is a wide range of variation with respect to the degree to which phonological tonal distinctions are preserved in songs. Also, within a language there may be a wide range of variation between different styles of song. In styles where a musical melody is 'imposed' on the lyrics (fixed, traditional melodies, or melodies that have been borrowed from outside), there is a greater likelihood that tonal distinctions are ignored in song as compared to styles where melodies are newly composed with the lyrics.

The sung recitation of Kalam Kohistani *rō* makes use of a standard, traditional melody, which is imposed on the text of the songs. My finding, then, that tones and sung pitches in this genre often do not match, fits within the more general scheme of things. There are other styles of song in Kalam Kohistani, including lullabies, children's songs and work songs. The tone-song relationship in those other genres is an obvious area for further research. Other factors, too, may determine to what extent tone is preserved in song, one of them being the functional load of tone in a language. The subject of functional load also leads to questions about the intelligibility of songs. For instance, if tone is often ignored in song, how do listeners still derive the meaning of the words of a song? Such questions are worthy of further study. As far as South Asia is concerned, we are not even beginning to scratch the surface.

References

Akbar, Mohammad (2000). North West Frontier Province. In Alison Arnold (Ed.), *South Asia: the Indian subcontinent* (pp. 785-791). New York: Garland Publishing. (The Garland Encyclopedia of World Music, Vol. 5).

Baart, Joan L. G. (1997). *The Sounds and Tones of Kalam Kohistani: With wordlist and texts*. Islamabad: National Institute of Pakistan Studies and Summer Institute of Linguistics.

Baart, Joan L. G. (1999). Tone rules in Kalam Kohistani (Garwi, Bashkarik). *Bulletin of the School of Oriental and African Studies*, 62/1, 87-104.

Baart, Joan L. G. (2003). Tonal features in languages of northern Pakistan. In Joan L.G. Baart & Ghulam Hyder Sindhi (Eds.), *Pakistani languages and society: problems and prospects* (pp. 132-144). Islamabad: National Institute of Pakistan Studies and Summer Institute of Linguistics. Available: http://www.geocities.com/kcs_kalam/tone.html

Baart, Joan L. G. (2004). Contrastive tone in Kalam Kohistani. *Linguistic Discovery*, 2/2, 1-20. Available: http://linguistic-discovery.dartmouth.edu/WebObjects/Linguistics

Barth, Fredrik, & Morgenstierne, Georg (1958). Vocabularies and specimens of some southeast Dardic dialects. *Norks Tidsskrift for Sprogvidenskap*, 18, 118-136. Oslo.

Bhatia, Tej K. (2002). *Punjabi: A cognitive descriptive grammar*. New York: Routledge.

Chao, R. C. (1956). Tone, intonations, singsong, chanting, recitative, tonal composition, and atonal composition in Chinese. In M. Halle, H. G. Lunt, H. McLean, & C. H. Van Schooneveld (Eds.), *For Roman Jakobson: essays on the occasion of his sixtieth birthday, 11 October 1956*. The Hague: Mouton.

Grierson, George A. (1919). *Linguistic Survey of India, Vol. 8/2*. Calcutta.

Howard, Wayne (2000). Vedic chant. In Alison Arnold (Ed.), *South Asia: the Indian subcontinent* (pp. 238-245). New York: Garland Publishing. (The Garland Encyclopedia of World Music, Vol. 5)

Losey, Wayne E. (2002). *Writing Gojri: Linguistic and sociolinguistic constraints on a standardized orthography for the Gujars of South Asia*. MA thesis, University of North Dakota. Available: http://www.und.nodak.edu/dept/linguistics/theses/2002Losey.htm

Masica, Colin P. (1991). *The Indo-Aryan Languages*. Cambridge: Cambridge University Press.

Radloff, Carla F. (1999). *Aspects of the Sound System of Gilgiti Shina*. Islamabad: National Institute of Pakistan Studies and Summer Institute of Linguistics.

Saurman, Mary E. (1999). The agreement of Thai speech tones and melodic pitches. *Notes on Anthropology,* 3/3, 15-24. Dallas, TX: Summer Institute of Linguistics.

Schmidt, Ruth Laila, & Kohistani, Razwal (1998). Paalus /kostyo~/ Shina revisited. *Acta Orientalia*, 59, 106-149.

Strand, Richard F. (1973). Notes on the Nuristani and Dardic languages. *Journal of the American Oriental Society*, 93/3, 297-305.

Strand, Richard F. (2004). *Indo-Aryan-Speaking Peoples of the Hindukush Region.* Available: http://users.sedona.net/~strand/IndoAryan/IndoAryas.html

Wong, Patrick C. M., & Diehl, Randy L. (2002). How can the lyrics of a song in a tone language be understood? *Psychology of Music*, 30, 202-209.

Yip, Moira. (2002). *Tone*. Cambridge: Cambridge University Press.

Yung, B. (1983). Creative process in Cantonese opera I: the role of linguistic tones. *Ethnomusicology*, 27, 29-47.

Zoller, Claus Peter (forthcoming). *A Grammar of Indus Kohistani, Vol. 1: Dictionary*. Berlin: De Gruyter.

Qualities of a voice emeritus

Gerrit Bloothooft and Peter Pabon

Utrecht University

Abstract

The effects of vocal ageing are investigated in a professional mezzo-soprano singer, for which the phonetogram, and 45 vowels, each sung at fundamental frequencies of 220, 392, and 659 Hz, were recorded at the age of 52 and 74 years. The comparison demonstrates a serious loss in the vocal range, dynamics and control: (1) a loss of half an octave in the highest fundamental frequency range, (2) a loss of 6 dB at the highest vocal intensities, (3) less accuracy in targeting of F_0, (4) no significant change in average vibrato frequency, but (5) much more instability in vibrato frequency and less vibrato modulation depth. The analysis of vocal vibrato is realized with a new method that allows computation of instantaneous vibrato frequency and extent (modulation depth).

1 Introduction

When a person comes of age, the vocal apparatus will not escape from physiological changes such as dehydration of tissues, ossification of cartilages, changes in muscular structure, reduction of the number of nerve fibers and reduction of the speed of action potentials, among others. A major acoustic effect of this ageing process is the gradual increase of the pitch of the speaking voice for males from about 120 Hz at the age of sixty, to 155 Hz (vowel /a/) or even 187 Hz (vowel /i/) for elderly over ninety years of age. Not so much change was found for females (Decoster, 1998), implying that male and female speaking voices become to some extent more similar at older age. In addition, Decoster reported an increasing acoustic instability, exemplified by an increase of the variability of the speaking fundamental frequency (F_0), an increase of jitter and shimmer (random fluctuations in frequency and amplitude of F_0), and a decrease of the harmonics-to-noise ratio of the spectrum. Spectral (envelope) changes over age seem to be less pronounced.

The lowest and highest possible fundamental frequency a voice can produce critically depends on the state of the vocal apparatus, specifically on subglottal pressure and muscular tension. As a consequence, it may be expected that the vocal tonal range is sensitive to ageing. And indeed, a decrease of the total fundamental frequency range sets in at the age of sixty, for both males and females, and reduces from about 24 semitones (two octaves) to 18 semitones at the age of ninety (Böhme & Hecker, 1970).

Whereas relatively little research has been devoted to the effects of ageing on the normal human voice, even less attention has been given to the special case of the professional singing voice. Singers optimally train their voice during their career, but of course at some moment also for them age will take its toll. For singers, the ageing process may be even much more critical than for non-singers, since they usually explore their voice to its dynamic and tonal extremes. Also voice control, the ability to follow precisely the musical score, is of eminent importance to a singer, and its precision may reduce with age. These capabilities are less

important in speech, where less vocal control or reduced pitch range will longer go unnoticed, but for singers the slightest loss may foreshadow the end of a career.

In this study we explore the effects of age on the acoustic properties of a professional mezzo-soprano singer's voice. In 1981, recordings were made in the framework of a doctoral thesis on the spectrum and timbre of sung vowels (Bloothooft, 1985). At that time the singer was 52 years of age. We were in the circumstances that the recordings could be repeated 22 years later, in 2003, when she was 74 years old. This opened the possibility of a longitudinal study of acoustic changes in the same voice. We concentrated on influences on vocal dynamics, as measured in the phonetogram, and on vocal precision and control, by means of an analysis of vibrato in sung vowels. For the latter study a new method was developed for the decomposition of the pitch trace into transient effects, drift, vibrato, and jitter.

2 Recordings

In 1981, recordings were made of professional singers, among which the mezzo soprano who is central in this study. In an anechoic room at the Free University of Amsterdam, the nine vowels /a, ɑ, i, u, ɔ, œ, y, ɛ, e/ were sung in /h/-vowel-/t/ context with a duration of one to two seconds each, at fundamental frequencies (F_0) of 220 Hz (A3 or a), 392 Hz (G4 or g') and 659 Hz (E5 or e"). The singers were asked to sing these vowels in several modes, out of which the following five were used in the present study: neutral, light, dark, soft and loud. Recordings were made with a microphone positioned at 0.3 m from the singer. The recordings in 2003 were made in the personal studio of the singer. A microphone pair, one at a distance of 0.3 m, the other close to the mouth, was fixed on a headset. The microphone close to the mouth was used for the softer phonations (but calibrated for the sound level at 0.3 m), while the other microphone was used for the loud phonations. In this way an optimal signal-to-noise ratio could be achieved.

The phonetogram recordings from 1981 were the first using a computerized measurement technique (Bloothooft, 1981). Under the same conditions as described above, the singer stood facing a monitor on which F_0 was presented on the x-axis and vocal intensity (I) on the y-axis. Real-time measurement of F_0 and I was translated to a position of a marker on the monitor. With the help of this visual feed-back the singer was asked to complete the phonetogram, i.e. to sing all possible combinations of F_0 and I. In subsequent separate sessions, the singer was asked to make a registration of a phonetogram while singing in a specific register only. Usually the terms chest register, middle register, head register, and falsetto register (or voice) were used. The singer was entirely free to use his/her own interpretation of these terms, but should strictly keep the intended register with no mixing of other registers. The mezzo-soprano singer made a distinction between chest voice, mid voice, and head voice (or falsetto). From the 1981 phonetogram recordings only graphical representations remained. In 2003, the phonetogram registration was repeated with the voice-profiler system by Pabon (www.voiceprofiler.com). In addition to fundamental frequency and vocal intensity, acoustic voice parameters were simultaneously recorded (Bloothooft & Pabon, 1999), out of which the crest factor has been used in this study. The crest factor describes the ratio of the maximum amplitude and the RMS value of a signal. Its value is maximal for a peaked signal and minimal for a sine wave. In this respect the crest factor relates to spectral properties and tends to indicate a flat spectrum for high values and a falling spectrum for lower values. Because of fatigue, the singer this time experienced difficulties to complete all the measurements for the full and separate register phonetograms. We will only present the result of all combined measurements.

3 Singer

The singer is a renowned Dutch mezzo-soprano, born in 1929. She had a long international career both in opera and in Lied singing. In 1981, at the age of 52, she was in her last years as a regular artistic performer. In 2003 she still worked as a vocal pedagogue[1].

4 F_0 decomposition

In general, the trace of the fundamental frequency of song can be thought to consist of three different components: (1) a base-line that follows the musical score, including effects of F_0 onset, F_0 transition between two tones, and over- or undershoot during this process, (2) a quasi-stationary, sinusoidal-shaped vibrato (modulation of F_0) with typical modulation frequencies between 4.5 and 9 Hz (Prame, 1995), and (3) fast irregular fluctuations, called jitter. Although each of these three components has its own specific frequency range, there can be some overlap which complicates the decomposition of the F_0 trace. Here we describe our method to accomplish the F_0 decomposition in the special case of vowels, individually sung at a prescribed F_0 frequency. Figure 1 shows a typical example of an original F_0 trace of a sung vowel, derived with Praat software (Boersma & Weenink, 1996), yielding fixed time samples, approximately corresponding to a waveform period[2].

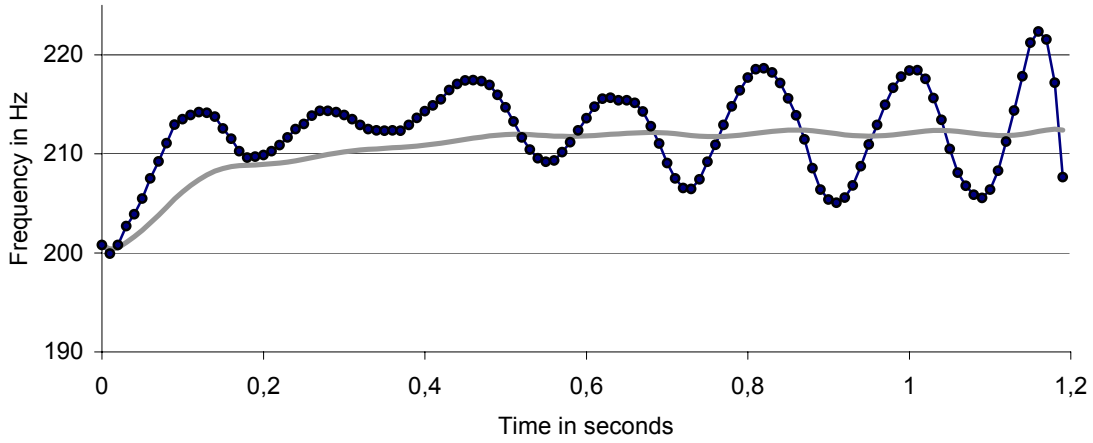


Figure 1. F_0 trace of the vowel /a/, sung by a mezzo-soprano in /hat/ context with an intended fundamental frequency of 220 Hz (dotted line). The grey line is the estimated transient component, produced by an adaptive low-pass filtering technique in which lowest cutoff frequency of 0.5 Hz is reached at 0.64 seconds.

The vowel onset often consists of a short time interval, between 100 and 200 ms long, in which F_0 starts at a low initial frequency (usually sung with low intensity) while increasing quickly towards the target value. During this onset the transient component 1 dominates. Usually within half a second from voice onset, vibrato (component 2) sets in. Although the singer will intend to keep the same average pitch, slow drift in F_0 may occur which should be considered as part of component 1. Finally, over the whole length of the F_0 trace small fluctuations may be observed (jitter, component 3, not so much present in professional singing). Computationally, we face a problem in the separation of the three F_0 components. If

[1] The singer suffered a stroke in 2000, but fortunately recovered well and retained full linguistic and vocal competencies. She could take up her pedagogical practice again.

[2] The Praat procedure includes an optimization procedure across a four-sample window as a result of which the F_0 values are less suited for the analysis of jitter, which was, however, no part of this study.

we separate vibrato by a fixed band-pass filter with cutoff frequencies of 2 Hz and 10 Hz, we will also include part of the F_0 transition during vowel onset. Moreover, vibrato usually is a quasi-sinusoidal modulation and includes higher frequency components, which will be lost in band-pass filtering. We therefore prefer to separate vibrato by removing the F_0 onset transition and subsequent possible drift by a dynamic low-pass filter with variable cutoff frequency. The latter is high during vowel onset — to capture the relatively rapid F_0 transient —but much lower soon thereafter — to capture average F_0 and slow F_0 drift. A running average procedure on the F_0 samples approximates this behavior. The grey line in Figure 1 shows the estimated transient and drift of F_0. The difference between both traces is an estimate of the true vibrato component, and is shown in Figure 2.

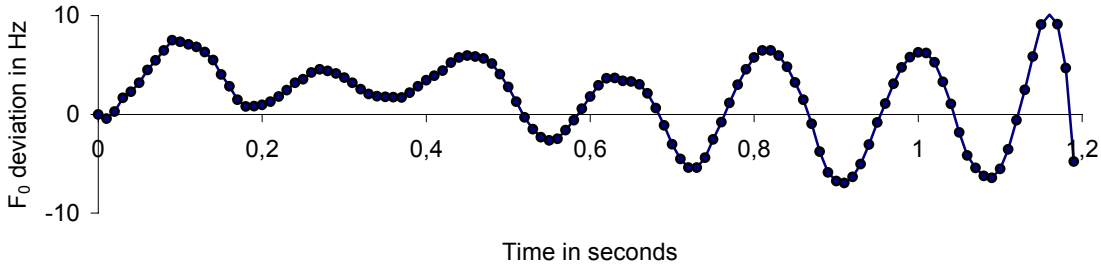


Figure 2. Remaining vibrato trace (F_0 deviation), after subtraction of the estimated F_0 transient, average, and drift from the original trace.

We would like to know not only the average vibrato frequency and vibrato extent (or modulation depth), but also the variation in both measures. For that, we need to compute an instantaneous estimate of the vibrato frequency and vibrato extent at all time samples. The method for this (an instantaneous frequency model) is described in the appendix. For the vibrato trace of figure 2, the resulting instantaneous vibrato frequency is shown in figure 3 and the instantaneous vibrato extent in Figure 4.

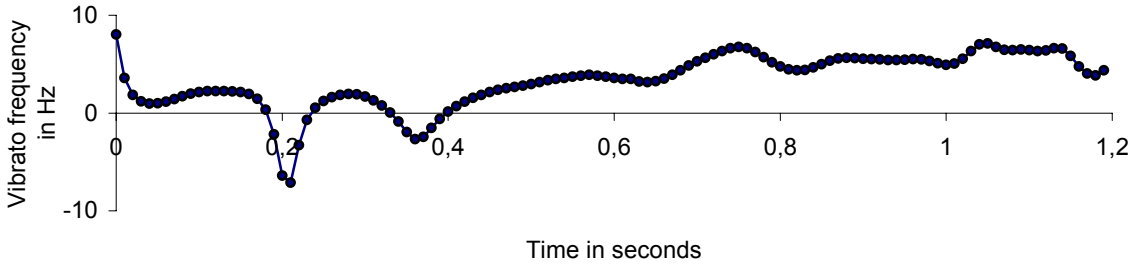


Figure 3. Instantaneous vibrato frequency of the trace of Figure 2.

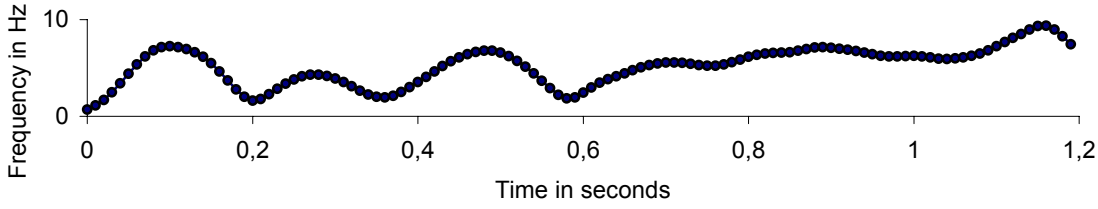


Figure 4. Instantaneous vibrato extent of the trace of Figure 2.

During the first 0.6 second of phonation, the instantaneous vibrato measures are quite unstable, and even have negative values. More study is needed to understand whether this is an intrinsic property of the voice or a computational artifact. Therefore, only data from two subsequent vibrato periods, selected from the second half of the vowels were used. For this interval we computed the average F_0 and the average and standard deviation of the instantaneous vibrato frequency. Instead of instantaneous vibrato extent, which is not further discussed in this paper, we used the difference in minimum and maximum F_0 during the chosen vibrato periods as a direct measure of vibrato modulation depth.

5 Results: phonetograms

In Figure 4 the phonetogram registration of 1981, at age 52 is shown. In the phonetogram the separate registrations of chest voice, mid voice and head voice are presented by vertical and horizontal hatching. The phonetogram at age 74 is presented in Figure 5 as an overlay of the one at age 52. Here, the grey-scale indicates the crest factor value: the darker, the higher the crest factor, which implies a more flat spectrum. When comparing both phonetograms, it is remarkable that the shape of the upper contour at high vocal intensity is quite similar (disregard the octave error at F_0 = 400 Hz in 1981), although there is a uniform loss in vocal intensity of about 6 dB at older age. At the high end of the frequency range, the singer has lost more than half an octave, while no losses are observed at low F_0. The differences at low vocal intensity may be due to the fact that the singer in 1981 probably sang at lowest acceptable singing level rather than lowest phonation level, which more likely is found between 55 and 60 dB SPL. In addition, measurement sensitivity at low vocal intensity was poorer in 1981.

The phonetogram from 1981 demonstrates that the chest voice ranged up to 300 Hz. This corresponds rather well to the highest crest factor levels seen in 2003, which up to 300 Hz mainly originate from the separate phonetogram recording in chest voice (not shown), and for higher F_0 values to a mixed register voice type.

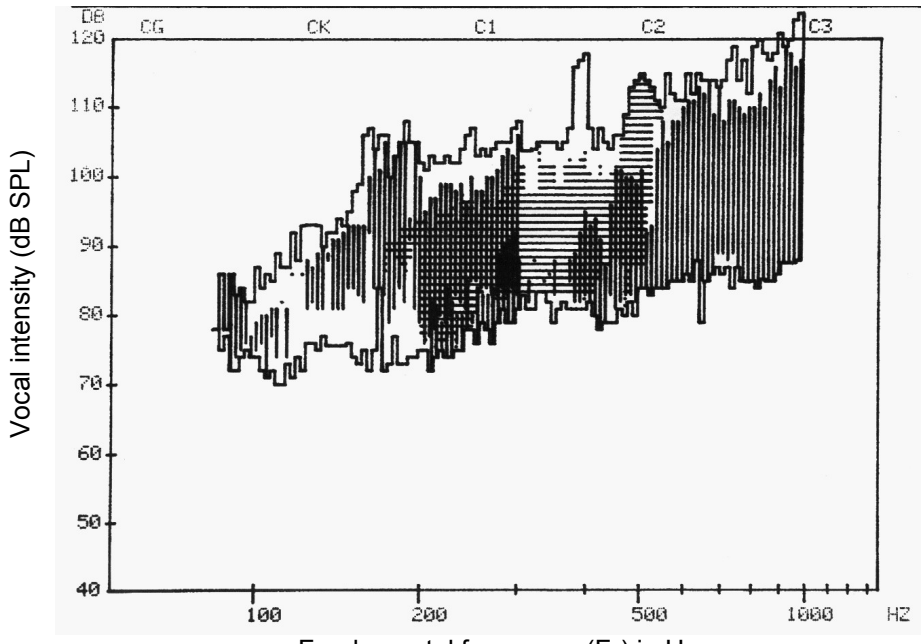


Figure 4. Phonetogram of the mezzo-soprano singer at age 52 (1981). Vertical and horizontal hatching indicates from left to right chest voice, mid voice and head voice.

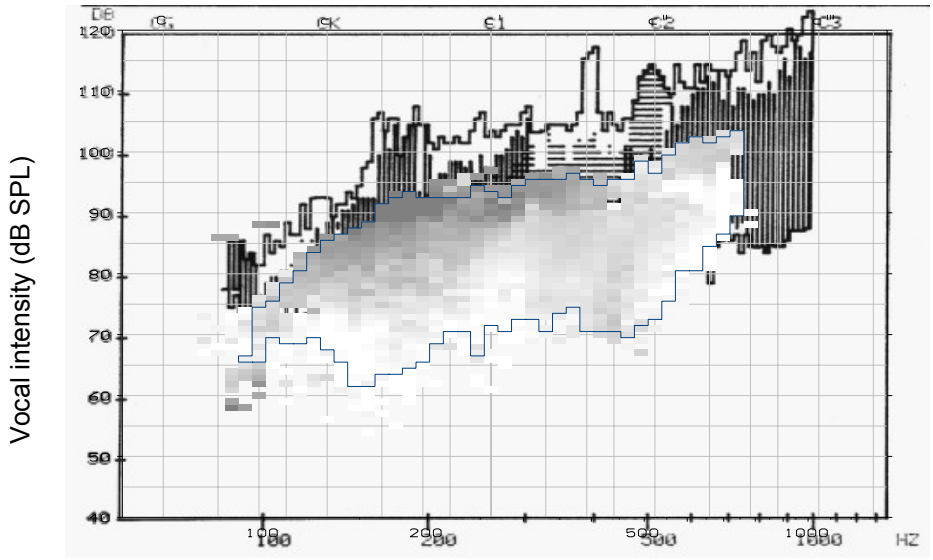


Figure 5. Phonetogram of the mezzo-soprano singer at age 74, as overlay of the phonetogram from 1981. The contour limits the area with more frequent phonations. The grey-scale indicates the level of the crest factor, which gradually changes from 3 dB (white) for a sinusoidal waveform, to 9 dB (dark) for a more peaked waveform.

The mid voice, as partly covering the higher chest voice and the lower head voice areas, roughly ranged from 200 to 600 Hz, which in 2003 still is a range with relatively high crest factors. The higher part of what the singer called head voice (falsetto) seems lost at later age.

6 Results: fundamental frequency and vibrato

At the fundamental frequencies of 220, 392 and 659 Hz, nine different vowels were each sung in five different modes, yielding 45 vowels at each pitch. Table 1 shows the target frequencies realized. The accuracy in reaching the target frequency was much better at age 52, although not really perfect. At later age, there was a serious undershoot, especially at higher fundamental frequencies. The singer noticed this herself during the recording of a vowel series. She also experienced some difficulties in register control during vowel onset at F_0 = 659 Hz. All this may relate to the fact that this frequency (E5 or e") was at age 74 at the top of the vocal range, while at age 52 it was in the centre of the head voice.

Table 1. Targeted and realized F_0 values in Hz, averaged over 45 vowels.

Target F_0	Realized F_0	
	age 52	age 74
220	211.5 ± 2.7	209.0 ± 4.5
392	385.0 ± 5.8	357.7 ± 9.1
659	632.2 ± 9.0	582.1 ± 26.7

In figure 6 we present a typical example of the distribution of the instantaneous vibrato frequency for a vowel /a/ sung at 392 Hz. The major difference between the distributions is exemplified by their standard deviation, which is 0.88 Hz at age 52 and 2.14 Hz at age 74, implying a more instable vibrato at later age. On average, across all vowels and F_0 values, the

standard deviation of the instantaneous vibrato frequency was 1.19 Hz at age 52 and 3.73 Hz at age 74.

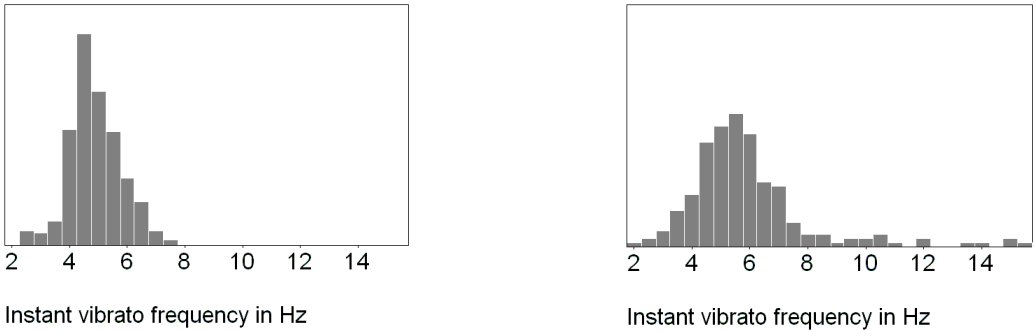


Figure 6. Normalized distributions of the instantaneous vibrato frequency for a typical vowel /a/, sung at 392 Hz in neutral mode. Left-hand panel at age 52 (n = 322), right-hand panel at age 74 (n = 205).

Not only is the stability of vibrato itself poorer at older age, also the variability in *average* instantaneous vibrato frequency per vowel is much larger. This is shown in the distributions presented in Figure 7. At age 52, the average vibrato frequency per vowel is well within a restricted range between 5 and 6.5 Hz (grand average 5.7 Hz), irrespective of the target F_0. At age 74, the distribution is much wider and even almost homogeneous at F_0 = 220 Hz, implying little control of vibrato, although the grand average (5.5 Hz) is not significantly different. Neither were any significant effects found of the five modes of singing.

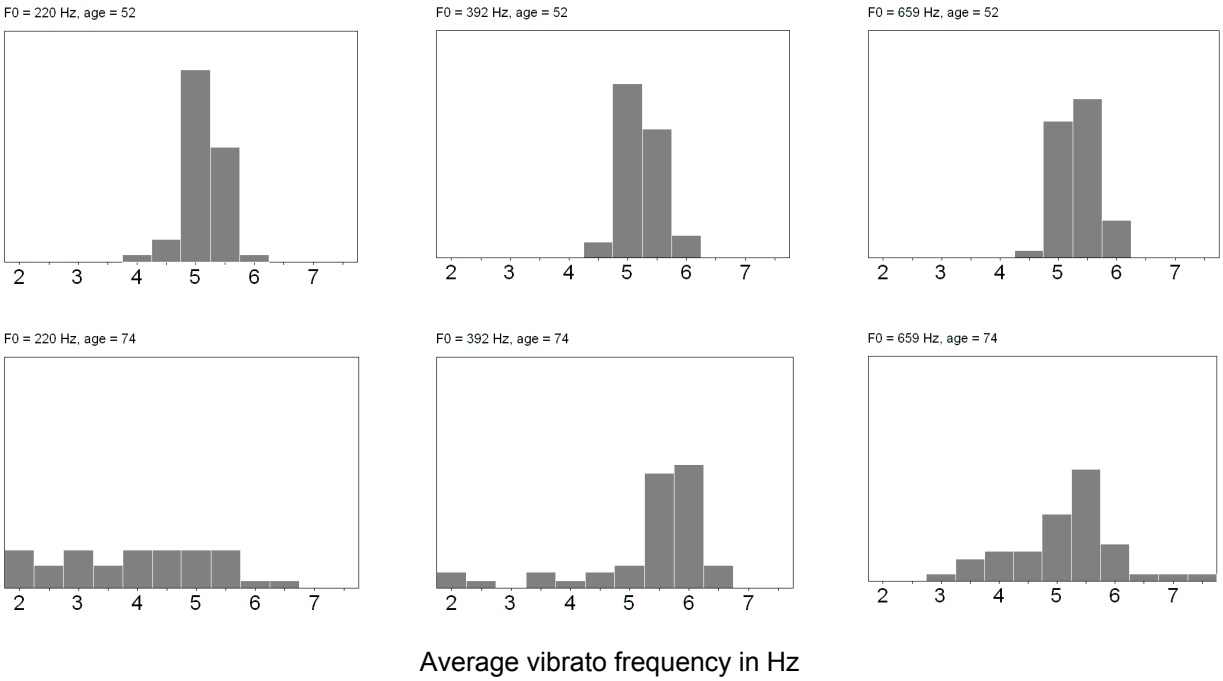


Figure 7. Distribution of the average vibrato frequency over all 45 vowels per age and per fundamental frequency.

Finally, we studied the modulation depth of the vibrato by taking the top-to-top frequency difference in F_0 during two vibrato periods. The result at F_0 = 220 Hz is presented in figure 8.

The mean at age 52 was 17.4 Hz; at age 72 this was reduced to 10.2 Hz. The same relative reduction was found for the other values of F_0. At age 52, these vibrato extent values correspond to those found by Seidner et al. (1995) for singers between 33 and 55 years of age, while those for our mezzo soprano at age 72 correspond in Seidner to values only found in (soft) singing with less expressive voice.

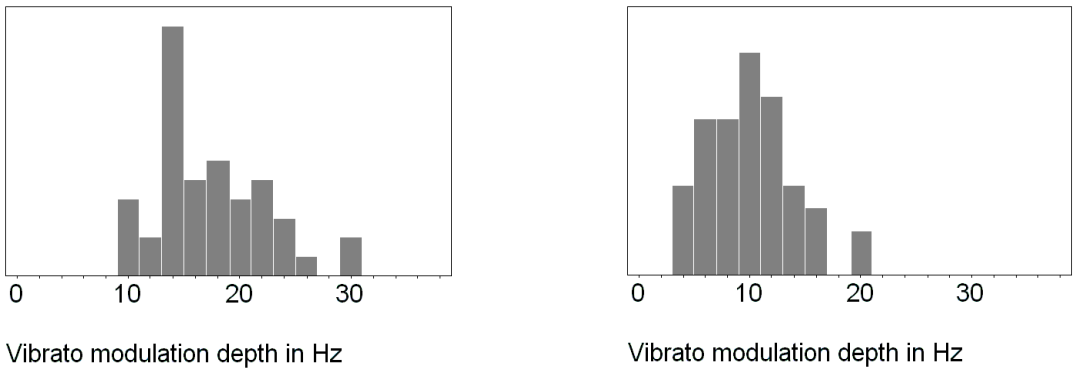


Figure 8. Distribution of the vibrato modulation depth (top-top) at F_0 = 220 Hz. Left-hand panel at age 52, right-hand panel at age 74, n = 45.

7 Discussion

A longitudinal analysis of singing is rare. This pilot study explored some possibilities for acoustic research on a mezzo-soprano singer for which comparable recordings were available at the age of 52 and 74. The literature shows that at least for the normal speaking voice noticeable acoustic changes in the voice set in at these decades. We found: (1) a considerable loss in the highest fundamental frequency range, (2) a loss of 6 dB at the highest vocal intensities, (3) less accuracy in targeting of F_0, (4) no significant change in average vibrato frequency, but (5) much more instability in vibrato frequency and less vibrato modulation depth. This implies a serious loss of the voice range and of vocal control. It should be realized that the effects of ageing can be manifold and that intersubject variability should be expected to be much higher than in young healthy voices, while fewer professional singers will be available for further study than non-singing subjects. Maybe the study of the ageing singing voice is deemed to be exploratory and difficult to generalize. A future attempt to record most of the 14 singers studied in 1981, will help us to come to grips with general and individual factors in the ageing professional singing voice.

Appendix: Derivation of the instantaneous vibrato frequency and amplitude

A special method has been developed that allows evaluation of the sinusoidal smoothness of vibrato on even fragments of a vibrato period. For this, we return to the circular base of sinusoidal motion, in which a time step corresponds to an angle increment. Change in the instantaneous sine frequency then corresponds to a change in the angle increment step. This is known as an instantaneous frequency model, in which the distance to the origin corresponds to the instantaneous amplitude value for every time sample. We will demonstrate the application to vibrato modeling using a synthetic F_0 trace with an average of 250 Hz, vibrato frequency increasing from 1 to 10 Hz, and vibrato modulation depth increasing from zero to ±1 Hz (Fig. A1).

A two-dimensional sinusoidal representation of this trace is realized by first computing a 90° phase-shifted version of the trace (grey line in figure A.1). Subsequently, the original F_0

values are presented as X co-ordinates and the values of the 90° phase-shifted version as co-ordinates on the (imaginary) Y-axis (see figure A.2). In the resulting graph, the time axis is not an independent variable, but it curls through this plane according to the X/Y pairs, starting and ending in the origin. The spacing between time points is small in the beginning, representing a low instantaneous vibrato frequency, while spacing (or vibrato frequency) increases over time. We now smooth the trajectory and compute the instantaneous vibrato frequency from the rate of change in spacing. The result is shown in figure A.3, in which the exponential increase in instantaneous vibrato frequency is clearly visible. The rather strong ripples on the curve indicate the limited precision in the discrete approximation of the various computational steps.

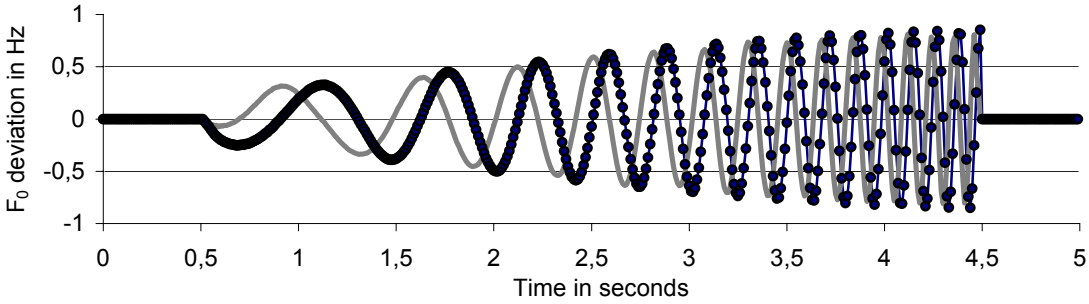


Figure A.1. Synthetic F_0 trace (fluctuations around an average of 250 Hz), with increasing modulation frequency and amplitude. The grey trace is the same signal but with a 90^o phase shift.

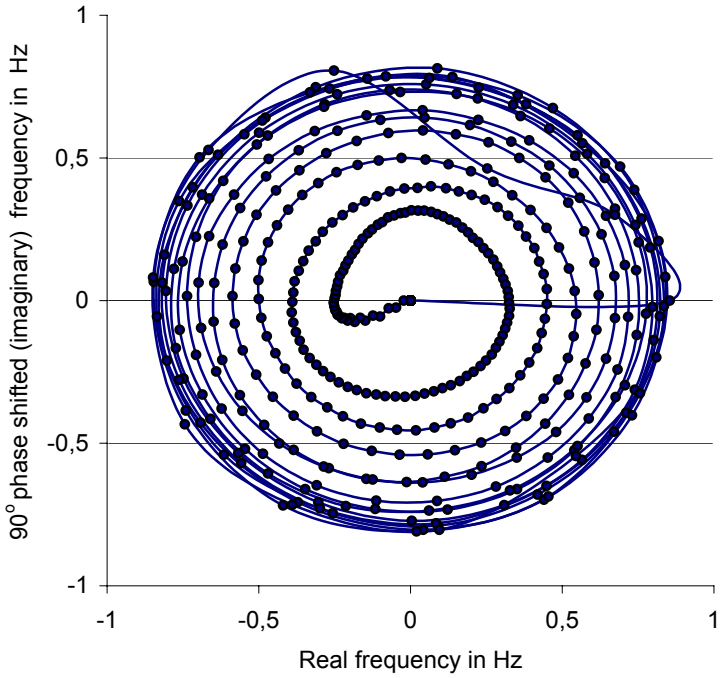


Figure A.2. Analytic signal plane with the values of the original F_0 trace along the X-axis and their 90° phase-shifted versions along the Y-axis, resulting in a circular trajectory.

The distance of the trajectory to the origin in Figure A.2 represents the instantaneous vibrato extent, which also increases gradually. The result is shown in Figure A.4. Notice that the given amplitude only has positive values and compares to half the top-to-top range of the vibrato modulations in the original F_0 trace. From the instantaneous vibrato and extent, a smoothed F_0 trace can be derived that approximates the original signal. Subtraction of the smoothed version from the original signal will reveal the short term differences, or jitter. In this example, this is not demonstrated.

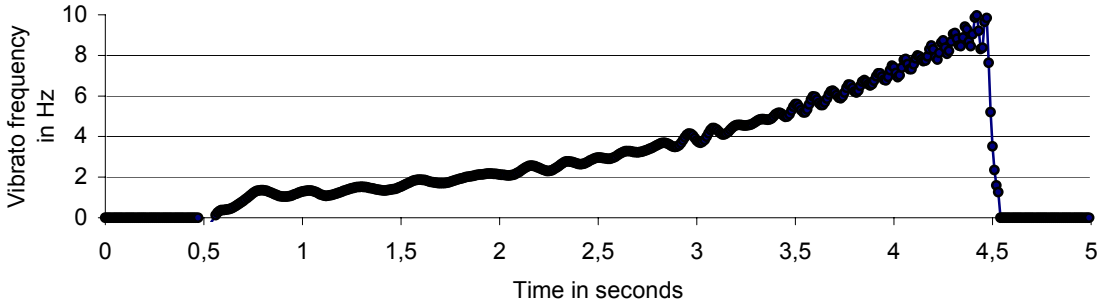


Figure A.3. Instantaneous vibrato frequency of the synthetic F_0 trace, derived from the smoothed instantaneous phase from the trajectory in Fig. A.2.

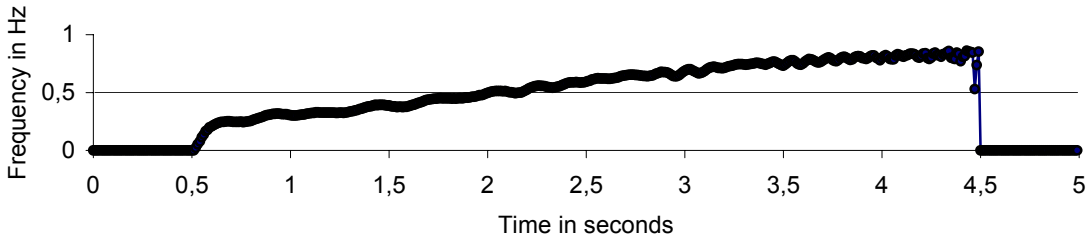


Figure A.4. Instantaneous vibrato extent of the synthetic F_0 trace.

References

Bloothooft, G. (1981). A computer-controlled device for voice-profile registration. In *Proceedings IX Conference of the Union of European* Phoniatricians, Amsterdam (pp. 83-85).

Bloothooft, G. (1985). *Spectrum and timbre of sung vowels*. Doctoral issertation, Free University, Amsterdam.

Bloothooft, G., & Pabon, P. (1999). Vocal registers revisited. In *Proceedings Eurospeech '*99, Budapest (pp. 423-426).

Boersma, P., & Weenink. D. (1996). *Praat: Doing phonetics by computer*. Report nr. 132. Amsterdam: Institute of Phonetic Sciences, University of Amsterdam.

Böhme, G., & Hecker, G. (1970). Gerontologischen Untersuchungen über Stimmumfang und Sprechstimmlage. *Folia Phoniatrica*, *22*, 176-184.

Decoster, W. (1998). *Akoestische kenmerken van de ouder wordende stem.* Leuven: Leuven University Press.

Prame, E. (1995). Measurement of the Vibrato Rate of Ten Singers. In P.H. Dejonckere, M. Hirano, & J. Sundberg (Eds.), *Vibrato* (pp. 121-140). San Diego: Singular.

Seidner, W., Nawka, T. & Cebulla, M. (1995). Dependence of the Vibrato in Pitch, Musical Intensity, and Vowel in Different Voice Classes. In P.H. Dejonckere, M. Hirano, & J. Sundberg (Eds.), *Vibrato* (pp. 83-92). San Diego: Singular.

On the role of the late rise and the early fall in the turn-taking system of Dutch

Johanneke Caspers

Universiteit Leiden

Abstract

The question posed in the present paper is whether subjects interpret a short utterance with a late non-prominent rise in pitch (LH%) as having a 'go on' function, prompting the current speaker to continue, whereas the same short utterance spoken with an accent-lending fall (H*L L%) is associated with finality, for example, with the answer to a yes-no question. A series of three perception experiments were run with natural data taken from Dutch Map Task dialogues. The results support the hypothesis that the LH% contour is associated with a 'go on' response, while the falling contour is associated with the answer to a question. Furthermore, LH% is preferred over H*L L% in contexts leading to backchannel responses, while there is no preference for either contour in question contexts. Finally, the LH% contour is acceptable in both context types, whereas the accent-lending fall is unacceptable in backchannel contexts.

1 Introduction

In everyday conversation there is generally a smooth and fast alternation of speaking turns, which can only be explained in terms of a highly complex system of interacting factors comprising syntax, semantics, pragmatics, prosody, visual cues, etc. My specific interest is in the function of one particular prosodic factor in the turn-taking process in Dutch, viz. speech melody.

In natural conversation so-called 'backchannels' (Yngve, 1970) are a common phenomenon. These are short optional utterances (for instance, *yes*, *hmhmm* or *okay*) produced by the current hearer to signal that (s)he is still engaged in the discourse, prompting the current speaker to go on. Communication is an interactional process, involving continuous feedback between interlocutors, and backchannels are important instances of responsive behavior. They signal that the information so far has been integrated into the common ground shared by speaker and listener (Clark & Brennan, 1991), and they also signal that the listener understands that the speaker has not finished yet. An utterance like *yes* can be used to indicate that the current listener has understood so far and that the speaker may continue with – for instance – giving directions. However, if the *yes* is a so-called conversational move, for example an answer to a yes-no question, it is not an optional utterance and therefore not a backchannel. It seems possible that the specific dialogue function of short utterances like *yes* is reflected in their suprasegmental characteristics.

In earlier investigations of a corpus of Dutch Map Task dialogues (task-oriented dialogues in which an 'instruction giver' has to explain to an 'instruction follower' how to draw a route on an unmarked copy of a map), backchannels were found to be often marked by a specific melodic configuration: a slight dip in pitch followed by a conspicuous rise, not lending overt prominence to the utterance, and therefore labeled as LH% (a label not present in the ToDI inventory, cf. Gussenhoven, Rietveld, Kerkhoff & Terken, 2003); the L stands for a low tone (not marking accent), and the H% for a high final boundary tone. The data revealed that the

majority of short utterances functioning as encouraging background signals carry such a LH% contour, while lexically identical 'real' turns – generally answers to yes-no questions – were marked by a pitch accent in the majority of the cases. Figure 1 presents the percentage of LH% contours versus the percentage of pitch accents as marked by two expert labelers (for more details see Caspers, 2000, 2003a, under revision).

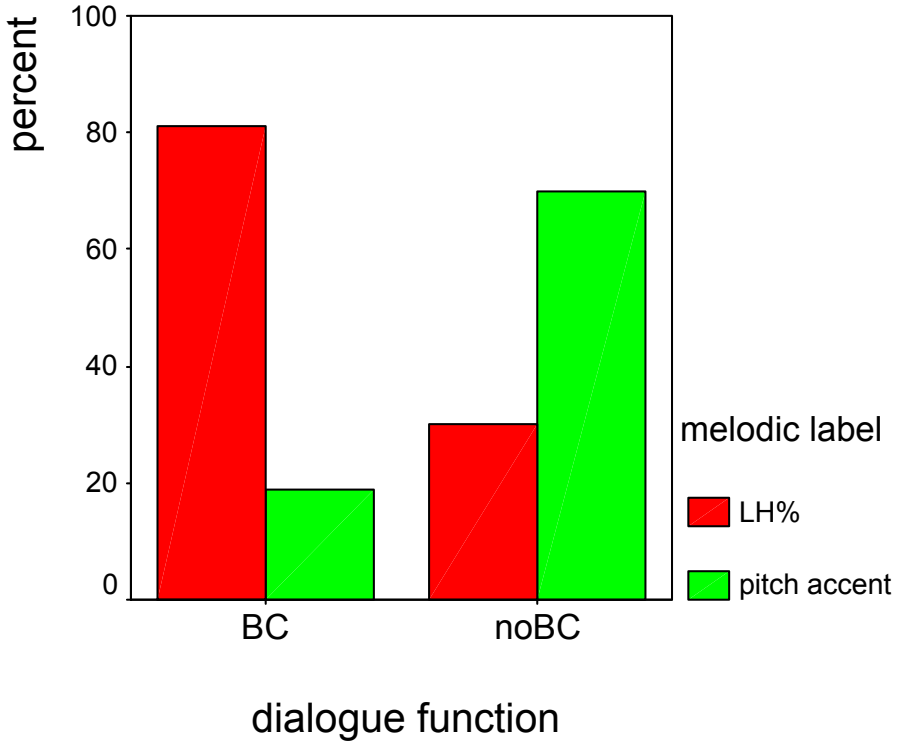


Figure 1. Percentage of LH% and pitch accent labels, broken down by dialogue function: BC (backchannel) versus noBC ('real' speaker turn).

This finding suggested that speech melody plays a role in signaling the dialogue function of short utterances like *yes* and *okay*, since there is a clear correspondence between backchannels and a non-prominent late rise and between 'real' speaker turns and a pitch accent. However, the LH% configuration does not seem to be an exclusive marker of backchannels, since it was found on approximately a third of the lexically identical 'real' turns as well. It could well be the case that LH% is essentially some sort of 'go on' signal, which suits backchannels in general, but may also fit certain 'real' speaker turns (for example, the answer to a yes-no question, which at the same time serves as an invitation to continue speaking; these kind of sequences are typical for Map Task dialogues, which essentially consist of one long instruction). The present perception experiment was designed to establish whether the LH% configuration is generally interpreted as a 'go on' signal in Dutch.

2 Approach

To be able to test the hypothesis that the LH% contour functions as a 'go on' signal in Dutch, it was contrasted with a contour that is supposedly not interpreted as such: the accent-lending fall (H*L L%), a contour typical for the positive answer to yes-no questions in the corpus

materials, and associated with finality (Caspers, 1998, 1999, 2003b; Ladd, 1996; Rietveld & Gussenhoven, 1995).[1]

Figure 2 presents examples of a typical H*L L% and a typical LH% contour.

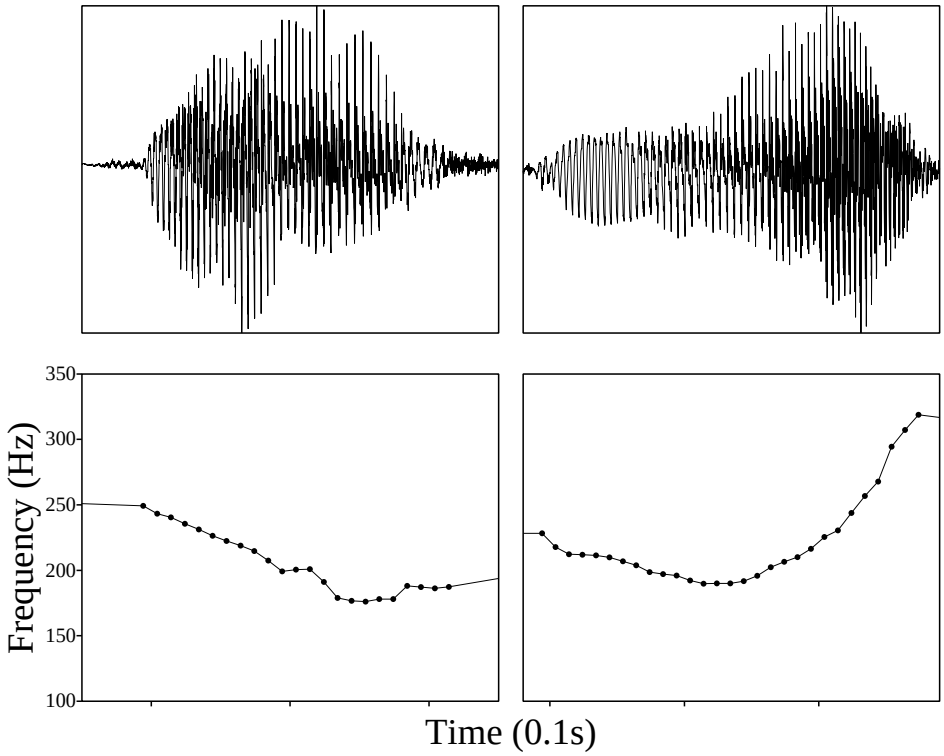


Figure 2. Examples of H*L L% (left) and LH% (right) contours on the word *ja* (yes); top: waveform, bottom: F_0-curve (in Hz).

Three sub-hypotheses were formulated:

1 Isolated utterances carrying LH% contours are associated with backchannels (since the main function of a backchannel is to signal to the speaker that the hearer is still there, which goes together naturally with 'go on'), in contrast with H*L L% contours, which will generally not be associated with a backchannel function.

2 In backchannel contexts there is a preference for LH% contours, in question contexts there is a preference for H*L L% contours.

3 LH% contours fit backchannel contexts as well as question contexts (because 'go on' is suitable as a positive reply to a yes/no question which is part of a larger instruction), while H*L L% contours will not fit backchannel contexts.

All combinations of the factors contour type (LH% vs. H*L L%) and dialogue function (BC vs. noBC) were available in the materials, albeit in different numbers. No manipulations of pitch were performed on the data, thereby preserving the naturalness of the stimuli. It also

[1] This fall is located early in the syllable, and is labeled 'A' in the Grammar of Dutch Intonation ('t Hart, Collier & Cohen, 1990). ToDI, the transcription system for Dutch intonation developed by Gussenhoven et al. (1999) uses the label H*L L% to refer to this type of contour, but H*L may also refer to a rising-falling pitch accent ('1&A' in the Grammar of Dutch Intonation).

means that, in addition to contour type, other prosodic and segmental information relevant to dialogue function may be present in the stimulus materials.

Only *ja*-utterances (yes) were used in the investigation, because they are the most frequently used lexical type of backchannel utterance (73% of the backchannels in the Map Task materials are *ja*'s), while they also occur most frequently as the (one-word) affirmative answer to a question (71%, Caspers, under revision).

For each hypothesis a separate experiment was conducted. To test hypothesis 1, isolated *ja*-utterances were presented to subjects, varying contour type (LH% versus H*L L%) as well as their original dialogue function (backchannel versus answer to a question). Subjects had to indicate whether they thought the *ja* was originally uttered as an optional background signal, prompting the current speaker to continue, or whether it was uttered as the positive answer to a yes-no question.

To test hypothesis 2, pairs of different *ja*-utterances were presented in a specific context, asking the subjects to select the utterance best fitting the given context. In the pairs of utterances either the contour type (LH% versus H*L L%), the original dialogue function (backchannel or answer), or both were contrasted; in each pair one of the two *ja*-utterances was originally produced in the presented context.

To test hypothesis 3, different combinations of a context and a *ja*-utterance were presented, asking subjects to rate the acceptability of each combination. As in the other parts of the experiment, contour type and dialogue function were systematically varied, leading to a combination of context and *original ja*-utterance in a quarter of the cases.

3 Method

3.1 Stimulus materials

All *ja*-utterances available in the Map Task materials (ca. 40 minutes of task-oriented Dutch dialogue, see Ladd, Schepman, Mennen & Lickley, 2001) were inspected for their usefulness in the present experiments. All cases of overlapping utterances were excluded, as well as all cases with immeasurable pitch contours.

The contexts were cut in such a way that they contained enough information for the subjects to determine the dialogue function of the immediately following utterance (i.e., to decide whether the *ja* was an optional backchannel or a non-optional answer to a yes-no question); their durations varied between 4.5 and 11 s.

3.2 Subjects

Twenty-four native speakers of Dutch participated in the experiment. They were paid a small fee. Fourteen were female, and their ages varied between 19 and 61.

3.3 Procedure

The interactive experiment was made accessible on the internet.[2] The contexts and stimuli to be judged were presented auditorily; subjects could press the relevant buttons as often as they found necessary.

Part A: subjects were presented with 24 different versions of *ja*. After (repeatedly) listening to each stimulus, they had to click a response button named either 'go on-signal' or 'answer

[2] The program for interactive stimulus presentation and response collection was written by Jos J.A. Pacilly of the Universiteit Leiden Phonetics Laboratory.

to a question'. The order of the stimuli and the spatial location of the two response buttons was blocked over subjects.

Part B: subjects were presented with 24 combinations of a context and two possible continuations, one of which they had to select as the best fitting. The order of the stimuli was reversed for half of the subjects.

Part C: subjects were presented with 16 combinations of a context and a *ja*-utterance, and judged the acceptability of each combination on a ten-point scale (in the Dutch educational system values 1 to 5 represent degrees of inadequacy, whereas values 6 to 10 represent degrees of adequacy; the boundary between acceptable and unacceptable is drawn at 5.5). The order of the stimuli was reversed for half of the subjects.

4 Results

4.1 Part A

In part A the subjects had to listen to a series of *ja*-utterances carrying either an LH% or an H*L L% contour, and indicate for each stimulus whether they thought it was uttered as a 'go on' signal or as the answer to a question, expecting an association between LH% contours and backchannels and between H*L L% contours and answers. The results are presented in table 1.

Table 1. Absolute (and relative) frequency of 'go on' and 'answer' responses for the LH% and H*L L% contours, broken down by original dialogue function (backchannel vs. answer to question).

		response		
contour type	original dialogue function	go on	answer	total
LH%	backchannel	117 (81%)	27 (19%)	144
	answer to question	125 (87%)	19 (13%)	144
	total	242 (84%)	46 (16%)	288
H*L L%	backchannel	52 (36%)	92 (64%)	144
	answer to question	26 (18%)	118 (82%)	144
	total	78 (26%)	210 (73%)	288
total		320 (56%)	256 (44%)	576

The table shows that in 84% of the cases a late-rising (LH%) contour is associated with a 'go on' function, while a falling contour (H*L L%) is associated with an answer in 73% of the cases (χ^2=189.11, p<.001), providing support for hypothesis 1. For the LH% contours there does not seem to be an additional influence of the original dialogue function of the stimulus: even when the stimulus functioned as the answer to a question in its original context, subjects associate it with a 'go on' function in 87% of the cases (χ^2=1.66, n.s.). However, for the accent-lending falls there does seem to be such an effect: when the stimulus originates from a *backchannel* context, the subjects associate it with an answer in 64% of the cases, but the association rises to 82% when the stimulus originally functioned as an answer to a question (χ^2=11.89, p<.001). This may mean that the association between an LH% contour and a 'go on' function is stronger than the association between H*L L% and the answer to a question. But it could also be the case that the *ja*-utterances carrying a falling contour contain more information referring to their original dialogue function than the utterances with a late rising contour.

4.2 Part B

In part B the subjects were presented with two different *ja*-utterances in a specific context. Their task was to select the one best fitting the given context, expecting the LH% contours to be preferred in backchannel contexts (part of a MapTask dialogue ending in a statement like *Then you take a right turn*) and the H*L L% contours to be preferred in answer contexts (a dialogue fragment ending in a question, e.g., *Do you have a stranded whale?*). Note that there is always a change of speaker between the end of the context and the *ja*-utterance.

In a third of the cases the two *ja*-stimuli from which the subjects had to choose did not differ in contour type, but only in their original dialogue function: backchannel or answer (and one of the two originally occurred in the presented context, as was the case for every stimulus pair). These cases were used to establish if subjects were able to hear which of the two stimuli is originally taken from the presented context. When the two *ja*'s differ in dialogue function (backchannel versus answer), but not in contour (*N*=192), there is a preference for the *ja* that was originally uttered in that context in 75% of the cases. This stimulus effect is larger than expected; it means that there is enough prosodic and/or segmental information available in the data for the subjects to connect the *ja* to its original context in three-quarters of the cases. At present it is not clear what kind of information this would be exactly.

The next question was: is there an *additional* effect of contour type? In table 2 the preference scores are presented for those cases where there was an opposition in contour type between the two alternative *ja*-utterances (*N*=384), making a distinction between utterances originating from the context presented and utterances originating from another context.

Table 2. Absolute (and relative) frequency of preference for LH% and H*L L% contours, for stimuli that were presented in their original context versus stimuli presented in a non-original context, broken down by context type.

	preference for contour			
	LH%		H*L L%	
context type	original	non-original	original	non-original
BC	88 (92%)	50 (52%)	46 (48%)	8 (8%)
noBC	68 (71%)	16 (17%)	80 (83%)	28 (29%)

Bearing in mind that there is a stimulus bias towards the original *ja*, an additional effect of contour type can be observed of approximately 20% for the LH% contours in backchannel contexts (a 92% preference when the LH% was produced in the presented context, i.e., 17% above the bias, and a 52% preference when the LH% was taken from another than the presented context, i.e., 27% above the bias). In the question (noBC) contexts the responses are roughly at the level of the stimulus bias (83% and 29%, which amounts to 8% and 4% above the stimulus bias respectively), which means that there is no clear additional influence of contour type in these cases. Presumably both contours fit these answers equally well and the original utterance – whose status as such is clear from other prosodic and/or segmental information – therefore wins. In the case of a backchannel there is indeed a preference for one of the two contour types, albeit a small one. This may mean that an LH% contour is acceptable as the answer to a question in the current materials.

4.3 Part C

In the third part of the test the subjects were presented with the same BC or noBC contexts, this time combined with just one *ja*-utterance, varying original dialogue function (BC or

answer) and contour type (LH% and H*L L%). The subjects had to rate the acceptability of each combination of context and *ja* on a scale from 1 to 10. The LH% contour was expected to be acceptable in both context types, whereas the H*L L% contours were predicted to be acceptable in answer contexts only.

Table 3. Mean acceptability (and standard deviation) of contour type, broken down by context type.

	contour type presented		
context type	LH%	H*L L%	total
BC	6.1 (3.0)	4.8 (2.2)	5.4 (2.7)
noBC	5.7 (3.3)	6.3 (2.6)	6.0 (3.0)

The results presented in table 3 support the hypothesis that an LH% contour is acceptable in a backchannel context (a mean score of 6.1) and that an H*L L% contour is acceptable in a question context (6.3); furthermore, an LH% contour is – marginally – acceptable in a question context (5.7), whereas an H*L L% contour is clearly unacceptable in a backchannel context (4.8). Overall, the acceptability scores are rather low, and further inspection of the data (again) shows a large effect of the originality of the stimulus:

Table 4. Mean acceptability (and standard deviation) of contour type for stimuli that were presented in their original context versus stimuli presented in non-original context, broken down by context type.[3]

	contour type presented			
	LH%		H*L L%	
context type	original	non-original	original	non-original
BC	7.9 (1.9)	5.5 (3.1)	5.7 (2.4)	4.5 (2.0)
noBC	8.6 (1.4)	4.7 (3.1)	8.8 (1.2)	5.4 (2.4)
count	46	142	48	141

Table 4 reveals that stimuli presented in a non-original context are rated much lower than stimuli presented in their original contexts. An analysis of variance with fixed factors *contour type*, *originality of stimulus* and *context type* shows a small main effect of *context type* [$F(1,375)=4.7$, $p<.05$], a large effect of *originality* [$F(1,375)=83.4$, $p<.001$], interaction between *context type* and *contour* [$F(1,372)=13.1$, $p<.001$] and between *context type* and *originality* [$F(1,372)=9.2$, $p<.005$].

The main effect of context type – *ja*'s presented in question contexts are on average perceived as more acceptable (a mean score of 6.0) than *ja*'s in backchannel contexts (a mean score of 5.4) – could well be caused by the fact that some of the backchannel contexts are clearly unfinished (in content, but sometimes prosodically as well), while the question contexts are neatly finished off by a positive reply.

The main effect of originality was not expected to be this large: original *ja*'s get a mean acceptability score of 7.8 (*sd* 2.2), while non-original *ja*'s have an average acceptability score of 5.0 (*sd* 2.7).

[3] Because of a mistake by one of the subjects, seven cases are missing from the dataset.

The interaction between *context type* and *contour type* was predicted, but the main effect and interaction regarding the originality of the stimuli were not. However, there is no three-way interaction between *context type*, *contour type* and *originality*, which indicates that the predicted effect is present in the data, irrespective of the influence of the originality of the stimulus.

Support for hypothesis 3 can only be found for the stimuli presented in their original contexts: stimuli with LH% contours are amply acceptable in backchannel contexts as well as question contexts (mean scores of 7.9 and 8.6, respectively), whereas stimuli with a falling contour are fully acceptable *only* in a question context (a mean score of 8.8); when appearing in a backchannel context, they are judged as – comparatively – unacceptable, despite the fact that the stimulus was presented in its original context (a mean score of 5.7). For the stimuli presented in non-original contexts there is a trend toward higher acceptability for LH% contours in backchannel contexts and for H*L L% contours in question contexts, but the overall acceptability of these stimuli is below 6. As in parts A and B, this means that the stimuli must contain information that reveals that they were taken from another context than they were presented in.

5 Discussion and conclusion

Summarizing the results, subjects clearly associate a late-rising contour (LH%) with a 'go on' function and an early accent-lending fall (H*L L%) with an answer to a question, the latter association being a little weaker. Furthermore, subjects prefer an LH% contour over an H*L L% contour in a context leading to a backchannel, which is compatible with the proposed function of contour LH% as signalling 'go on', while there is no clear preference for either contour type in a question context. Finally, an LH% contour is acceptable in both context types, whereas the accent-lending fall is unacceptable in a backchannel context (even if it was originally produced there), supposedly because the fall puts the *ja* itself in focus, and because the fall is associated with finality, which does not suit a real backchannel very well.

The results showed an unexpectedly large influence of prosodic and/or segmental characteristics other than contour type: in the turn-changing contexts there was no clear preference for a specific contour type, but there was a clear preference for the stimulus originally uttered in that specific context, and only the original stimuli were judged to be generally acceptable. This means that subjects were able to determine whether or not a stimulus was presented in its original context in a majority of the cases, probably on the basis of varying combinations of prosodic characteristics of the stimulus itself (voice quality, loudness contour, pitch range, duration, segmental characteristics, etc.), as well as information contained in the transition between the end of the presented context and the following stimulus (despite the fact that there was always an intervening pause).

However, the predicted association between LH% and 'go on' is clearly visible in the data, which may well explain why the LH% contour, which typically appears on backchannels, may also appear on certain non-optional 'real' turns. In contrast, the accent-lending fall does not seem to be a very appropriate contour for backchannels, supposedly because this contour is associated with prominence and finality.

The proposed functions of the two contours investigated appear to be opposites at first sight, but they certainly do not exclude one another. In light of this fact the presented results seem quite clear.

References

Caspers, J. (1998). Experiments on the meaning of two pitch accent types: the 'pointed hat' versus the accent-lending fall in Dutch. In *Proceedings of the International Conference on Spoken Language Processing, Sydney* (pp. 1291-1294).

Caspers, J. (1999). The early versus the late accent-lending fall in Dutch: Phonetic variation or phonological difference. In *Proceedings of the International Congress of Phonetics Sciences, San Francisco* (pp. 945-948).

Caspers, J. (2000). Melodic characteristics of backchannels in Dutch Map Task dialogues. In *Proceedings of the International Conference on Spoken Language Processing, Beijing* (Vol. II, pp. 611-614).

Caspers, J. (2003a). Local speech melody as a limiting factor in the turn-taking system in Dutch. *Journal of Phonetics*, *31*, 251-276.

Caspers, J. (2003b). Phonetic variation or phonological difference? The case of the early versus the late accent-lending fall in Dutch. In: J. van de Weijer, V.J. van Heuven & H. van der Hulst (Eds.), *The Phonological Spectrum. Volume II: Suprasegmental structure* (pp. 201-223). Amsterdam/Philadelphia: John Benjamins.

Caspers, J. (under revision). Melodic characteristics of backchannels in Dutch task-oriented dialogues. *Speech Communication*.

Clark, H.H., & Brennan, S.E. (1991). Grounding in communication. In: L.B. Resnick, J.M. Levine, & S.D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127-149). Washington DC: American Psychological Association.

Gussenhoven, C., Rietveld, T., Kerkhoff, J., & Terken, J. (2003). *ToDI, Transcription of Dutch Intonation* (second edition). Available: http://todi.let.kun.nl/ToDI/home.htm.

Hart, J. 't, Collier, R., & Cohen, A. (1990). *A perceptual study of intonation*. Cambridge: Cambridge University Press.

Ladd, D.R. (1996). *Intonational Phonology*. Cambridge: Cambridge University Press.

Ladd, D.R., Schepman, A., Mennen, I., & Lickley, R.J. (2001). Final report on research activities and results for ESRC project No. R000-23-7447 'Alignment of Fundamental Frequency Targets in English and Dutch'. Available: www.ling.ed.ac.uk/eprints/.

Rietveld, T., & Gussenhoven, C. (1995). Aligning pitch targets in speech synthesis: effects of syllable structure. *Journal of Phonetics, 23*, 375-385.

Yngve, V. (1970). On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*. Chicago: Chicago Linguistic Society.

There's many a slip 'twixt the cup and the lip

Anne Cutler

Max Planck Institute for Psycholinguistics

and

Caroline G. Henton

University of California, Santa Cruz

Abstract

The retiring academic may look back upon, *inter alia*, years of conference attendance. Speech error researchers are uniquely fortunate because they can collect data in any situation involving communication; accordingly, the retiring speech error researcher will have collected data at those conferences. We here address the issue of whether error data collected in situations involving conviviality (such as at conferences) is representative of error data in general. Our approach involved a comparison, across three levels of linguistic processing, between a specially constructed Conviviality Sample and the largest existing source of speech error data, the newly available Fromkin Speech Error Database. The results indicate that there are grounds for regarding the data in the Conviviality Sample as a better than average reflection of the true population of all errors committed. These findings encourage us to recommend further data collection in collaboration with like-minded colleagues.

1 Life just fings through your slippers

A long lifetime in academia translates to a long history of convivial occasions. As David Lodge so accurately portrayed in *Small World* (Lodge, 1984), the successful academic life is not passed exclusively in the laboratory, library, or university classroom. Academics travel, above all to attend conferences; at conferences they meet colleagues whom they have not seen since the last such occasion, and the pleasure of these reunions naturally warrants the raising of a celebratory glass. The retiring academic can therefore look back not only on armies of past students, compendia of experimental results, and stacks of published papers, but also on a rich array of memories of happy conviviality, sharing with old friends a select Sancerre in Stockholm, a parade of Pinots in Philadelphia, or a superlative Sebastiani in San Francisco.

Conferences count as work time - academic CVs list conference attendance - and so these agreeable happenstances can be reckoned a perk of the job. All academics enjoy this fortune. But some academics possess a further advantage: their research field is one that can be pursued outside the lab, for instance during a conference, just as well as it can in the normal work setting. Speech researchers are among these happy few. Speaking is a necessary component of conference communication. Thus whatever the communicative occasion, and most especially at conferences, speech researchers find themselves surrounded by a flow of speech — that is, by a flow of research data. Not only are speech researchers at conferences working by making professional contacts, presenting their newest results, and learning about developments in the field, but they may also be working in amassing the very material on which a future presentation will be based.

Most fortunate of all, even among speech researchers, are those who study slips of the tongue. Where there is speech, there are slips. But some occasions produce more slips than others. Factors known to induce an increase in error frequency include, for instance, anxiety and stress — the stress which might accompany an oral presentation at a major conference, for example. Thus a speech researcher with an interest in slips of the tongue might be even more likely to be presented with data at a conference than at home in the lab! Yet another factor exhibiting a very strong causal relationship with speech error frequency is alcohol. Consider, then, the uniquely blessed situation of the speech error researcher: the wines in Stockholm, Philadelphia, and San Francisco not only contribute to wonderful memories of conferences past, they generate wonderful data! Who would choose any other field?

2 Through a glass darkly?

The only cloud hanging over this idyllic prospect is the possibility that errors produced under conditions of conviviality might constitute a seriously biased set, quite unlike errors produced in less lively circumstances. 'Drunk', as *Webster's Unafraid Dictionary* (Levinson, 1967) reminds us, is the future tense of 'drink'. The wine might generate intoxication, in short, and the intoxication might generate speech errors that tell us much about the difficulty of articulating after drinking, but nothing about the speech error researcher's main focus of interest, i.e. the process of speech production in general. We consulted the literature on this question, in an attempt to determine the degree to which speech errors perpetrated by talkers who have consumed alcohol are representative of speech errors in general.

The principal difference between errors in alcohol-affected speech and errors in sober speech is to be found, as we forecast above, in their quantity: the former are far more numerous. Stemberger, Pisoni and Hathaway (1985) conducted an experiment to quantify this, using a laboratory method for error induction developed by Motley and Baars (1975). In this method, participants read aloud visually presented word pairs under time pressure. If several pairs sharing an aspect of phonological structure (e.g., *big dog*, *bad deal*, *both days*) are followed by a pair with a contrasting structure (e.g., *dim bar*), an error is quite likely to result (in this example, a phoneme exchange in word-initial position between the two words of the pair). Stemberger et al.'s subjects undertook this task twice, once when they were sober and once when they were intoxicated (with blood alcohol levels of 0.1 promille or higher); they made nearly twice as many errors in the latter condition.

The articulatory differences between intoxicated and sober speech have been extensively studied, despite the repeated finding that these differences show great individual variability (Klingholz, Penning & Liebhardt, 1988; Hollien & Martin, 1996). All researchers report that the most noticeable difference is a slower rate of speech in the intoxicated (Lester & Skousen, 1974; Johnson, Pisoni & Bernacki, 1990; Behne, Rivera & Pisoni, 1991; Hollien & Martin, 1996; Chin, Large & Pisoni, 1997). Fundamental frequency has been reported to rise with intoxication (Hollien & Martin, 1996), but also simply to become more variable (Behne & Rivera, 1990; Chin, Large & Pisoni, 1997). Differences in voice quality after drinking are also noticeable to trained observers (Künzel, Braun & Eysholdt, 1992); but again, individual differences are so great that it is not possible to predict the specific effects of alcohol on phonation. Place of articulation may be affected, with more posterior articulations being favored (Behne & Rivera, 1990), presumably because opening the mouth and articulating clearly demands more energy than an intoxicated talker can muster.

None of these observations, however, can provide definitive evidence on the issue of whether speech errors under intoxication are merely quantitatively, or also qualitatively different from

speech errors perpetrated by sober talkers. Some relevant evidence emerged from Stemberger et al.'s (1985) error induction study. Although all the subjects in this study made errors of the predicted, i.e. 'induced', kind, they also made other errors, and the greater part of the increase in error rate under intoxication occurred in this latter category. In particular, Stemberger et al. observed many cases of perseveration, with certain phoneme sequences persisting (erroneously) across a sequence of trials. Stemberger et al. related these findings to several aphasic syndromes resulting from brain damage, suggesting that intoxication may be viewed as induced cortical dysfunction. However, they did not report any types of error that do occur under intoxication but are not observed under other circumstances. This suggests that there are no qualitative differences between errors that occur in intoxicated versus sober speech.

There are, of course, some ways in which a percipient speech error researcher may distinguish informative from uninformative errors. It has often been lamented (see, e.g., Cutler, 1981, 1988) that many speech errors are ambiguous as to how they should be classified. Thus a speaker who utters the word *dignify* when the intended word was *signify* may be making a word substitution error, but might just as well be making a phoneme substitution error. In the case of intoxicated speech, ambiguity may arise between a phoneme substitution versus an inability to articulate a (correct) target; did the utterance *shlip* arise because of a substitution of the phoneme /ʃ/ for /s/, or because the speaker simply could not achieve an [s]? Such ambiguity makes statistical study of error types difficult, and it leads further to theoretical complication. For instance, it has been argued (e.g., by Dell and Reich, 1981) that phonemic slips are more likely than would be expected by chance to lead to real words, and that this is evidence for interaction between levels of processing in speech production. This argument cannot be put to the test at all unless there is some way of ascertaining for a slip like *dignify* for *signify* whether it arose at the lexical or phonemic level of the production process. Shattuck-Hufnagel and Cutler (1999) pointed out that one potential way to distinguish between these two alternative source levels is to look at error correction patterns, which have long been known as a source of useful insights (Nooteboom, 1981). The prosodic characteristics of corrections of word-level and phoneme-level errors differ, in that speakers apply greater contrastive accent to corrections of the former type of error. In an analysis of a small corpus of errors which were ambiguous between these two levels, together with errors unambiguously arising from each level, Shattuck-Hufnagel and Cutler found that their ambiguous set strongly resembled the phoneme-level set and differed from the word-level set. This approach to distinguishing the source level of ambiguous errors has been continued more recently by Nooteboom (in press).

Of course, such analyses require that sound files of the errors in question be available. When this is the case, there are also analogous criteria by which researchers can distinguish phonemic from articulatory sources of errors made under intoxication. (An extended discussion of potential criteria for making this distinction is to be found in Chin and Pisoni, 1997: Chapter 4.) Lester and Skousen (1974) showed that word-final devoicing errors were common in English spoken by the intoxicated. However, one of the main correlates of voicing distinctions in word-final position in English is preceding vowel duration. As Lester and Skousen showed, the vowel duration of, for example, *tease* spoken as *teace* by an intoxicated talker more closely resembled that speaker's other finally-voiced syllables than the finally-voiceless syllables, suggesting that the devoicing did not indicate a phonemic substitution (/s/ for /z/) but rather arose from difficulty of articulation. In a similar manner, one could use detailed phonetic analysis to clear up multiply ambiguous cases, such as an English speaker's pronunciation of *sip* as [si:p], which allows the potential interpretations that

it is a word substitution (*seep* for *sip*), a phoneme substitution (/i/ for /ɪ/), or a contextually prompted intrusion of the name of the drinking partner (*Sieb*).

3 Sips and slips

There's many a slip 'twixt the cup and the lip; but the cup is in fact not the serious researchers' instrument of choice when conducting research into alcohol-induced behaviour. A cup, be it finest bone china, delicate finger-crooking punchbowl variety, cheap party plastic, or a silver-plated and double-handled trophy (the capacity is attractive but the metal taints the taste), is in fact our least favourite vessel. We prefer a glass, bottle, flagon, magnum, Jeroboam, Rehoboam, Methusulem, Salmanazar, Balthazar, case, Nebuchadnezzar, firkin, barrel, or hogshead; or, best of all, a butt (108 imperial gallons) from which to obtain as many sips, and as many slips of the tongue, as we can. Nor, of course, is the lip the only organ involved in slips of the tongue. For that matter, neither is the tongue. Slips of the tongue are by no means simply lingual: they are made with many other places and organs of articulation. It is possible to have many a slip 'twixt the cup and the lip(s), upper teeth, the tongue tip, the tongue blade, the alveolar ridge, the tongue front, the post-alveolar region, the tongue dorsum, the hard palate, the tongue back, the velum, the uvula, the pharynx, and even as far down the vocal tract as the glottis. (It is, of course, hardly coincidental that Les Chevaliers du Vin and master sommeliers world-wide use most of the lingual landmarks above when classifying the initial burst of flavours and the after-taste in a wine-tasting.) Others might wish to add further exotica such as the tongue root, and the epiglottis; although all things being equal, slips involving the epiglottis would be quite hard to swallow.

Phonetic and phonological slips can further involve substitutions of airstream mechanisms and manners of articulation. It would be fascinating to find documented speech errors involving Sindhi implosives, Amharic ejectives, or Bantu clicks — e.g. the production of a velaric airstream bilabial click /⊙/ instead of an alveolar lateral click /‖/. To our knowledge, no such data is available — yet. But we can posit that were an ejective error made while sipping, the results could be quite messy.

The consumption of alcohol can lead to what Trojan and Kryspin-Exner (1968) aptly termed "general lingual dissolution". Instrumental analysis of dissolute speech has shown that lay terms for drunken speech, such as 'thick', 'drawn out', 'drawled', and 'slurred' correlate well with articulatory abnormalities, partial articulations, and maladjustments. 'Slurred' speech is slower, weakened (lenided), palatalized, and segmentally longer and imprecise. In the next section we examine phonetic misfits more closely, with particular regard for the phonetic features of the most common phonological errors.

4 The pH value of slips

Phonetic and phonological analyses (jointly known as the 'Ph' disciplines within linguistics) yield many examples of speech errors; the Ph domains also yield the most widely-studied speech errors. By far the most frequent Ph error unit is the single segment, which often corresponds to a phoneme; the error may also be attributed to a feature switch. Although recent research (Mowrer & Mackay, 1990; Nooteboom, 2004) has cast doubt on the integrity of phonemes as units in erroneous performance, many oft-reported patterns are established. Nooteboom (1967, 1973) and Shattuck (1975) reported early in the study of segmental errors that consonant slips are 33-66% more common than vowel slips. Errors involving consonant clusters are less frequent than those involving single consonants, largely because the syllable

structure of languages favours CVC rather than CC(C)VC(CCC). Clusters are nevertheless often split. VC errors occur more frequently than CV slips; feature errors (such as tranposing voicing in stops) are quite rare. Dispute has raged for 30 years about the syllable as an error unit. Syllables may be lost, or repeated, but only rarely are they reversed.

Phonetic similarity between the affected elements characterizes speech errors in all languages for which there is substantial data. The frequency hierarchy for errors is place of articulation (most errors), manner of articulation, voice (fewest). Among the manners of articulation, errors in stops are most frequent, followed by fricatives. The lower the frequency of occurrence of a unit, the more likely will it be error-prone. The liquids /l/ and /r/ interact in a particularly facile way (Nooteboom, 1967; we note that no study has yet investigated whether the liquids are disproportionately affected by liquid intake).

Like beer and wine, vowels and consonants rarely mix. If vowels are substituted for one another, then monophthongs replace monophthongs, and diphthongs replace diphthongs. Seldom do diphthongs disintegrate into their vocalic parts. The most productive region (the 'hautes côtes') for all Ph slips is the word- or syllable-initial position, where 70–80% of all errors can be found. Identical syllable position in exchanges is another powerful factor: Nooteboom (1967) and others have found this to hold for 80–100% of exchange cases.

So in trying to select an error with an optimal Ph value, the speech researcher should look for single consonants in initial position. These will yield the best, the oldest, and the most representative varietals, er, variants. As a sommelier would say, they are the most rewarding 'in the mouth'.

5 I'm not as thunk as people drink I am

Dr. William Spooner, 'founding slipper' of the speech error field, is said to have raised his glass and proclaimed the toast "Let us drink to the queer old Dean." This error has not gone into the literature marked "drunk" — who would dare assume that of the venerable Reverend Doctor? But some speech error collections do contain errors marked "drunk". And all speech error collections contain errors that occur in utterances having to do with eating, drinking, and general conviviality. Rudolf Meringer (Meringer & Mayer, 1895/1978), for example, reports a similar example involving the proposing of a toast: *Ich fordere Sie auf, auf das Wohl unseres Chefs aufzustossen!* ('I call on you to belch to the health of our boss'; the target verb was *anzustossen,* 'raise your glasses'). We have conducted a search in the collection of the first author and compiled a list of as many such errors as we could find. We included all errors that involved an eating or drinking topic, plus all errors marked to indicate that the speaker was drunk at the time (or at least on the way to becoming so). In this section we attempt to compare the characteristics of this list (which we will refer to henceforth as the Conviviality Sample) with characteristics of a much larger collection taken to be the standard for the field.

Our work in this section drew on the publicly available web version of the Fromkin Speech Error Database (and was thus the first study to make use of this version of the Fromkin corpus). The database contains over 7500 slips of various kinds collated from collections made by many of the leading speech error researchers, most notably Vicki Fromkin's collection of errors. It is possible to search in this database using a number of search criteria. For instance, one can search by speaker name; Table 1 lists the complete output which the database currently produces for a random three-name sample which we chose to input. (The sample consisted of two native English speakers and one non-native speaker, whose native language is Dutch. All errors found in this search were in English. We note that a study by

Poulisse [1999] found that non-native speakers' errors outnumbered native speakers' errors by 14.5 to 1; on the basis of this comparison, we may conclude that the speech of Table 1's non-native speaker is atypically low in errors).

Table 1. Sample output from the Fromkin Speech Error Database.

Speaker	Target	Error
Nooteboom	EVERYTHING you hear	EVERYHING you hear
Nooteboom	to CUT him SHORT	to SHUT him COURT
Nooteboom	I prefer to RESERVE	I prefer to PRESERVE
Nooteboom	what does that SIGNIFY	what does that DIGNIFY
Cutler	In the form of three CRIPPLED human beings	In the form of three KIPPLED human beings
Henton	and peter said, caroline, couldn't you find something yellow in your WARDROBE?	and peter said, caroline, couldn't you find something yellow in your AUDIENCE?

It is also possible to search on error type, and we used this latter function in comparing the contents of our Conviviality Sample with those of the database as a whole. We tallied the number of errors in English (plus a very few in German) in the database which were labelled as unambiguously phonological, morphological, lexical or supralexical (the latter category being an amalgam of several categories involving phenomena above the word level). Because the morphological category contained (both in the database and in our Conviviality Sample) relatively few members, we collapsed it together with the phonological category into a new category, sublexical, giving a three-way split corresponding to phenomena arising below, at, and above the level of the word.

The size of the sample which we extracted from the Fromkin Speech Error Database was 3601 errors, of which 3588 were in English and 13 in German (this does not mean that all the rest of the errors in the database are ambiguous; some are ambiguous, but some are not classified; further, quite a large number are in other languages such as French or Italian, and quite a number also fall into categories which we excluded here, such as Tip of the Tongue). The size of our Conviviality Sample was 98 errors (95 in English and three in German). Table 2 shows the proportions of errors in each sample falling into the three broad categories we used in our comparison.

It can readily be seen that while the proportion of lexical-level errors in the two samples is quite comparable, the samples differ in that the Conviviality Sample contains relatively fewer sublexical errors and relatively more supralexical errors than the sample we extracted from the database. What does this pattern imply regarding the representativeness of errors made under conditions of conviviality?

Table 2. Proportions of errors in the Fromkin Speech Error Database and in the Conviviality Sample categorized as sublexical, lexical, or supralexical.

	Sublexical	Lexical	Supralexical
Fromkin Speech Error Database	62.6	27.7	9.7
Conviviality Sample	55.1	29.6	15.3

First, it suggests that the pH value of convivial slips is *not* higher than desirable. Just as a wine's stability and colour are dependent on its pH value (ideally in the range 3.0–3.5), so is our analysis critically affected by the balance of sublexical (Ph) errors compared with the other types of errors. The overall Ph proportion in our sample is lower than in the database; as a result, we can rest assured that the errors will keep well, and that no 'acid tongue' will result from over-indulgence in liquid pleasures. It is also noteworthy that the ratio of sublexical to supralexical errors in the Conviviality Sample is roughly 3.6:1; a number that equates with a near-ideal pH value for wine! If the Ph value were too high, then greater oxidation, undesirable bacterial fermentation, and poor colour might occur. Whether these effects would be detectable in the convivial slips, or in the speakers themselves, we can only conjecture. We are however certain that convivial speech may disintegrate phonetically just as badly as a wine with too high a pH value.

Second, the pattern we have observed prompts a very interesting conclusion. Note that there is always an element of reporter bias in collections of slips of the tongue. Speech error researchers write down the errors that they hear, but not necessarily all the errors they hear — speech error researchers are only human, and they write down only the errors they notice. An attempt to compare the patterns of errors which are noticed by researchers with the patterns which actually occur, and thus to quantify the extent of the reporter bias, was made by Ferber (1993). She had collected a quite substantial corpus (roughly 1000 errors) by the usual means of jotting down everything that had come to her and her colleagues' attention over a period of approximately a year. She compared this corpus with another corpus of closely comparable size created from a painstaking transcription of recorded radio discussions. Her results, analysed according to the three broad categories which we have used here, showed that the distributions in her two samples differed. The proportion of lexical errors exhibited the least difference across the samples; but the proportion of sublexical errors was significantly lower, and the proportion of supralexical errors significantly higher in the 'true' corpus (based on the recorded material) than in the corpus collected by traditional methods.

This is exactly the asymmetry that our Conviviality Sample displays in comparison to the sample from the larger database. We therefore propose that errors collected under conditions of conviviality are actually *better* than average reflections of the true population of all errors committed. Speech error researchers may thus justifiably devote particular care and attention to the collection and evaluation of such corpora. In consequence, it is with undiluted pleasure that we recommend further empirical investigation of this kind, in the firm anticipation that the preliminary findings reported here will be corroborated. Future studies should include sampling from a larger population of bibulous utterances, and extending the international, cross-linguistic, and cordial sources of the database. Cheers!

Acknowledgements

We thank Jack Fromkin and Bill Sloman for comments on this paper and for extensive assistance with essential preliminary research on the topic under investigation. Research support was provided by numerous institutions and funding agencies over three decades.

This paper is for Sieb Nooteboom, but we are sure that he will approve a further dedication: to the memory of Vicki Fromkin, and countless occasions of shared conviviality with her.

The Fromkin Speech Error Database referred to in section 5 was collected over many years, and was converted (along with eight other error corpora) to computer-readable form at UCLA with support from a National Science Foundation grant to Vicki Fromkin. At the time of Vicki's death in January 2000, the wider availability of the database was in doubt because

there was no longer support for the software format used to convert it. As described above, the database is now publicly available (www.mpi.nl/world/corpus/sedb), thanks to a grant from the Max Planck Society enabling its further conversion to XML format. The work was carried out by Hansje Braam under the supervision of Sieb Nooteboom.

References

Behne, D. M., & Rivera, S. M. (1990). Effects of alcohol on speech: Acoustic analysis of spondees. *Research on Speech Perception*, Progress Report, 16 (pp. 263-291). Bloomington, IN: Speech Research Laboratory, Department of Psychology, Indiana University.

Behne, D. M., Rivera, S. M., & Pisoni, D. B. (1991). Effects of alcohol on speech: Durations of isolated words, sentences, and passages in fluent speech. *Research on Speech Perception*, Progress Report, 17 (pp. 285-301). Bloomington, IN: Speech Research Laboratory, Department of Psychology, Indiana University.

Chin, S. B., Large, N. R., & Pisoni, D. B. (1997). Effects of alcohol on words in context: A first report. *Research on Spoken Language Processing,* Progress Report, 21 (pp. 403-420). Bloomington, IN: Speech Research Laboratory, Department of Psychology, Indiana University.

Chin, S. B., & Pisoni, D. B. (1997). *Alcohol and Speech*. San Diego: Academic Press.

Cutler, A. (1981). The reliability of speech error data. *Linguistics*, 19, 561-582.

Cutler, A. (1988). The perfect speech error. In L.M. Hyman, & C.S. Li (Eds.), *Language, speech and mind: Studies in honor of Victoria A. Fromkin* (pp. 209-223). London: Croom Helm.

Dell, G. S., & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning & Verbal Behavior*, 20, 611-629.

Ferber, R. (1993). *Wie valide sind Versprechersammlungen?* Bern: Verlag Peter Lang.

Hollien, H., & Martin, C. A. (1996). Conducting research on the effects of intoxication on speech. *Forensic Linguistics: The International Journal of Speech, Language and the Law*, 3, 107-127.

Johnson, K., Pisoni D.B., & Bernacki, R.H. (1990). Do voice recordings reveal whether a person is intoxicated? A case study. *Phonetica*. 47, 215-237.

Klingholz, F., Penning, R., & Liebhardt, E. (1988). Recognition of low-level alcohol intoxication from the speech signal. *Journal of the Acoustical Society of America*, 84, 929-935.

Künzel, H. J., Braun, A., & Eysholdt, U. (1992). *Einfluss von Alcohol auf Sprache und Stimme*. Heidelberg: Kriminalistik-Verlag.

Lester, L., & Skousen R. (1974). The phonology of drunkenness. In A. Bruck, R. Fox, & M.W. LaGaly (Eds.), *Papers from the Parasession on Natural Phonology* (pp. 233-239). Chicago, IL: Chicago Linguistic Society.

Levinson, L.L. (1967). *Webster's Unafraid Dictionary*. New York: Collier Books.

Lodge, D. (1984). *Small World*. New York: Penguin Books.

Meringer, R., & Mayer, K. (1895/1978). *Versprechen und Verlesen: Eine psychologisch-linguistische Studie.* Amsterdam: John Benjamins.

Motley, M. T., & Baars, B. J. (1975). Encoding sensitivities to phonological markedness and transitional probabilities: Evidence from spoonerisms. *Human Communication Research*, 1, 353-361.

Mowrer, R., & Mackay, I. (1990). Phonological primitives: electromyographic speech error evidence. *Journal of the Acoustical Society of America*, 88, 1299-1312.

Nooteboom, S. G. (1967). Some regularities in phonemic speech errors. *Eindhoven: IPO Annual Progress Report*, 2, 65-70.

Nooteboom, S. G. (1973). The tongue slips into patterns. In V. A. Fromkin (Ed.), *Speech errors as linguistic evidence* (pp. 144-156). The Hague: Mouton.

Nooteboom, S. G. (1981). Speaking and unspeaking: Detection and correction of phonological and lexical errors in spontaneous speech. In V. A. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand* (pp. 87-95). New York: Academic Press.

Nooteboom, S. G. (2004). *Waar komen de letters van het alfabet vandaan?* Utrecht: Universiteit Utrecht.

Nooteboom, S. G. (in press). Listening to oneself: Monitoring speech production. In R. Hartsuiker, R. Bastiaanse, A. Postma, & F. Wijnen (Eds.), *Phonological encoding and monitoring in normal and pathological speech.* Hove, UK: Psychology Press.

Poulisse, N. (1999). *Slips of the Tongue: Speech Errors in First and Second Language Production*. Amsterdam: John Benjamins.

Shattuck, S. (1975). *Speech Errors and Sentence Production*. Unpublished doctoral dissertation, MIT.

Shattuck-Hufnagel, S., & Cutler, A. (1999). The prosody of speech error corrections revisited. In J.J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A.C. Bailey (Eds.), *Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, August 1-7, 1999* (Vol. 2, pp. 1483-1486).

Stemberger, J. P., Pisoni, D. B., & Hathaway, S. N. (1985). Effects of alcohol intoxication on phonological errors in normal speech. *Research on Speech Perception,* Progress report, 11 (pp. 95-404). Bloomington, IN: Speech Research Laboratory, Department of Psychology, Indiana University.

Trojan, F., & Kryspin-Exner, K. (1968). The decay of articulation under the influence of alcohol and paraldehyde. *Folia Phoniatrica*, 20, 217-238.

The tongue slips into (recently learned) patterns *

Gary S. Dell and Jill A. Warker
University of Illinois at Urbana-Champaign

Abstract

Speech errors reflect linguistic knowledge. For example, phonological errors follow language-wide phonotactic constraints such as the fact that [h] must be an onset in English. We review five experimental studies that demonstrate that errors adhere to artificial experiment-wide constraints (e.g. [f] must be an onset during this experiment) as well as language-wide constraints. These studies show that the language production system adapts to its recent experience with phonological patterns.

1 Introduction

Speech errors, or slips of the tongue, are an important source of data in psycholinguistics. It is not uncommon for textbooks in the field to devote entire chapters to them and their implications for theories of speech production (e.g. Carroll, 1999). The prominence of slips in the field, however, is a recent phenomenon. When psycholinguistics was in its infancy during the 1960's, speech errors were dismissed, both by linguists who were simply not interested in performance data and by experimental psychologists who saw their study as a suspect relic of Freudian theory.

Nooteboom was among the first to see the potential of speech errors as data for production theory. His article (Nooteboom, 1969, "The tongue slips into patterns") antedated the influential error analyses of Fromkin (1971) and Garrett (1975) and went beyond the classic study of Meringer & Mayer (1895) by linking errors to the information processing requirements of speaking as well as to the properties of spoken language. For example, he argued that one must consider the limited nature of short-term memory when explaining the distance that misplaced speech sounds move in errors (e.g. example 1).

(1) everything you hear → *everyhing you hear*

We (and just about everyone else nowadays) agree with Nooteboom that error patterns can be explained through cognitive and perceptual mechanisms acting on linguistic knowledge. Here, though, we put forth a more specific claim: Errors reflect recent experience with linguistic regularities as well as long-term linguistic knowledge. In other words, the tongue slips into recently learned patterns as well as those acquired through a lifetime of speaking.

2 The syllable-position and phonotactic-regularity effects

Consider error example (1) again. The [h] in *hear* moves to the onset of the nearby syllable *thing*, replacing the [ϴ]. Why does it move specifically to the *onset* of this syllable rather than some other spot? One explanation, termed the *syllable-position effect*, is that speech

* This research was supported by National Institutes of Health grant HD-44458 to Cynthia Fisher (principal investigator) and the first author.

sounds may have a tendency to preserve their position when they move in an error. The consonant [h] was an onset in *hear* and so it emerges as an onset in *hing*. An alternative explanation appeals to the fact that [h] is *always* an onset in English syllables. That is, the principles of English sound combinations, its *phonotactic* constraints, license [h] for onset position, but not for syllable-final (coda) position. More generally, we may explain the movement of [h] to an onset rather than a coda slot by hypothesizing a *phonotactic regularity effect* on errors: Errors do not create phonotactically illegal sequences.

The syllable-position and the phonotactic-regularity effects are both genuine influences on error patterns. The syllable-position effect has been observed in analyses of error collections in several languages (e.g. Garcia-Albea, del Viso, & Igoa, 1989; MacKay, 1970; Nooteboom, 1969; Stemberger, 1983). There is a clear tendency for onsets to move to onset positions and codas to move to coda position. Thus, *napkin* might slip to *kapkin* ([k]-onset moves to an onset position) but would be less likely to be mispronounced as *papkin* (movement of [p]-coda to onset position). The syllable-position effect is a bit difficult to disentangle from confounded influences of word position because word-initial sounds have a strong tendency to slip to other word-initial locations (Shattuck-Hufnagel, 1983). But when word-onset influences are removed from the data, a syllable-position effect remains. For example, in one analysis of English slips, 77% of English non-word-initial consonant movements retained their syllable positions (Vousden, Brown, & Harley, 2000).

The phonotactic regularity effect is even stronger than the syllable-position effect. The claim that phonological speech errors create only legal sound sequences was originally made by Meringer & Mayer (1895) and was called as the "first law" of speech errors by Wells (1951). Although this effect is often characterized as exceptionless, illegal slips do occur. For example, Stemberger (1983) collected 37 slips that violated the phonotactic-regularity effect, e.g. (2) below. These violations, however, were from a large collection of phonological slips, over 99% of which were phonotactically legal.

(2) first floor dorm → *first floor dlorm* (onset [dl] is illegal)

The standard view of the syllable-position and phonotactic-regularity effects is that they are separate influences on the form of speech errors. The syllable-position effect derives from a process of inserting consonants labeled as either onset or coda into labeled slots in a syllable structure or "frame". Each slot only takes appropriately labeled consonants (e.g. Shattuck-Hufnagel, 1979). So, the [k] in *napkin* is [k]-onset, not just [k]. If it slips to the first syllable of *napkin*, it must be inserted into that syllable frame's onset slot, creating the erroneous syllable [kæp]. The phonotactic-regularity effect is assumed to have a different mechanism. It reflects a set of phonotactic rules (e.g. *[h] must be an onset*) that are acquired early in life. If a potential slip violates a rule (e.g. an [l] is about to be inserted into *dorm* to create *dlorm*) some unspecified process either prevents the error or corrects it so that it is no longer illegal (Fromkin, 1971). Consequently, errors that violate the rules do not occur.

Our contention is that the syllable-position and phonotactic-regularity effects are not the result of distinct mechanisms. Rather they reflect closely related constraints on phonological sequences. The only difference is that the syllable-position effect arises from what we call *local constraints* and the phonotactic-regularity effect is a consequence of *language-wide constraints*. To illustrate, consider the target phrase, *king of hearts*. To produce this phrase, one must retrieve its constituent speech sounds and assign them to syllables and syllable

positions (Levelt, Roelofs, & Meyer, 1999). In particular, one must retrieve and assign [k] to the onset of the first syllable, [ɪ] to its vowel slot, [ŋ] to its coda slot, and so on. These processes are subject to constraints. Some of these constraints apply only in specific situations; for example, *[k] is onset | KING*. (The symbol "|" should be read as "in the context of"). Thus, this constraint would be active when *king* is spoken. Other constraints are more general, even language-wide; for example *[ŋ] is coda | all of English*. That is, English phonotactics require that [ŋ] be a coda. We assume that these constraints, local and general, are invoked during production and that they bias the assignment of speech sounds to syllable positions. The actual mechanisms of this bias are not important here. But it is worth noting that spreading activation through the network is a psychologically plausible way to represent constraints and their satisfaction, and that many models of production employ such mechanisms (e.g. Berg & Schade, 1992; Dell, 1986; Harley, 1984; Levelt et al., 1999; Stemberger, 1985).

An error that obeys the syllable-position effect, such as *king of karts* for *king of hearts*, is facilitated by the relevant local constraint, *[k] is onset | KING*. The erroneous [k] sticks close to *king* and it maintains it status as onset. However, as we mentioned before, the syllable-position effect is often violated (23% of the time in Vousden et al.'s, 2000, analysis of English word-internal errors). There are a couple of reasons for the frequency of these violations from our constraint perspective. First, the location of the error itself is outside of the relevant context; [k] should be an onset when saying *king*, not when saying *hearts*. Second, there are other local constraints that could work against the relevant constraint, constraints such as *[k] is coda | JACK* or *[k] is coda | HARK*. If any of these opposing constraints were to become active, they would counteract *[k] is onset.* Perhaps thinking about the king of hearts calls to mind other playing cards, such as the jack of hearts. Or possibly, planning the word *heart* activates the phonologically similar word *hark*. Speech-error studies (including Nooteboom's original paper) suggest that preparing to say a word leads to the activation of its semantic and phonological neighbors. All things considered, there are good reasons to expect local constraints such as *[k] is onset | KING* to have only limited influence and, hence, to expect violations of the syllable-position effect to be common.

Language-wide constraints such as *[h] is onset | all of English* or *[ŋ] is coda | all of English* are inherently more powerful than the local constraints. Because they are true for all words of the language, there are no opposing constraints. For example, there is no *[ŋ] is onse...* constraint to stimulate the error *king of ngearts*. Moreover, their influence is not limited to the context of a particular word. Their context is all of English. These properties of language-wide constraints help us understand why phonotactic regularity is a much stronger influence on errors than the syllable-position effect.

In summary, the syllable-position and phonotactic-regularity effects are both products of constraints on how speech sounds are positioned in syllables. The syllable-position effect results from errors adhering to a local constraint in the vicinity of a particular word. Adherence to language-wide phonotactics is due to more general constraints. In this way, the two error effects can be thought of as two ends of a continuum of breadth of constraint, with the syllable-position effect at the narrow end and the phonotactic regularity effect at the wide end.

3 The role of experience in speech errors: Where do the constraints come from?

Language-wide and local constraints must be learned. We are not born knowing that [k] is an onset in the word *king*, or that [ŋ] must always be a coda. Undoubtedly, much of this learning occurs early in life (e.g. Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993). We contend, though, that learning about the positions of sounds in syllables does not stop after childhood. On the contrary, language learning never stops. Language users continually adjust their knowledge of constraints based on their experience with phonological forms. This learning is implicit and serves to adapt the production and recognition systems to the current situation.

We have carried out several experiments demonstrating that recent experience can affect knowledge of constraints and thus change speech error patterns (Dell, Reed, Adams, & Meyer, 2000; Warker & Dell, in preparation). In particular, we tested the idea that experience can strengthen the syllable-position effect, turning it into a kind of phonotactic regularity effect. In these studies, English speakers recited syllables whose consonants were artificially restricted to particular syllable positions. For example, throughout the first experiment described below, [f] only occurred as an onset and [s] only occurred as a coda. Hence, [f] and [s] exhibited what we call *experiment-wide constraints*. Experiment-wide constraints (e.g. *[f] is onset | this experiment*) are a kind of middle ground between local constraints (e.g. *[k] is onset | KING*) and language-wide constraints (e.g. *[h] is onset | all of English*). The data of interest were the extent to which errors adhered to the constraints. We expected that errors would adhere to language-wide constraints nearly all of the time (the phonotactic-regularity effect). So, any slip of an [h], for example, should be to an onset position. We also expected that slips would adhere to local constraints (the syllable- position effect) around 70-80% of the time. Onset-[k]'s should tend to slip to onset positions, but there would be a fair number of exceptions. What about [f] and [s], the sounds subject to experiment-wide constraints? If learning occurs, slips of [f] and [s] should come to obey these constraints. For example, onset-[f]'s should move to onset, rather than coda, positions at a greater rate than consonants not subject to experiment-wide restrictions.

4 Learning experiment-wide constraints: First-order effects

4.1 Methods

Eight English-speaking participants recited four-syllable sequences in time with a metronome (e.g. example 3). They repeated 96 of these sequences four times in a row in an experimental session. Moreover, each participant did four separate sessions on separate days.

(3) *feng keg hem nes*

Each sequence contained one each of eight consonants, [h, ŋ, f, s, k, g, m, n], and the vowel was always [ɛ]. The consonants fell into three groups: language-restricted [h, ŋ], experiment-restricted [f, s], and unrestricted [k, g, m, n]. Within each sequence, [h] was always an onset and [ŋ] was always a coda, respecting English phonotactics. The consonants in the unrestricted group could be either onsets or codas. In one sequence, [k] might appear as an onset (as in 3 above), but in the next, it could be a coda. Experiment-restricted consonants

maintained their target syllable positions throughout the four sessions for the experiment. Half of the participants experienced [f] only as an onset and [s] only as a coda, and half experienced the reverse assignment. In total, 384 such sequences were prepared for each participant, divided into four sets of 96.

Each sequence was presented visually one at a time. The participant first repeated it slowly (1 syllable/sec) and then three times without pause at a faster rate (2.53 syllables/sec). Half of the participants were informed about the experiment-wide constraints at the beginning of each session ("when you see an 'f' it will be at the beginning of a syllable and when you see an 's' it will be at the end of a syllable"), and half were told nothing about these constraints. This manipulation was designed to test whether explicit knowledge of the experiment-wide constraints influences the extent to which errors adhere to them.

A second experiment used [k] and [g] as the experiment-restricted consonants; [f] and [s] were then put along with [m] and [n] in the unrestricted group. Another group of eight English speakers participated. In all other respects, this [k-g] version of the study was the same as the [f-s] version.

4.2 Results

The data of interest were slips in which consonants moved from one location to another. When such movement occurred did the consonants maintain their syllable positions? Table 1 shows the percentage of slips that did so, as a function of consonant group, for both the [f-s] and [k-g] experiments. First, consider slips of the unrestricted consonants. These maintained their syllable positions 68% of the time in the [f-s] experiment and 77% of the time in the [k-g] experiment. This clear syllable-position effect provided a baseline against which to compare effects for restricted consonants. Next, consider the language-restricted consonants. In both experiments, movements of [h] and [ŋ] kept their positions 100% of the time, demonstrating the robustness of the phonotactic-regularity effect. The key findings came from the experiment-restricted consonants. Slips of these maintained their positions at a greater rate than did the unrestricted consonants. In fact, these slips nearly always kept their positions (98% and 95% of the time), much like those of language-restricted consonants, which always stayed in position.

Table 1. Percentage of consonant movement errors that maintained syllable positions (from Dell et al., 2000)

	Unrestricted	Language-restricted	Experiment-restricted
Experiment 1	68% (*n*=1941)	100% (*n*= 640)	98% (*n*= 484)
	[k, g, m, n]	[h, ŋ]	[f, s]
Experiment 2	77% (*n*=1850)	100% (*n*=1016)	95% (*n*= 718)
	[f, s, m, n]	[h, ŋ]	[k, g]

Our interpretation of these results is that speakers learned something about the distribution of the experiment-restricted consonants and this knowledge affected their errors. The effect of this learning was strong in a number of respects. It developed on the very first day of testing

in both experiments; 98% for [f-s] and 93% for [k-g] of the experiment-restricted consonants kept their positions on first-day slips. The effect also appeared regardless of how the consonants were restricted (whether [f] was always an onset or whether [f] always was a coda). In fact, all 16 subjects exhibited high rates of position maintenance for the restricted consonants, including those who had not been explicitly informed about their distribution (see Dell et al., 2000, for detailed analysis of the data).

4.3 Discussion

We have suggested that speakers may be learning a new constraint, something like *[f] is onset | this experiment*. This constraint has a wider influence than local constraints and, in some ways, can be thought of as an experimentally induced analogue of language-wide phonotactic constraints. An alternative interpretation of the data is that previously existing constraints are being affected. One possibility is that particular local constraints are being strengthened. Perhaps every time that one says the syllable *fen*, the constraint *[f] is onset | FEN* grows stronger. If we then assume that stronger local constraints exert greater influence on errors in their vicinity, we can perhaps explain why slips of experiment-restricted consonants kept to their positions more than those of unrestricted consonants. Another possibility is that existing *general* constraints are subject to learning. For consonants that distribute freely across onsets and codas in the language such as [f], we hypothesized the existence of local constraints, only. But suppose that in addition to these local constraints, there exist general constraints of the form *[f] is onset* and *[f] is coda*. If the general constraints are affected by experience, our speakers who only get [f] in onset position may have their general [f]-onset constraint strengthened. The corresponding general constraint promoting [f] as a coda would not be strengthened or may even weaken. Thus, slips of [f] would show an enhanced tendency to stick to onset positions.

The data in Table 1 do not allow us to discriminate among these interpretations. What is required is a systematic investigation of the kinds of constraints that can and cannot be acquired and the conditions that promote their acquisition. As an initial step in this investigation, we have tested the learning of "second-order" experiment-wide constraints (Warker & Dell, in preparation). The constraint *[f] is onset | this experiment* is first-order; a particular consonant is associated with a particular syllable position. A second-order constraint is one in which the position that a consonant is associated with is conditioned on another property of the context. In the next section, we describe three experiments in which the experiment-wide constraint is a second-order vowel contingency. For example, in Experiment 3, some participants experienced syllables in which [f] must be an onset and [s] must be a coda, if the vowel is [ɪ], but [f] must be a coda and [s] must be an onset if the vowel is [æ]. This kind of experiment-wide constraint allows us to test whether the speech production system is capable of acquiring new arbitrary relations. Recall that one interpretation of the first-order findings is that existing general positional constraints (e.g. *[f] is onset*) are strengthened. If that is the only kind of learning that happens in these experiments, a second-order vowel-contingent constraint could not influence the error pattern. Experiment-restricted consonants such as [f] and [s] occur both as onsets and codas in this kind of second-order study and, hence, there is no overall correlation between

consonants and particular syllable positions. If there is a tendency for slips of the restricted consonants to stick to their positions beyond that of the unrestricted consonants, this tendency could not be attributed to strengthening an existing general constraint.

5 Learning experiment-wide constraints: Second-order effects

5.1 Methods

Eight English-speaking participants recited four-syllable sequences, such as those in Example (4), in time with a metronome on four separate days, producing 96 four-syllable sequences a day.

(4) *han fak mas gang*
sim ghin kif hing

As in Experiment 1 and 2, each sequence had 8 consonants, which fell into three groups: language-restricted [h, ŋ], experiment-restricted [f, s], and unrestricted [k, g, m, n]. The vowels in each sequence were either all [æ]'s, or all [ɪ]'s. Throughout the four sessions of the experiment, the experiment-restricted consonants adhered to the following constraints: *[f] is onset, [s] is coda* | *[æ]* and [s] *is onset, [f] is coda* | *[I]*. Half of the participants received the reverse constraints. Also, as before, half of the participants were informed of the constraints at the beginning of the experiment and half were not informed. The procedure was identical to that of Experiment 1 and 2.

An additional experiment used [k] and [g] as the experiment-restricted consonants, and a third one used [m] and [n]. These experiments were similar to the [f-s] version except that only four speakers participated in each.

5.2 Results

Errors where consonants moved from one location to another in a given sequence were analyzed to see if slips involving the experiment-restricted consonants would maintain their syllable position more often than slips involving the unrestricted consonants. In all three experiments, slips of [h] and [ŋ] followed the language-wide constraint and kept to their respective positions 100%. Unrestricted consonants maintained their position 77% in the [f-s] version, 74% in the [k-g] version, and 77% in the [m-n] version. However, the main findings concern errors involving the experiment-restricted consonants. On the first day of testing, slips of these consonants only maintained their syllable position about as often as slips of the unrestricted consonants: 86% in the [f-s] version, 72% in the [k-g] version, and 81% in the [m-n] version. But for the later testing sessions, day two through day four, experiment-restricted consonants kept their syllable position more often than the unrestricted consonants: 93% in the [f-s] version, 84 % in the [k-g] version, and 96% in the [m-n] version. Figure 1 shows the breakdown of percentages by day and by experiment.

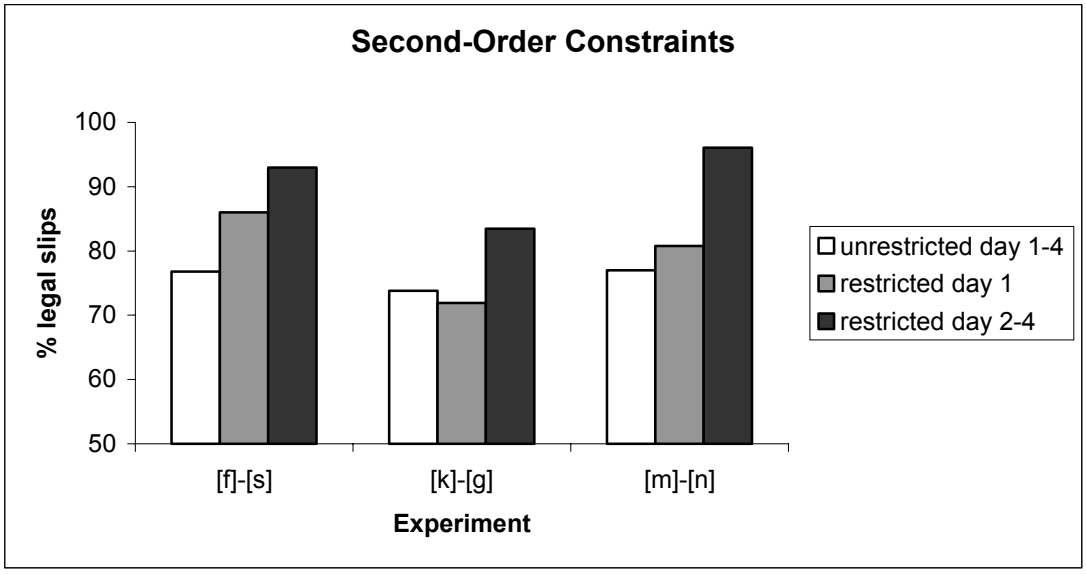


Figure 1. Percentage of experiment-restricted consonant movement errors that maintained syllable position.

The interpretation of these results is that participants learned something about the positions of the experiment-restricted consonants with regard to the identity of the vowel and, in turn, this learning influenced their speech errors. However, the pattern of learning differed from that found in Experiment 1 and 2. In the first-order constraint experiments, learning was found on the first day of testing. But in the second-order constraint experiments, it took until the second day of testing for participants' errors to adhere to the constraint. This pattern surfaced regardless of whether participants were explicitly informed of the constraints beforehand.

5.3 Discussion

Our results suggest that people can learn vowel-consonant dependencies and that these second-order constraints take longer to acquire than first-order constraints. This may occur because the second-order constraint is more complex and thus, harder to learn. As a result, people require more exposure to the constraint before being able to implicitly assimilate the rules. The learning of a second-order rule also indicates that people are not simply strengthening existing constraints, such as *[f] is onset*, since the experiment-restricted consonants occurred as both onsets and codas throughout the experiment. Rather, the results imply that the participants are adding a new constraint to their existing collection.

6 General Discussion

Our experiments show that speech errors respect the patterns present in an experiment as well as those that are true for the entire language. Participants acquired sensitivity to both first- and second-order constraints on the positions of consonants within syllables, although the second-order constraints were learned more slowly. We suggest that experience in producing syllables changes the speech production system, adapting it to its current circumstances.

The learned constraints have been described in terms of particular phonological categories—onsets, codas, and phonological segments. In a recent speech-error experiment, Goldrick (2004) demonstrated that features must be represented as well. Goldrick restricted a consonant to a particular position (e.g. *[f] is onset* | *this experiment*), but included a similar

unrestricted consonant (e.g. [v]) in the study. The presence of the unrestricted [v] weakened the tendency for [f] slips to be onsets. Furthermore, slips of [v] exhibited an increased tendency to be onsets, even though [v]'s were equally likely to be codas and onsets. These findings point to the acquisition of constraints about features such as *labial* or *fricative*, as well as constraints about entire segments.

Changes in speech-error patterns point to changes in production mechanisms. What about speech perception? Can the perceptual system rapidly acquire new constraints? Onishi, Chambers, and Fisher (2002) showed that it can. Their participants listened to CVC syllables that exhibited the same kinds of experiment-wide constraints that were tested in our production experiments. Then they presented participants with novel test syllables that followed the constraints and found that they could respond to these more quickly in an auditory naming task than novel syllables that violated the constraints. Both first-order and second-order vowel-contingent constraints could be learned.

Onishi et al. (2002) also identified a second-order constraint that could not be learned in their perceptual study. The constraint involved a speaker-voice contingency rather than a vowel contingency, for example, [b] is an onset if spoken by speaker A and [b] is a coda if spoken by speaker B. Onishi et al. hypothesized that constraints can only be rapidly learned if their elements are internal to the phonological system. Vowels and consonants are both internal to this system, and languages commonly exhibit phonotactic constraints in which consonants and vowels are dependent (e.g. in American English, the diphthong [iu] follows a restricted set of onsets, such as [k] or [m]). Speaker identity, though, would not normally be considered part of the phonological system. We are currently investigating this hypothesis in production by testing whether speech errors can become sensitive to a second-order constraint involving consonant position and a factor that can be argued to be extra-phonological, speech rate.

A final issue concerns language acquisition. We have been characterizing these experiments as learning experiments, inviting the conclusion that adult participants are learning artificial phonological constraints in much the same way that children acquire the phonotactics of their native language. We cannot assert this conclusion with confidence. We note, though, that 16-month olds learn first-order consonant-position constraints from listening to CVC syllables about as quickly as adults do (Chambers, Onishi, & Fisher, 2003). Studies with second-order constraints in children's perception are ongoing (K. Chambers, personal communication). Our speech-error results suggest that children will learn the second-order constraints, but perhaps more slowly than they do the first-order constraints.

7 Conclusions

Back in 1969, Nooteboom predicted that speech errors "may be of some use for the future construction of an explicit theory of language use" (p. 132), and proceeded to demonstrate just how this could be done. Nooteboom's understated prediction has been fully realized. Not only are speech errors profoundly important to theories of language production, but also, as we have seen, error patterns suggest useful avenues of exploration in perception and acquisition, in short, in all of psycholinguistics.

References

Berg, T., & Schade, U. (1992). The role of inhibition in a spreading-activation model of language production: I. The psycholinguistic perspective. *Journal of Psycholinguistic Research*, 22, 405-434.

Carroll, D.W. (1999). *Psychology of Language* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Chambers, K.E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, 87, B69-B77.

Dell, G.S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93, 283-321.

Dell, G.S., Reed, K.D., Adams, D.R., & Meyer, A.S. (2000). Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 6, 1355-1367.

Fromkin, V.S. (1971). The nonanomalous nature of anomalous utterances. *Language*, 47, 27-52.

Garcia-Albea, J.E., del Viso, S., & Igoa, J.M. (1989). Movement errors and levels of processing in sentence production. *Journal of Psycholinguistic Research*, 18, 145-161.

Garrett, M.F. (1975). The analysis of sentence production. In G.H. Bower (Ed.), *The psychology of learning and motivation* (pp. 133-177). New York: Academic Press.

Goldrick, M. (2004). Phonological features and phonotactic constraints in speech production. Manuscript.

Harley, T.A. (1984). A critique of top-down independent levels models of speech production: Evidence from non-plan-internal speech errors. *Cognitive Science*, 8, 191-219.

Jusczyk, P.W., Friederici, A.D., Wessels, J.M.I., Svenkerud, V.Y., & Jusczyk, A.M. (1993). Infants' sensitivity to sound patterns of native language words. *Journal of Memory and Language*, 32, 402-420.

Levelt, W.J.M., Roelofs, A., & Meyer, A.S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences,* 22, 1-75.

MacKay, D.G. (1970). Spoonerisms: The structure of errors in the serial order of speech. *Neuropsychologia*, 8, 323-350.

Meringer, R., & Mayer, K. (1895). *Verspechen und Verlesen*. Stuttgart: Goschensche.

Nooteboom, S.G. (1969). The tongue slips into patterns. In A.G. Sciarone, A.J. van Essen, & A.A. van Raad (Eds.), *Nomen: Leyden studies in linguistics and phonetics* (pp. 114-132). The Hague: Mouton.

Onishi, K. H., Chambers, K.E., & Fisher, C. (2002). Learning phonotactic constraints from brief auditory experience. *Cognition*, 83, B13-B23.

Stemberger, J.P. (1983). *Speech errors and theoretical phonology: A review*. Bloomington IN: Indiana University Linguistics Club.

Stemberger, J.P. (1985). An interactive activation model of language production. In A. Ellis (Ed.), *Progress in the psychology of language* (Vol 1, pp. 143-186). Hillsdale, NJ: Erlbaum.

Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial-ordering mechanism in sentence production. In W.E. Cooper & E.C.T. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett* (pp. 295-342). Hillsdale, NJ: Erlbaum.

Shattuck-Hufnagel, S. (1983). Sublexical units and suprasegmental structure in speech production planning. In P.F. MacNeilage (Ed.), *The production of speech* (pp. 109-136). New York: Springer-Verlag.

Vousden, J.I., Brown, G.D.A., & Harley, T.A. (2000). Serial control of phonology in speech production: A hierarchical model. *Cognitive Psychology*, 41, 101-175.

Warker, J.A., & Dell, G.S. (in preparation). Speech errors reflect newly learned phonotactic constraints.

Speech synthesis and electronic dictionaries

Arthur Dirksen

Fluency, Amsterdam

Abstract

Speech synthesis software can be used in electronic dictionaries to generate audible renditions of phonetic transcriptions, idioms and example sentences. This paper discusses advantages and disadvantages of using synthesized rather than recorded speech for these purposes, and examines one approach in some detail.

1 Introduction

Speech synthesis came out of the laboratory when sound cards, CD-ROMs, large hard disks and fast processors appeared on standard consumer PCs, sometime around 1995. These advances in technology also enabled electronic dictionaries and encyclopedias of (almost) unlimited size and scope. Since one of the purposes of a dictionary is to indicate how a word is pronounced, an electronic dictionary has an advantage over its paper equivalent in that it can do this using sound rather than symbols. The easiest and most straightforward way to add sound to a dictionary is to use recorded natural speech, and this is certainly the most common approach. However, speech synthesis provides a more general, flexible approach, which is well worth the additional investment.[1]

The most obvious advantage of using recorded human speech in an electronic dictionary is the naturalness and liveliness that can be obtained. But the newer speech synthesizers sound surprisingly natural, to the extent that it becomes difficult to distinguish synthesized from natural speech. Also, naturalness is but one aspect of speech. It is certainly the most important because it dominates other aspects: if speech sounds highly unnatural it will also be rather unintelligible, not just unpleasant to listen to. But once a sufficient level of naturalness is obtained, other aspects matter as well. In the context of an electronic dictionary, the user will want to know which sounds make up the pronunciation of a word, which syllable carries the main stress, how the distribution of stress influences the rhythm of the pronunciation (in stress-timed languages like English and Dutch), and not much else. If a speech synthesizer can do a convincing imitation of this, it is certainly good enough.

The advantages of using a speech synthesizer include the following:

- *scalability* With recorded speech, the amount of storage needed ultimately becomes problematic (and becomes problematic rather quickly on small devices). With synthetic speech, as each word is generated on the fly (using some kind of phonetic transcription as input), the same software can generate one million words just as easily as 10,000.

---

[1] On a more personal note, I came out of the laboratory when in 1996, after several years of academic research into linguistics, speech synthesis and text-to-speech, I started developing commercial text-to-speech software for Dutch, and founded Fluency (www.fluency.nl). Sieb Nooteboom helped in many ways with this "coming out," for which I am very grateful. Sieb was especially intrigued by the collaboration between Fluency and Van Dale Lexicografie on the topic of this paper, and always liked to mention that he had already discussed this possibility with Van Dale many years ago, when sound cards were still the size of a refrigerator. Which is why I think this paper is a suitable tribute to Sieb. In addition, I would like to thank Ludmila Menert, Josée Heemskerk, Johan Zuidema, Rik Schutz and Bram Wolthoorn for their various contributions to the work discussed here.

- *extensibility* New words can be added by adding a phonetic transcription (which will typically be done by the editors of the dictionary). There is no need to go back to the recording studio. Also, it is possible to include the pronunciation of not just head words, but all inflected forms of a word, or to offer audible pronunciation of idioms and example sentences. The latter is especially useful in a learner's dictionary, allowing the user to exercise the pronunciation of full sentences.
- *customizability* A speech synthesizer may allow the user to indicate preferences for a variety of aspects of the synthesized speech, such as the voice to be used (male, female), pitch and speech rate, or even speaking style (rather formal or somewhat informal). In addition, the dictionary publisher has the possibility to tune the pronunciation details to the intended audience of a given title (examples will be given in section 4).

A rather more subtle advantage, which need not be appreciated by users, is that a speech synthesizer establishes a formal and consistent relation between a phonetic transcription on the one hand, and an audio signal on the other. With recorded human speech, this can never be guaranteed, and the recording can only be regarded as an example of how a word is pronounced. With a speech synthesizer, to the extent that it operates correctly, what you see (in terms of phonetic symbols) is what you get (in terms of audio samples). For this reason, a dictionary publisher might wish to use speech synthesis to develop and evaluate a database of phonetic transcriptions.

This paper discusses collaborative research and development by Van Dale Lexicografie, a Dutch dictionary publisher, and Fluency, which produces text-to-speech software for Dutch. Section 2 briefly summarizes the development of a large pronunciation database for Dutch at Van Dale. Next, section 3 gives a rapid overview of the Fluency text-to-speech software, and its application in electronic dictionaries. Section 4 discusses the rule-based system that translates a phonetic transcription into a full specification for the MBROLA diphone synthesizer (Dutoit, 1997), which is used by the Fluency software for audio generation. These rules implement phonological and allophonic processes, and assign durations and pitch. Subtle (and admittedly non-modular) interactions between phonological and phonetic rules in our system define a formal relation between phonetic transcriptions and speech which is highly flexible and can easily be customized to suit a variety of applications. Section 5 indicates how example sentences can be synthesized correctly using annotations for pitch accent placement.

2 A database of phonetic transcriptions

The most economic way to produce dictionaries of varying size and scope for different audiences (children up to language professionals) is to derive them from a large database. Rather than updating each dictionary title for each new release, editors update the database directly, which saves not only time and money, but also improves consistency among the various titles. The construction and maintenance of such a database, however, is no small matter, especially if a dictionary publisher has a lot of legacy material. In the past decennium, Van Dale made a leap forward in this respect by constructing a database encoding the linguistic properties of over 1.5 million Dutch word forms. The properties encoded include spelling (as well as common misspellings), hyphenation and syllabification, part of speech, syntactic features, morphological structure, frequency of occurrence (in a large and frequently updated corpus of newspaper text), and, of course, pronunciation. In the construction of the database, language technology software and tricks-of-the-trade were heavily used, but everything was manually checked and edited. The database is used in-house to automatically enrich dictionaries, but it also serves as a source of data for language and

speech technology applications of Van Dale (Froon, den Hartog & Zuidema, 1999) and third parties.

The phonetic transcriptions in the database are not tied to a particular set of phonetic symbols such as IPA or SAMPA. Instead, they use a more abstract notation, from which actual transcriptions are derived: IPA for the professional dictionary, respelling for the children's dictionary, or whatever is needed for a particular application. Also, the master transcriptions include morphophonological information, such as secondary stresses, syllable and compound boundaries, and applications of phonological rules. Some examples:

- In a case such as *woordenboek* 'dictionary', a compound boundary is indicated in the transcription between *woorden* 'words' and *boek* 'book', and a secondary stress is indicated on the second part of the compound.
- In the transcription of the word *spinnenweb* 'spider web', it is indicated that the final consonant, which is pronounced [p] due to Final Devoicing, is underlying /b/ (which surfaces as [b] in the plural *spinnenwebben*).
- Although both *komen* 'come' and *gaan* 'go' end in /n/, the /n/ in the plural suffix *-en* may be deleted or reduced (unless /n/ is resyllabified as in *ze komen en gaan* 'they come and go'). Hence, in the transcription for *komen*, it is indicated that we are dealing with a deletable /n/.
- In *postbode* 'mail man', the /t/ may be deleted in casual speech, and this deletion would trigger Progressive Voice Assimilation: /s/ → [z]. Again, the transcription indicates the special nature of such consonant clusters.

All this extra information, which is not usually found in phonetic transcriptions for a dictionary, is a boon for applications and research in the area of language and speech technology. For example, in a lexicon for a morphophonological parser one will want to include underlying rather than surface forms. But also, the extra information allows the phonetic transcriptions to be tuned to a particular audience.

3 From text to speech

Conversion of text to speech is usually a three-stage process. In the initial stage, input text is assigned a phonetic transcription by looking up words in a lexicon, applying morphological or grapheme-to-phoneme rules for unknown words, as well as special rules for numbers, punctuation, dates, e-mail addresses, and so on. In the second stage, the phonetic transcription is modified by phonological rules (to the extent needed), each phoneme is assigned a contextually appropriate duration, and pitch targets are set to create a fitting intonation pattern. Finally, in the third stage the actual speech synthesis occurs, using one of several available methods (e.g., formant synthesis, diphone concatenation or unit selection).

Figure 1 gives a visual impression of how the Fluency text-to-speech software implements the three stages (for audio examples, go to www.fluency.nl). The second panel shows the phonetic transcription. The third panel displays the allophonic transcription (top), durations (bottom) and pitch targets (middle).

The software uses a lexicon of approximately 180,000 word forms. These were selected from the database described in section 2, using the lemma frequency of the word forms as the main criterion, with manual additions to optimize the lexicon for text-to-speech conversion. In order to cope with the fact that Dutch text is littered with English words and terminology, we also included some 15,000 English word forms, transcribed as well as one can expect using

the Dutch phoneme symbols. This lexicon covers arbitrary Dutch text rather well, and is augmented by compound analysis and grapheme-to-phoneme rules as backup mechanisms.

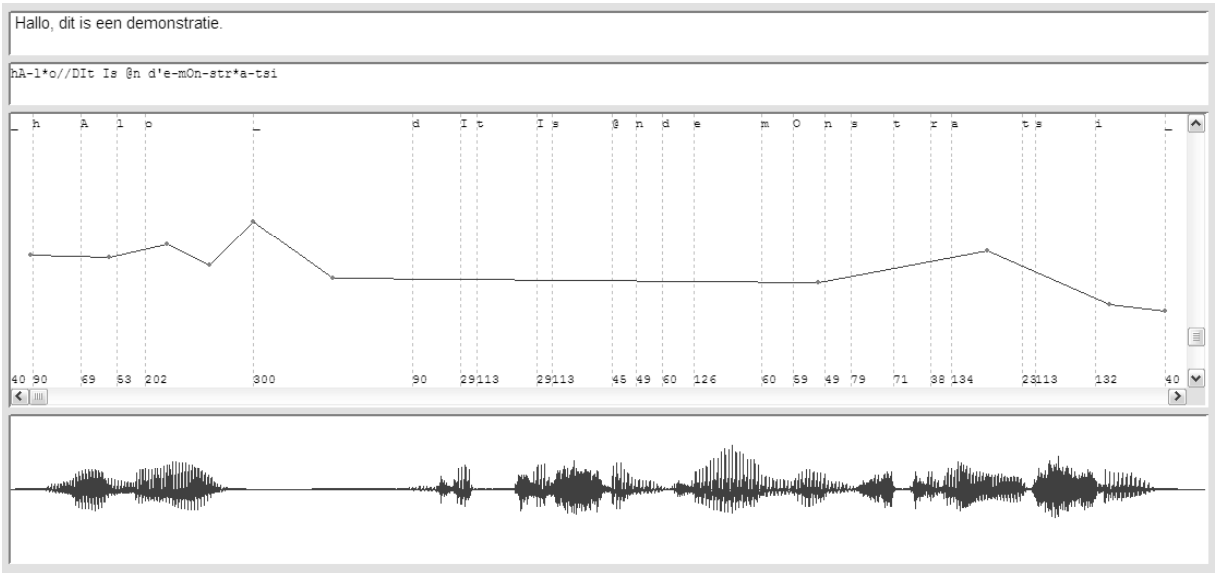


Figure 1. The Speech Editor application of Fluency TTS showing the intermediate stages in the generation of the sentence *Hallo, dit is een demonstratie* 'Hello, this is a demonstration'.

If the software is used to synthesize phonetic transcriptions of words in an electronic dictionary, the first stage is by-passed and the dictionary software provides the transcription. This is necessary, as the text-to-speech software cannot predict the pronunciation of ambiguous words such as *VOORkomen* 'happen' versus *voorKOmen* 'avoid' (small capitals indicate stress). It is also useful, as it allows us to define a more careful, emphatic pronunciation of the isolated dictionary words than is necessary for continuous speech. For example, the word *onAARdig* 'unkind' might be transcribed with a glottal stop before the second vowel in the dictionary, whereas for text-to-speech it might be preferable to have the /n/ in the first syllable resyllabify with the second syllable.

Regarding the third stage, a developer of text-to-speech software has a choice to either generate the audio from scratch (using articulatory or formant synthesis), or concatenate recorded speech samples (usually diphones). Excellent results can be obtained both ways, but with the latter approach it is easier to achieve the degree of naturalness required for commercial application.[2] The Fluency software uses MBROLA diphone concatenation, with diphones for both a male and a female voice, and approximately 2500 diphones for each voice. Included are diphones for French nasalized vowels (e.g. *ch<u>an</u><u>on</u>*), so the many French loans in the Dutch vocabulary can be synthesized correctly.

4 Phonological rules

In the second stage of the conversion of text to speech in the Fluency software, prosody rules are applied to the phonetic transcription to derive a specification for the synthesizer. In our approach, the term prosody applies not only to durations and intonation, but also covers

[2] With formant synthesis one can control almost every aspect of the synthesis process. This makes it rather suitable for phonetic experimentation (e.g. Dirksen & Coleman, 1997).

allophonic variation and phonological rules, which are integrated with the duration rules. The rules do not implement a full phonology of the Dutch language[3]: some of the effects of phonological rules are implicitly present in the diphones, other effects are assumed to be present in the phonetic transcriptions. In each case, we have looked at what was needed (for example, rules which operate across word boundaries need to be dealt with), and how we could obtain the best result in terms of speech output.

For example, there is no need to implement Homorganic Glide Insertion for words such as *du[w]o* 'duo' and *be[j]o* 'beo', or to require that these glides are included in the phonetic transcriptions. The reason is that these intervocalic glides are implicitly (and systematically) present in our diphones. On the other hand, we did need to implement Schwa Epenthesis for cases such as *mel[ə]k* 'milk' and *ker[ə]k* 'church', as the diphones for our male voice are somewhat infelicitous in this respect. However, we only needed a partial implementation of the rule, as many cases, for example *her[ə]fst* 'autumn', sound excellent without an epenthetic schwa[4]. In cases where the insertion does take place, we make the schwa very short, to account for the fact that it is part of a particular transition rather than a full segment.

We also found no necessitiy for Nasal Assimilation. Within morphemes, as in *ba<u>n</u>k* 'bank', we assume that a velar nasal is indicated in the phonetic transcription. Across morphemes, e.g. *i<u>n</u>komen* 'income' or *be<u>n</u> klaar* 'am ready', the diphone for [n-k] supplies a sufficient amount of natural assimilation (see also Jongenburger & Van Heuven, 1993). In fact, such natural assimilation is to be preferred in our view, as it provides a more subtle hint of what is going on than can be expressed in terms of phoneme symbols or phonological features.

In fact, when we cut our diphones, we tried to include these natural assimilations as much as possible. As a general rule, coarticulation between two consecutive sounds A and B is anticipatory in nature, that is, A will adapt to B much more than B to A. This can easily be seen in the spectrogram for a vowel which is in between consonants. The consonant-to-vowel transition shows very rapid formant transitions. In the vowel-to-consonant transition, on the other hand, formant transitions are slow, and they start early in the vowel. As a result, very early on in the vowel (as soon as the vowel targets are reached), very little evidence remains of the preceding consonant, but the effects of the following consonant can be seen (and heard) early on. So, although the common rule of thumb for cutting diphones is to cut somewhere in the middle of a sound, we prefer to cut very early in most sounds, so coarticulatory effects are taken into account as much as possible. Plosives are a notable exception, in that here we cut directly before the burst, i.e. very late rather than early in the time course of the segment, because the silent interval, whether voiced or not, is coarticulated with the preceding sound. As a result, in the sequence [V-n-k] the vowel-to-nasal diphone will signal nasality, and the nasal-to-plosive diphone will signal velarity, which is just right for these assimilations.

For the same reason, Regressive Voice Assimilation, which says that obstruents are voiced before a voiced plosive, can safely be left in the hands of the diphones. If the word *ee<u>t</u>baar* 'edible' is synthesized with a /t/ in the transcription, the result is more subtle than when a /d/ is transcribed, as the latter sounds overly casual (especially for a dictionary). A duration rule for intervocalic consonant clusters will make the [t] rather short, which also helps to indicate

[3] We have used Booij (1995) as our main reference for Dutch phonology.

[4] The diphone transition from a (rolling) [r] to [f] provides a subtle hint of a schwa.

an assimilated voiced plosive, as voiced obstruents are much shorter in duration than their voiceless counterparts. Finally, it is interesting to note that in clusters of two plosives there will generally be one (assimilated) burst rather than two identifiable segments.

However, perhaps as a result of our strategy for cutting diphones, we do need to implement Progressive Voice Assimilation, which says that a fricative is devoiced after a voiceless obstruent, both within and across words. And we do need to take care of interactions between the voice assimilation rules and Degemination.

The Degemination rule says that two consecutive identical consonants merge into one. This happens at morpheme boundaries, e.g. *nachttrein* 'night train', but also across words, e.g. *hij kan niks* 'he can do nothing'. Rather than merely deleting one of the two, we make the remaining one almost twice as long, so it still signals it twin brother. Again, this sounds more subtle and careful than a single consonant, and it provides an interesting example of the interplay between phonology and phonetics.

In derivational phonology, the output of the voice assimilation rules may feed Degemination, as in *afval* 'garbage' and *klapband* 'flat tire'. Although the [f-v] and [p-b] diphones do not sound too bad in these cases, we did incorporate these assimilations in our rules. However, as our rule system is not derivational, we had to adapt our Degemination rule with special cases for these assimilated obstruent clusters.

Our rules are also sensitive to the extra information encoded in the Van Dale transcriptions (section 2). The secondary stresses and compound boundaries are a useful guide to duration rules that define the rhythm of a word or sentence. Also, the special nature of the plural suffix /n/ in *komen* 'come' is dealt with by our rules. If it resyllabifies, as in *ze komen en gaan* 'they come and go', it is treated as any other onset /n/. If it does not, it is not deleted (as a straightforward phonological rule would demand), but given very short duration (less than 15 ms.). Such a short [n] is not perceived as such, but presents itself as a hint of nasality on the preceding schwa, which sounds more refined than a simple deletion, or, for that matter, a full [n], which sounds overarticulated or even wrong.

The fact that we do not always need to explicitly define a phonological rule in our software, does not mean that we cannot do so, should this be needed. All rules in our system can be parameterized, and these parameters can be changed at run-time. In an early version of the rule set, the explicit application of phonological rules could be selected by the end-user. The idea was that explicit assimilations and deletions define a somewhat casual speaker, whereas implicit, natural assimilations, and reductions instead of full deletions, define a more formal, refined speaker, which the end user might expect from a talking dictionary. This approach allowed us, for example, to synthesize three versions of the word *postbode* 'mail man' (see section 2):

- a version with a rather short [t] (about 40 ms, a single intervocalic [t] receives a duration of 90 ms).
- a version without the /t/, but with natural assimilation of [s] to [b].
- a version without /t/, and with explicit assimilation.

We found that we could amuse an audience of linguists and phoneticians with such demonstrations, but for most end users the differences were just a bit too subtle. So, in the end, we decided to remove the clutter of if-then-else's and select the more refined speaking style as the only option.

However, one optional rule remained, as users seemed to have rather pronounced preferences. It concerns the allophony of /r/, which can be a rolling [r] or a more vowel-like [R], which is not unlike /r/ in American English. In our analysis of Dutch, the [r] is selected if it appears directly before a vowel (and in cases of Epenthetic Schwa like *ker[əɹ]k* 'church' or *her[əɹ]fst* 'autumn'), whereas the [R] is typical of the coda position. But many users prefer a rolling [r] in all positions, so we offer this as an option.

5 Sentences

It is one thing to hear a word spoken in isolation, quite another to hear the pronunciation of that word in a sentence. A dictionary usually gives one or more examples of the usage of a word, and of how the word may be used in idiomatic expressions. As an extra service to the user, an electronic dictionary may offer audible pronunciations of these sentences as well. With the use of text-to-speech software, this can easily be achieved. It is even possible, and especially useful for non-native users, to highlight each word as it is being spoken (a standard feature of modern text-to-speech software). In this case, the input to the speech synthesis software is not a phonetic transcription but plain text.

However, if the text-to-speech software has to decide on the pronunciation of a sentence all by itself, it cannot be guaranteed that it will always be correct. Ambiguous words present problems that cannot always be solved by the text-to-speech software, even if it uses contextual analysis. Also, the distribution of pitch accents in a sentence is a thorny problem which involves not only syntactic structure, but also the (presumed) discourse context (Quené & Kager, 1993; Dirksen & Quené, 1993; Hoekstra, 2004).

The most straightforward solution to these problems is to (have the editors of the dictionary) annotate sentences where necessary. Ambiguous words can be solved by giving the exceptions a special code and providing additional lexicon entries with the correct pronunciation of these forms. Some examples:

hij kon niet voorkomen dat het fout ging 'he could not prevent it to go wrong'
het kan voorkomen\1 dat de deur gesloten is 'it can happen that the door is closed'

het regent de hele dag 'it rains the whole day'
de regent\1 is afwezig 'the governor is absent'

In the first example, the code '\1' is used to refer to an extra lexicon entry with the pronunciation for the word *VOORkomen* 'happen', with stress on the first syllable rather than the second (the speech synthesis software selects *voorKOmen* 'prevent' as the default). In the second example, the software knows the more frequent *regent* /rˈe-ɣənt/ 'rains', but not the rather arcane *regent* /rə-ɣˈɛnt/ 'governor', so the latter is added to the lexicon with a special code.

Pitch accent patterns can be indicated to the Fluency text-to-speech software by using the following tags: '\+' to add a pitch accent to a word, '\–' to remove a pitch accent, and '\<' to indicate a rhythmic stress reversal. Some examples (accented syllables are capitalized):

de tele**VI**sie is de **HE**le \–avond niet \+**AAN** geweest 'the TV has not been on all evening'
ik heb het \<**HE**lemaal met hem \+ge**HAD** 'I've really had it with him'

In the first example, *aan* 'on', which is in the synthesizer's list of unstressed function words, is used as a particle and should receive the main focus of the sentence. A pitch accent on *avond* 'evening' is not wrong, but the sentence is rhythmically much better without it. In the second example, the normal stress pattern of *helemaal* 'really' is with the main stress on the final syllable, and a secondary stress on the first. The '\<' tag reverses this pattern to improve the rhythm of the sentence.

Using these tags, a collection of examples and idiomatic expressions can easily be optimized so they are spoken with adequate prosody by our text-to-speech software. Example sentences are usually short and simple, so most will not need any modification at all. The extra codes can easily be hidden by the dictionary software.

6 Conclusion

Text-to-speech software can be used in electronic dictionaries to demonstrate to the user the pronunciation of words and example sentences. We have seen that some amount of experimentation and fine-tuning is both useful and necessary. Phonetic transcriptions embedded in the dictionary software should not attempt to provide pronunciation details that can be added with more subtlety by the speech synthesis software (section 4). On the other hand, the text-to-speech software sometimes needs a helping hand with the pronunciation of sentences (section 5).

The close collaboration between Fluency and Van Dale Lexicografie has resulted in high-quality text-to-speech software, that is used as a stand-alone product, but also as the pronunciation module in several of Van Dale's electronic dictionaries.

References

Booij, G. (1995). *The Phonology of Dutch*. Oxford: Oxford University Press.

Dirksen, A., & Quené, H. (1993). Prosodic Analysis: The Next Generation. In V. J. van Heuven, & L. C. W. Pols (Eds.), *Analysis and Synthesis of Speech* (pp. 131-144). Berlin: Mouton de Gruyter.

Dirksen, A., & Coleman, J. S. (1997). All-Prosodic Speech Synthesis. In J. P. H. van Santen, R. W. Sproat, J. P. Olive & J. Hirschberg (Eds.), *Progress in Speech Synthesis* (pp. 91-108). New York: Springer.

Dutoit, T. (1997). *An Introduction to Text-To-Speech Synthesis*. Dordrecht: Kluwer.

Froon, J., Den Hartog, J., & Zuidema, J. (1999). Beter verbeteren: slimme spellingchecker. *Natuur en Techniek*, *67* (12), 7-15.

Jongenburger, W., & Van Heuven, V. J. (1993). Sandhi Processes in Natural and Synthetic Speech. In V. J. van Heuven, & L. C. W. Pols (Eds.), *Analysis and Synthesis of Speech* (pp. 261-276). Berlin: Mouton de Gruyter.

Hoekstra, H. (2004). (De-)accenting and discourse structure. In this volume.

Quené, H., & Kager, R. (1993). Prosodic Sentence Analysis without Parsing. In V. J. van Heuven, & L. C. W. Pols (Eds.), *Analysis and Synthesis of Speech* (pp. 115-130). Berlin: Mouton de Gruyter.

Perceived vowel duration

Carlos Gussenhoven

Radboud University Nijmegen

1 Introduction

Two subjects that lie close to Sieb Nooteboom's professional heart are speech perception and vowel duration, and I am therefore pleased that I can combine these two interests in my contribution to this volume. I am concerned with the difference between *acoustic duration* and *perceived duration*, which concepts differ in a similar way to fundamental frequency and pitch. I will claim that vowel height affects perceived duration, in the sense that higher vowels sound longer than lower vowels when acoustic durations are equal. That is, vowel height and perceived duration are positively correlated. In section 2, I present the results of a perception experiment with Dutch listeners which shows this correlation. In sections 3 and 4 I deal with the two main questions that this finding raises. The first concerns the reason for this correlation. I will argue that it is to be sought in a mechanism of compensatory listening, and will cite other cases in the literature that have been given parallel explanations. The second question is whether the correlation is of any significance for the phonetics or phonology of languages. Here, I will argue that it solves a two cases of vowel raising, one phonetic and one phonological, in English and Limburgian Dutch, respectively.

2 Perceived duration of high and mid vowels

The research reported in this section was carried out in collaboration with Wilske Driessen, who wrote her MA thesis on this topic (Driessen 2004). A female speaker of Dutch recorded a number of isolated pronunciations of the vowels [i, y, u, ɛ, œ, ɪ, ɔ] as in *wie* 'who', *nu* 'now', *koe* 'cow', *bed* 'bed', *oeuvre* 'works', *pit* 'kernel', *bot* 'blunt', respectively, on digital audiotape, pronouncing them with a weakly falling intonation. Good tokens of these six vowels were stored at a 16 kHz sampling rate and trimmed so as to end up with a complete number of periods that were the closest to 180 ms in duration. With the help of the option for the manipulation of the fundamental frequency in the Praat package (Boersma & Weenink, 2002), they were provided with contours starting at 160 Hz and ending at 110 Hz, with a 40 ms plateau of 220 Hz beginning after 40 ms. Each of these standardized speech files then served as the basis for six further manipulated vowels with 15 ms increments in duration. That is, we ended up with seven durational versions of each vowel: 180, 195, 210, 225, 240, 255, and 270 ms. The same treatment was applied to tokens of four further vowels, [a] as in *na* 'after' and the diphthongs [ɛi, œy, ʌu], as in *ei* 'egg', *ui* 'onion', *kou* 'cold'. The diphthongs served as fillers for the purposes of this report, but the result of [a] will be reported here. The set of 10×7 stimuli was randomised three times and divided into blocks of ten. Thirty-four Dutch listeners rated each stimulus for duration on a 7-point scale, with the shortest duration appearing on the left of the scale. Each block was preceded by an anchor stimulus of 225 ms with a schwa-like vowel quality, which corresponding to a scale on the answer sheet in which the fourth scale category had been crossed off. Listeners were told that this stimulus represented the mid-point on the scale.

An analysis of variance with Duration (7 levels) and Vowel Height (2 levels) showed that acoustic duration and vowel height significantly affected the perceived duration.

Figure 1 shows that the effect of acoustic duration is consistently present in the case of all six vowels, while [a], which had been excluded from the analysis, shows the same positive correlation. More interesting in the context of this contribution is the finding that the high vowels are consistently rated longer than the equivalent mid vowels [$F(1,32)=18.11$, $p<.01$]. Interestingly, mid back [ɔ] is rated as longer than low back [a]. The results for the high and mid vowels have meanwhile been replicated in a second experiment with different stimuli and a different group of Dutch listeners.

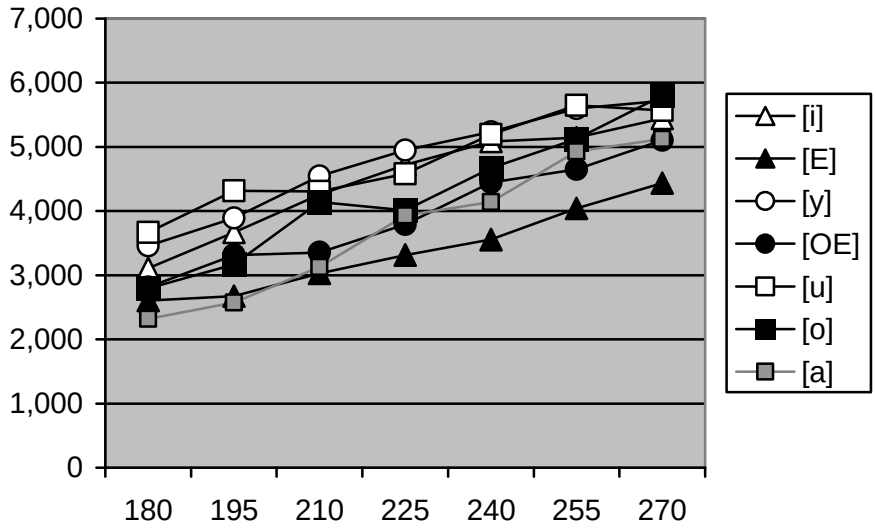


Figure 1. Perceived durations (on a 7-point scale) by Dutch hearers of seven vowel stimuli, each with seven acoustic durations (in ms) .

3 The explanation of the correlation between vowel height and perceived duration

Higher vowels are shorter than lower vowels. This is a universal tendency which has been explained on the basis of the distance between the roof of the mouth and the articulatory excursion of the tongue(-cum-jaw) made for the vowel: the greater this distance, the longer the vowel. In Dutch, this tendency has become phonologized. Originally long /i, y, u/, as in *wie* 'who', *nu* 'now', *koe* 'cow', have become short /i, y, u/, thus merging their quantity with /ɪ,ʏ/, as in *pit* 'kernel' and *put* 'sink, well' (Moulton, 1962; Nooteboom, 1972; Gussenhoven, 2004). It is suggested that, paradoxically, the negative correlation between vowel height and acoustic duration explains why vowel height and *perceived* duration are *positively* correlated. The hearer knows that low vowels require more time and thus be inherently longer than high vowels. When assessing the duration of a vowel he will therefore subtract this inherent portion in the duration before constructing the perceived duration. Putting it differently, high vowels do not, but low vowels do include a component in the *articulatory duration* which is obtained as an unintended bonus for producing the vowel *quality* in question, and by way of compensation, the hearer reduces the acoustic duration accordingly.

The explanation readily generalizes to other cases of 'compensatory listening'. First, Pierrehumbert (1979) found that accent-lending fundamental frequency peaks in English have more prominence if they come later in the utterance. The effect was attributed to the existence of a descending, abstract reference line marking equal pitch, which mimicked the declination found in production studies. By employing this reference line instead of the fundamental frequency scale when measuring the pitch of the peak, the hearer thus compensated for the declination in production, by bringing late low peaks back up to the level it would have had if there had been no declination. A second case is Silverman (1984), who demonstrated that the pitch of the same fundamental frequency peak was lower when combined with the British English vowel /iː/ than when it combined with the vowel /ɑː/. He did this by having hearers judge the pitch of the accented words in utterances like *They only FAST before FEASTing* and *They only FEAST before FASTing*, in which the only difference is the spectral composition of the accented words, and demonstrating that *feast-* had lower pitch than *fast-* when fundamental frequency contours were identical.

Both the declination effect and the intrinsic pitch effect occur because the hearer subtracts the effect of an articulatory advantage from the acoustic value. In the first case, the high subglottal pressure affords the speaker an advantage in the fundamental frequency domain, since higher subglottal pressure will lead to increased rates of vocal fold vibration. In the second case, the articulation of the high vowel causes an upward pull on the tongue root, the hyoid and the thyroid, causing some tensing of the vocal folds, which as a result vibrate a little faster (see Maddieson, 1997). Figure 2 presents this idea of subtraction of an articulatory advantage in a graphic form. The shaded area represents the boost in fundamental frequency due to high subglottal pressure in the case of the declination effect, the boost in fundamental frequency due to high tongue position in the case of the 'intrinsic pitch' affect, and lastly, the boost in vowel duration due to low tongue(-cum-jaw) position in the case of the correlation which is the subject of this contribution.

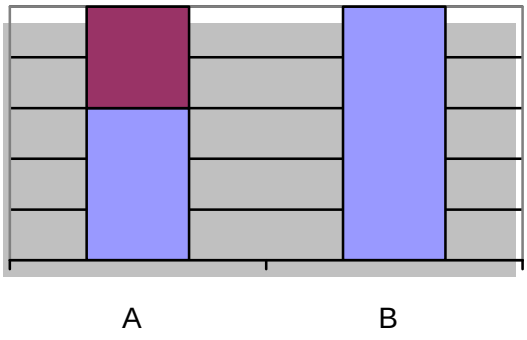


Figure 2. Compensatory listening explained as the hearer's subtraction of an articulatory advantage enjoyed by the speaker. Sounds A and B have identical acoustic values, but the hearer reduces A's value by subtracting the boost of the fundamental frequency or duration, which he assumes occurred as a by-product of the articulation.

4 The correlation between vowel height and perceived duration in phonetics and phonology

It is unlikely that the main finding reported in Figure 2 is a psycho-acoustic effect, unrelated to speech. This is because its explanation generalizes to other correlations that have been reported. Clearly, hearers bring detailed phonetic knowledge of speech production to bear on

the interpretation of the acoustic data. Not only are speakers in control of their speech production, hearers too know how speakers go about producing speech. By itself, of course, this merely shows that the correlation between vowel height and perceived duration is known to the participants in speech communication, but it has not been shown that it has any significance other than the significance which a known underwater rock at the entrance of a harbour has for a wary ship captain. The purpose of this section is to suggest that there is more to it. Section 4.1 summarizes certain regularities in pre-obstruent vowels in English and and an explanatory account of them by Moreton (2004), and introduces similar facts that are used to enhance a tonal contrast in Limburgian Dutch. Section 4.2 proposes an alternative explanation in which the observed correlation between vowel height and perceived duration plays a crucial role.

4.1 Enhancing the laryngeal contrast in English syllable codas

In a recent article, Moreton (2004) summarizes a number of research findings showing that in English low vowels are opener before [–voice] obstruents than before [+voice] obstruents. That is, *cat, boss* have a higher F_1 than *cad, Bozz*. Moreton assumes that this is an effect of hyperarticulation, and bases this assumption on the even more widely reported finding that the second element of English closing diphthongs is higher before [-voice] obstruents than before [+voice] obstruents. The latter effect has been phonologized in the case of /aɪ/ in Canadian English, where *writer*, for instance, has a categorically different vowel from *rider*, and where the distribution of these vowels is not entirely predictable (Wells, 1981:495). Although I will argue that Moreton's conclusion is incorrect, the fact that low vowels are lower and high second elements of diphthongs are higher before [–voice] indeed suggests hyperarticulation of the vowel. Moreton reports new measurements showing that the first elements of diphthongs are also higher before [–voice] consonants, though not to the same extent as second elements, and additionally demonstrates that F_1 and F_2 in the second elements of diphthongs may influence the perception of [voice] in the following consonant. After presenting and convincingly rejecting a number of explanations that have been proposed in the literature for these findings (not all of which are based on the assumption that the difference results from hyperarticulation), he tentatively advances the explanation that the vowels are hyperarticulated because the voiceless consonants are hyperarticulated, as if by a leakage of articulatory effort preceding the consonant in question. Since the greatest difference with the pre-voiced context is achieved in the latter half of the vowel, also in the case of the low monophthongs, this explanation would at first sight seem to be on the right track.

There are two problems with Moreton's suggestion of his *Spread-of-Facilitation* hypothesis. The first concerns the status of hyperarticulation as way of enhancing contrasts. It is not clear why hyperarticulation should apply to only one of two members of an opposition, rather than to both. Related to this is the question why, if one is to be chosen, this should be the voiceless member. Moreton suggests that the answer to this question is that the closing gesture of the voiceless obstruent is what is specifically used to enhance the contrast, rather than the opening gesture, and that therefore the hyperarticulation is only found before the voiceless consonant. This explanation is not unreasonable; yet, as far as I know it introduces a new conception of contrast enhancement, *viz*. that of selectively hyperarticulating one member of an opposition instead of both, and thus using the hyperarticulation itself as creating a contrast with the non-enhanced member of the opposition. The second problem with Moreton's conjecture is that there is no evidence that high vowels are higher before [-voice] obstruents than before [+voice] obstruents. As he observes, this is one of the predictions that his theory

makes. Like Wolff (1978) and Van Summers (1987), Moreton investigated the behaviour of *low* vowels before the voicing contrast.

While the first arguments against the *Spread-of-Facilitation* hypothesis might be countered with an observation that apparently this is the way things go, and the second by observing that there the jury is still out as long as the data are not available, there is a third problem which cannot as easily be dismissed. Vowel quality differences similar to those that have been observed before the laryngeal contrast in English have been found in syllables with a tone contrast in Dutch (Limburgian) dialects spoken in the northeast of Belgium and the southeast of the Netherlands (Gussenhoven, 2004). The tone contrast is known as Accent 1 vs Accent 2, or in the Dutch dialectological literature as *stoottoon* or *valtoon* ('pushing tone/falling tone') vs *sleeptoon* ('dragging tone'), respectively. Phonologically, the contrast has been described as the absence of a lexical tone vs the presence of H, respectively (Gussenhoven & Aarts, 1999). Phonetic realizations vary across the dialects, but frequently reported differences are that Accent 1 tends to be shorter, to have wider F_0 movements, and, less systematically, to have a firm amplitude decrease towards the end of the sonorant segments in the rhyme. For instance, in the dialect of Mechelen-aan-de-Maas, mid vowels split into an opener and closer vowel in syllables with Accent 1 and Accent 2, respectively, as shown in (1) (Verstegen, 1996).

(1)

Accent 1		*Accent 2*	
ɣɛɛl	'yellow-ATTR'	ɣeel	'yellow-PRED'
wɛɛx	'road-PL'	weex	'road-SG'
ɣɔɔn	'go-1SG,PRES'	ɣoon	'go-1PL,PRES'
nɔɔl	'needle-SG'	noolə	'needle-PL'

And in the dialect of Maastricht, the diphthongs /ɛi, œy, ɔu/ have markedly different allophones depending on whether they cooccur with Accent 1, as in (2a), or Accent 2, as in (2b) (Gussenhoven & Aarts, 1999). When combining with Accent 1, the diphthong's end point is very close, while in syllables with Accent 2 the end point is only weakly approximated, so much so that these vowels may lose their diphthongal character. The difference is non-discrete, and native speakers regard the allophones as the same vowel in each of the three cases.

(2) a. Accent 1: / bɛi/ 'bee', /lœy/ 'people', /dɔuf / 'pigeon'; [bɛj, lœj, dɔwf]

b. Accent 2: /bɛi/ 'near', /lœy/ 'lazy', /dɔuf / 'deaf'; [ˉbɛː$^{(i)}$, ˉlœː$^{(y)}$, ˉdɔː$^{(u)}$f]

Strikingly, the closer second elements of the diphthongs and the opener realizations of the monophthongs go hand in hand, as in the cases reviewed by Moreton. However, equally strikingly, the contrast that is to be enhanced by these quality differences is not a laryngeal contrast, but a tonal one.

4.2 An explanation for English and Limburgian Dutch

Before continuing to speculate on the explanation for the quality differences, we need to consider the question what the English coda voicing contrast and the Limburg tone contrast have in common. The answer is that both contrasts are enhanced, in Stevens & Keyser's (1998) sense of aided by some non-primary phonetic parameter, by a durational difference. I therefore propose that a *Duration Enhancement* hypothesis should take the place of Moreton's *Spread-of-Facilitation* hypothesis. Specifically, [–voice] obstruents are preceded by shorter sonorant portions in the rhyme than [+voice] obstruents, while also the sonorant portions in rhymes with Accent 1 are shorter than those with Accent 2. Since Limburgian

Dutch Accent 1 patterns with English [–voice] codas in this respect, it is difficult to escape the impression that the vowel quality differences are there to enhance the duration differences.

If this is so, the next question is how the features associated with the shorter vowels, closer second elements of diphthongs and opener low vowels, can contribute to the impression of shorter vowel duration. It would appear that they do so in different ways. The high second element is there to change the second element of the diphthong from a vowel into a consonant, and in that way reduce the perceived vowel duration. That is, [ɛj] sounds as if it has a shorter vowel than [ɛi], which in turn may well sound sorter than [ɛe]. In this interpretation, Canadian Raising is a trick to transfer part of the vowel to a consonantal percept, thereby reducing the perceived vowel duration.

Second, the more open vowels before [–voice] obstruents and in syllables with Accent 2 are likewise there to suggest longer vowel durations, in this case by exploiting the compensatory listening effect reported in section 2. That is, vowel lowering and vowel raising are ways of making vowels sound shorter and longer, respectively. The correlation between vowel height and perceived duration therefore appears to have a crucial role in the phonetics and the phonologies of languages.

5 Conclusion

The hypothesis that vowel lowering and closer diphthong off glides are ways of making vowels sound shorter has a number of advantages over Moreton's *Spread-of-Facilitation* hypothesis. First, it is no longer the case that *one* of the two terms in the phonological contrast to be enhanced is selected for having the privilege of a more canonical articulation bestowed upon it. In the *Duration Enhancement* hypothesis the two terms receive in principle equal treatment. Indeed, the number of phonological repercussions of the durational enhancement in the Limburgian Durch dialects is quite varied, and Accent 1 may just as easily be targeted for a change as Accent 2. Second, the *Duration Enhancement* hypothesis is capable of explaining the vowel quality adjustments in both as an enhancement of the English laryngeal contrast and of the Limburgian Dutch tone contrast.

Importantly, the choice between the two theories can be decided by a simple experiment. First, if the *Spread-of-Facilitation* hypothesis is correct, British English words like *bit*, *niece*, *look*, *belief* should have higher vowels than *bid*, *knees*, *Luke*, *believe*, but if the *Duration Enhancement* hypothesis is correct, the former should have lower vowels than the latter. Second, if the *Spread-of-Facilitation* hypothesis is correct, British English words with mid vowels like *bet*, *Bert*, *boss* should have more peripheral vowels than *bed*, *bird*, *Bozz*, but if the *Duration Enhancement* hypothesis is correct, the former should, again, have lower vowels than the latter. Research that addresses these questions is in progress.

References

Boersma, P. & D. Weenink (2002). Praat: Doing Phonetics by Computer. Computer program, available at http://www.praat.org .

Driessen, W. (2002). Compensatory Listening: The Effect of Vowel Quality on the Perception of Vowel Duration. MA thesis, Department of English, University of Nijmegen.

Gussenhoven, C. & W. Driessen (2004). Explaining two correlations between vowel quality and tone: The duration connection. Paper presented at the ISCA workshop on *Prosody 2004*.

Gussenhoven, C. & F. Aarts (1999). The dialect of Maastricht. *Journal of the International Phonetic Association, 29*, 55-66.

Maddieson, I. (1997). Phonetic universals. In W.J. Hardcastle & J. Laver (Eds.) *The Handbook of Phonetic Sciences* (pp. 619-639). Oxford: Blackwell.

Moreton, E. (2004). Realization of English postvocalic [voice] contrast in F_1 and F_2. *Phonetica, 32*, 1-33.
Moulton, W. (1962). The vowels of Dutch: Phonetic and distributional classes. *Lingua, 11*, 294-312.
Nooteboom, S.G. (1972). *Production and Perception of Vowel Duration: A Study of Durational Properties of Vowels in Dutch*. PhD dissertation, Rijksuniversiteit Utrecht.
Stevens, K. & J. Keyser (1989). Primay features and their enhancement in consonants. *Language, 65*, 81-106.
Van Summers, W. (1987). Effects of stress and final consonant voicing on vowel production: Articulatory and acoustic analysis. *Journal of the Acoustical Society of America, 82*, 847-863.
Verstegen, V., 1996. Bijdrage tot de tonologie van Oostlimburgse dialecten. In H. van de Wijngaard (Ed.) *Een eeuw Limburgse dialectologie* (pp. 229-234). Hasselt/Maastricht: VLDN/Vereniging Veldeke Limburg.
Wells, J.C. (1981) *Accents of English*. Three volumes. London: Longman.
Wolf, C.G. (1978). Voicing cues in English final stops. *Journal of Phonetics, 6*, 299-309.

The perceptual development of a British-English phoneme contrast in Dutch adults

Willemijn Heeren

Utrecht University

Abstract

How does the perception of a new phoneme contrast develop? In answering this question we consider two hypotheses: i) Acquired Distinctiveness: before learning, differences between and within phoneme categories are hardly discriminable. Through training, the phoneme boundary is learnt. ii) Acquired Similarity: before learning, differences between and within phoneme categories are well discriminated. Through training, only the phoneme boundary remains discriminable. In a pretest-training-posttest design, Dutch adults learnt the British-English pseudowords *thif* and *sif*: the first consonant in *thif* is not a phoneme of Dutch. Between pretest and posttest with materials from one speaker, participants were trained with speech from five other speakers. This forced listeners to form abstract phoneme categories. The results show that trained listeners performed better in the posttest than control listeners. However, in general the control group, who received no training, was difficult to distinguish from the trained listeners. With respect to the research question we found that discrimination levels increased as a result of training.

1 Introduction

The world's languages differ in their phoneme inventories. While learning a first or second language, a listener must acquire such a set of phonemes. In the continuous flow of speech sounds, however, two instances of one phoneme may show great diversity, due, for example, to speaker characteristics or speech rate. Moreover, two acoustically similar speech sounds may actually be realizations of two different phonemes. Native listeners tend to put sharp boundaries between phoneme categories in their language. Generally, discrimination of sounds taken from different sides of the phoneme boundary is higher than discrimination of within-category tokens. Assuming that this discrimination pattern is a native listener's end state, this would also be what a language learner is trying to achieve. But how does one learn to perceive novel phonemes?

Early studies that tried to change phoneme perception through laboratory training were not very successful (see Strange & Jenkins, 1978 for a review). Later studies, however, have shown that nonnative phoneme contrasts can be learnt through relatively short laboratory training (e.g. Jamieson & Morosan, 1986; Lively, Logan, & Pisoni, 1993). But there are also nonnative contrasts for which training is unnecessary. Best, Traill, Carter, Harrison, & Faber, (2003) showed that nonnative phoneme contrasts that do not fall within the acoustic range exploited by one's native language, can be discriminated quite well.

To study the perceptual development of novel phoneme contrasts, we consider two opposing hypotheses, Acquired Distinctiveness and Acquired Similarity, that have been proposed to explain the learning of native phonemes (Liberman, Harris, Kinney, & Lane, 1961). Both hypotheses deal with the degree to which within-category and between-category differences can be discriminated by a listener. The first hypothesis, Acquired Distinctiveness, says that listeners learn to perceive differences between those speech sounds that they are trained to

categorize differently, although they were unable to hear any differences before learning this new phoneme contrast. After training, discrimination of stimuli on different sides of the phoneme boundary has improved. In support of this hypothesis, a training study by Jamieson and Morosan (1986) reported increased discrimination at the phoneme boundary without within-category improvement.

The second hypothesis, Acquired Similarity, states that both within-category and between-category speech sounds can be distinguished well before training. As a result of training, however, perceptual sensitivity to speech sounds that belong to the same category decreases, in such a way that only above-chance discrimination of the stimuli straddling the phoneme boundary remains. This type of learning seems similar to the way infants treat speech sounds during their first year of life (Pisoni, 1991). Werker & Tees (1984), for example, have shown that infants of six to eight months of age can discriminate a natural speech contrast that their language environment does not contain, and that is not discriminated by adults from that language. But after about ten to twelve months infants lose the ability to discriminate these phoneme pairs. It is not probable, however, that the infants' representations lie at a phonemic level.

The question this paper aims to answer is: how do Dutch listeners learn the British English /θ-s/ contrast, where /θ/ is not a phoneme of Dutch? Do Dutch adults learn this contrast in accordance with Acquired Distinctiveness or Acquired Similarity? We tried to answer this question by means of a laboratory training study, run with Dutch adult listeners. Both before and after training, the perception of the nonnative phoneme continuum was assessed in absolute identification and discrimination tests. A control group that did not participate in training sessions, was also tested.

A subquestion that was addressed in the present study is whether perceptual learning is influenced by whether or not subjects know which language they are learning. Since most Dutch have learnt some English in primary and secondary school, the /θ-s/ contrast may not be entirely new to them. However, knowing that /θ/ is different from /s/ does not imply that a Dutch listener can differentiate the two acoustically. Moreover, the Dutch often produce /s/ when trying to pronounce /θ/ (Collins & Mees, 1999), which shows that they have difficulties with the English phoneme. We wanted to find out whether listeners who are told that they are learning a contrast from English, a language they are not completely unfamiliar with, benefit from this knowledge as opposed to learners who are told that they are learning a phoneme contrast from a foreign language.

2 Method

2.1 Materials

Eight-step continua were synthesized for the British-English phoneme contrast /θ-s/ by means of linear spectral interpolation (van Hessen, 1992). The phonemes occurred in the onset position of a pair of (both in Dutch and in English) nonsense words: *thif* ~ *sif*. Nonsense words were used to exclude word frequency effects or lexical bias (cf. Ganong, 1980). The phoneme continua were based on speech from six speakers of Standard English, both males and females. Figure 1 shows spectrograms of the eight-step continuum from one of the six speakers. The location of the phoneme boundary in each continuum had been determined from a classification study with 31 native British English listeners.[1]

[1] Collection of the materials was supported by grant R30-579 from the Netherlands Organisation for Scientific Research (NWO).

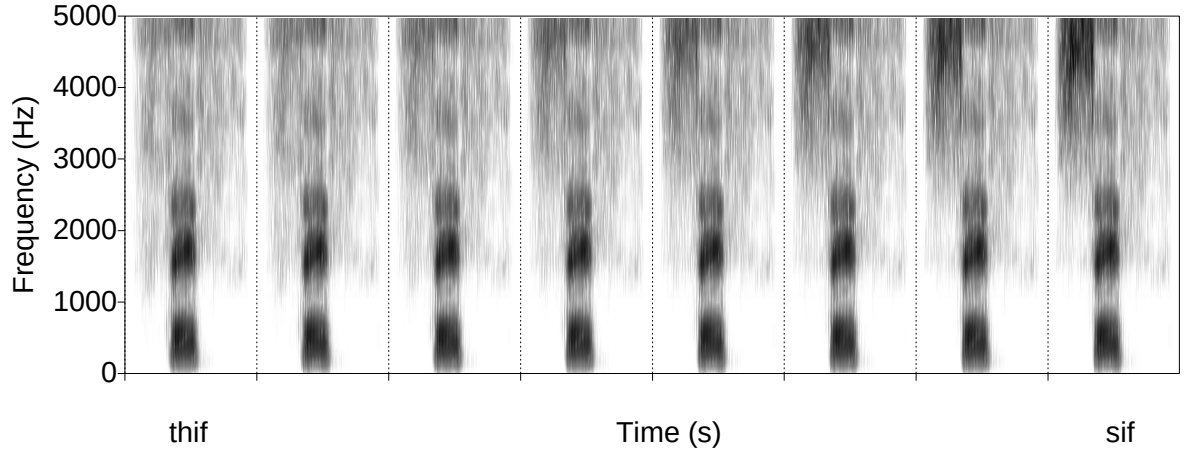


Figure 1. Spectrograms of the *thif-sif* continuum from one of the speakers.

2.2 Participants

Thirty-four students, native speakers of Dutch reporting normal hearing, participated in the test. None of them were or had been students of English. Seventeen subjects received training between pretest and posttest. The other half, the control group, participated only in pretest and posttest, with a time interval of approximately a week.

2.3 Design

The experiment was a pretest-posttest design. In both pretest and posttest four interval AX (4IAX) discrimination and absolute identification with speech from one male speaker were administered. At the end of the posttest a short questionnaire was given.

On each trial of 4IAX discrimination, two different stimuli A and B were presented in one of eight possible orders: AB-AA, AA-BA, BA-AA, AA-AB, BB-AB, BB-BA, AB-BB or BA-BB. The listener had to indicate whether the first or the second pair consisted of the same stimuli. Both one-step and three-step discrimination were included. Learning by Acquired Distinctiveness would become most apparent in the one-step test, since the discrimination level in this test was low from the beginning. On the other hand, learning by Acquired Similarity would show in the three-step test, since its initially high discrimination levels could fall as a result of training. The 4IAX discrimination test was expected to reflect both phonemic and auditory perception of the stimulus pairs (Pisoni, & Lazarus, 1974; Gerrits, 2001).

In absolute identification the listener has as many response options as there are stimuli. Since the listener is asked to indicate exactly which stimulus he heard, this test gives us insight into the listeners' control over the continuum.

For training, a classification design with trial-by-trial feedback on the correctness of the participants' responses was used. Classification was preferred over discrimination, since it directs the participants' attention towards the existence of two categories (e.g. Jamieson, & Morosan, 1986). The training materials consisted of phoneme continua from five speakers, both males and females, other than the test speaker. Speaker variation was included to encourage robust category formation (Lively, Logan, & Pisoni, 1993).

2.4 Procedure

Listeners were tested individually. Half of them were told that they would hear English nonsense words, the other half of the participants were told that they would hear words from a foreign language. All experiments were run in a relatively quiet room at the Utrecht Institute of Linguistics OTS. A laptop computer was used both to present the stimuli at random and to register responses. Stimuli were presented over Beyerdynamic DT 770 headphones at a comfortable listening level.

The pretest was completed on the first day. On subsequent days, training sessions were run until the listener classified the new phoneme contrast correctly in at least 85 % of the trials. On the final day the posttest took place. This test was of a similar content as the pretest, apart from a short questionnaire that was given after the listening tests. In this, the listeners were asked about the spelling of the newly learnt words. The following subsections will discuss the procedures of the listening tests in more detail.

2.4.1 4IAX Discrimination

The first test administered during pretest and posttest was 4IAX discrimination. This test was given twice, once with one-step stimulus pairs (i.e. 1-2, 2-3, …, 7-8) and once with three-step stimulus pairs (i.e. 1-4, 2-5, …, 5-8). The order of these tests was balanced across subjects.

Listeners received written instructions in which they were asked to indicate whether the first or the second word pair they heard consisted of the same stimuli. It was stressed that differences could be small. Responses were given by striking one of two keys on the computer keyboard. A short task introduction was given, consisting of eight four-step stimulus pairs from the same continuum. Inter-stimulus intervals were set at 300 ms, inter-pair intervals at 500 ms, and response times were unlimited. The eight different orders per stimulus pair were each presented four times, resulting in 224 trials for the one-step and 160 trials in the three-step test. Trial order was randomized and there were three short breaks at regular intervals.

2.4.2 Absolute Identification

The second test in pre- and posttest was absolute identification. Listeners received written instructions, telling them to indicate exactly which stimulus they had heard from the continuum. The instructions included a picture of a row of eight buttons numbered ‘1’ to ‘8’ from left to right. Over the first and eighth buttons, pictures of a man wearing differently colored headphones were shown. It was explained that each button hid a unique word. The words changed in steps from the name of the first man (behind button 1) to the name of the second man (behind button 8). Next, the stimuli were introduced five times in increasing, decreasing and random sequences. Listeners were instructed to listen very carefully and to try to remember which button corresponded to which word. During testing, stimuli were presented 20 times in random order, resulting in a total of 160 trials. Listeners responded by mouse-clicking one of eight on-screen buttons. Response times were unlimited and there were two short breaks at regular intervals.

2.4.3 Classification with feedback

During training, a classification design was used. The categories were represented by the pictures of the men and were introduced as the men’s names. Listeners had to reach a mean score of at least 85% correct identifications over two subsequent training tests before

proceeding to the posttest. The test was introduced by the pronunciation of both endpoint stimuli by each of the five training speakers.

Listeners received immediate feedback on each trial, informing them of the correctness of their choice. The percentage of correct responses so far was shown regularly. Training tests each contained twelve repetitions per stimulus, resulting in 480 trials. There were three short breaks at regular intervals.

3 Results

Training results were represented as percentages of *sif*-responses per stimulus for each of the five speakers. Next, the phoneme boundary was determined at the 50%-point for each of the five speakers in the training set and for each listener separately. Also, boundary widths were determined by calculating the 25% - 75% range. Boundary widths reflect the steepness of the boundary. For 4IAX discrimination, percentages of correct responses were determined per stimulus. From absolute identification results, the mean responses to each of the stimuli and their variances were determined.

3.1 Training

The participants' results from the first and the last training session were compared with the English norm defined by 31 native listeners. The mean boundaries in the first and last training sessions differed significantly from those defined by the English listeners [$F(1,221)=7.9$, $p=.005$ and $F(1,222)=7.7$, $p=.006$, respectively]. Post-hoc analyses showed, however, that the mean boundary of only one speaker differed from the English norm in both training sessions. The Dutch listeners perceived more stimuli as /s/. In addition, the boundary value of one more speaker in the last training differed from that of the English listeners; again more stimuli were judged as /s/.

The results from the first training session did not lead to a full cross-over in 36% of the cases: usually, no 25%-points were found, which means that listeners did not consistently choose the /θ/ category for stimuli at that end of the continuum. In the last training session, this number had decreased to only 6.3%, approximately equal to the number of cases the natives missed (6.5%). The boundary widths of the listeners for whom cross-overs were defined, differed significantly from the native English ones in the first training session [$F(1,186)=52.4$, $p<.001$]. In the last training session, however, these differences were no longer present.

3.2 4IAX Discrimination

For one-step 4IAX discrimination, a repeated measures ANOVA was run with within-subjects factors Test (pretest vs. posttest) and Stimulus Pair (1 through 7), and between-subjects factors Listener Group (trained vs. control) and Language (English vs. foreign language). The results are shown in figure 2.

First of all, a main effect of Test was found, $F(1,30)=14.9$, $p=.001$. Listeners gave more correct answers in the posttest than in the pretest. Secondly, a main effect of Stimulus Pair was found, $F(6,180)=10.2$, $p<.001$. This means that listeners were not equally good at distinguishing all pairs of stimuli. No main effects of Listener Group or Language were found. The absence of these effects means that the test listeners did not perform considerably better than the controls, and that listeners in the English condition did not benefit from their knowledge of what language they were attending. A post-hoc ANOVA on the posttest data

revealed only an effect of Listener Group, $F(1,224)=7.5$, $p=.007$): trained listeners scored better than those in the control group.

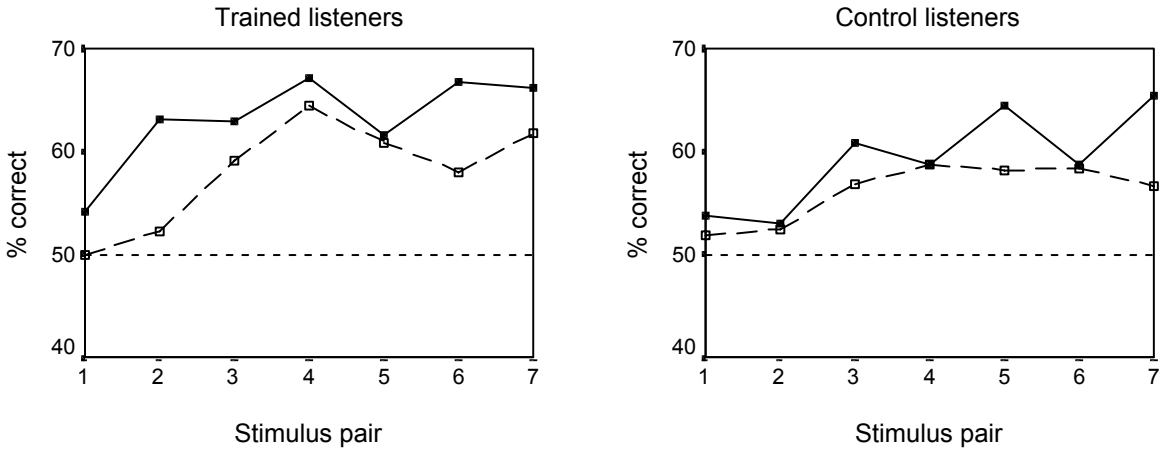


Figure 2. One-step 4IAX discrimination results for both pretest (dashed) and posttest (solid) with a reference line at chance level.

Three-step 4IAX discrimination showed similar results. Main effects of Test [$F(1,30)=13.4$, $p=.001$) and Stimulus Pair [$F(3.4,101.4)=9.3$, $p<.001$) were significant. Again, no main effect of Listener Group or Language was found. An ANOVA on the posttest data showed that test listeners performed better than controls, $F(1,155)=10.8$, $p=.001$.

3.3 Absolute identification

A repeated measures ANOVA was run on the mean responses shown in Figure 3 below, with within-subjects factors Test (pretest vs. posttest) and Stimulus (1 through 8) and between-subjects factors Listener Group (trained vs. control) and Language (English vs. foreign language).

Firstly, a Test × Stimulus interaction [$F(4.5,136)=3.9$, $p=.003$] was found. At the /θ/-end of the continuum, participants became better at identifying the stimuli, while this was not the case at the /s/-end. This can be seen in figure 3 by the trend towards the ideal case of absolute identifications at the /θ/-end only. Furthermore, main effects of Test [$F(1,30)=28.2$, $p<.001$] and Stimulus [$F(2.6,78.3)=794.1$, $p<.001$] were found. But again, no effects of the between-subjects factors Listener Group and Language were present.

The mean variances showed main effects of Test [$F(1,30)=20.6$, $p<.001$] and of Stimulus [$F(4.3,129.7)=5.5$, $p<.001$). Posttest variances were smaller than those in the pretest, meaning that participants had become better at identifying the stimuli.

In summary, differences between pretest and posttest were present in the three tests. Trained listeners' progress, however, was hard to discriminate from that of control listeners. The fact that some test listeners needed only a few training sessions to reach criterion may partly account for this phenomenon, which will be investigated in subsection 3.4.

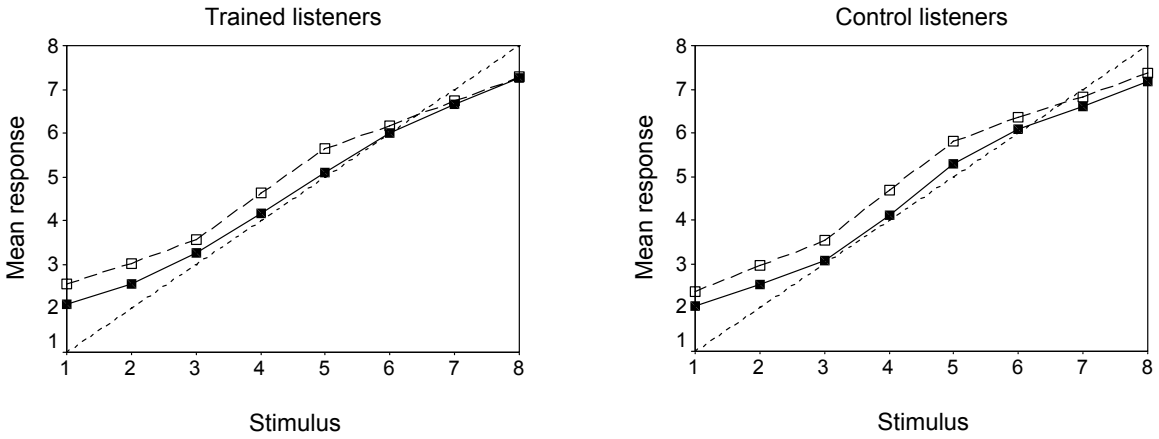


Figure 3. Absolute identification results for pretest (dashed) and posttest (solid). The ideal case of error-free perception is shown by the short-dashed line.

3.4 The effect of amount of training on pretest-posttest differences

The amount of training needed to reach criterion varied among test listeners: from 960 to 7,680 training trials before reaching criterion. It is conceivable that those listeners who needed more training also show larger differences between pre- and posttest results. Therefore, the trained listeners' data were tested in repeated measures ANOVAs, this time with Training Amount as a covariate. Only the findings that differ from the effects described in sections 3.2 and 3.3 will be reported here.

Three-step 4IAX discrimination showed a Test × Training Amount interaction [$F(1,15)=4.8$, $p=.044$] as well as a main effect of Training Amount [$F(1,15)=16.4$, $p=.001$]. The rise in correct responses between pretest and posttest tended to be larger as the amount of training increased. The mean variances of the stimuli in absolute identification also showed a main effect of Training Amount, $F(1,15)=5.5$, $p=.033$.

3.5 Posttest Questionnaire

After the perception tests, listeners wrote down the words they had heard as the two extremes during absolute identification. The reported sets varied greatly. From the test group, six listeners reported the correct pair of words, but none of the control listeners succeeded in doing this. Remember that acoustically only the first consonant varied between stimuli *thif* and *sif*. Three types of errors were identified. Listeners (a) replaced target phonemes with other phonemes, (b) added phonemes that were not present in the signal or (c) reported hearing differences between the vowels or between the coda consonants.

In total, control listeners made more errors: 55 as opposed to 37 for the trained listeners. The controls made the majority of these, 49%, in their responses to the first consonant, i.e. the target phoneme. They used almost twice as many substitutions, and also additions that resulted in complex onsets such as *stif* instead of *sif*. Furthermore, they more often replaced the coda consonant with another one, as in *striss* instead of *thif*.

4 Discussion

Listeners participated in training sessions until they learnt the contrast to the predetermined level of 85% correct. The training results showed that the phoneme boundaries were already close to the English norm during the first training session. But the widths of the phoneme

boundaries became much smaller and even indistinguishable from the English norm as a result of training.

The pre- and posttest revealed main effects of Test for both discrimination and absolute identification tests. However, the absence of Test × Listener Group interactions showed that listeners — whether in the test or control group — all performed better when they did the tests for the second time. Differences between the test and control groups were only present within the posttest data of three-step discrimination and of the variance results in absolute identification. As for three-step discrimination, test listeners performed better than controls, especially at the /θ/-end of the continuum in the posttest. Response variances in absolute identification were smaller for the test group.

By taking into account the amount of training participants received, we found that listeners who needed many training sessions to reach the criterion of 85% correct responses, showed a larger difference between their pretest and posttest results for three-step discrimination and for the variances in absolute identification.

The main research question of this paper was: do Dutch adults learn the British-English /θ-s/ contrast in a way compatible with Acquired Distinctiveness or with Acquired Similarity? We found no support for learning by Acquired Similarity. For this to be the case, perceptual sensitivity to speech sounds that belong to the same category should decrease, but we found no such effect at all. On the contrary, the improvement we found for the discrimination tests mainly occurred within instead of between categories. So these findings do not strongly support our other hypothesis, Acquired Distinctiveness, either.

Despite the progress during training, little evidence of an increase in discrimination levels at the phoneme boundary was found, contrary to earlier findings (Jamieson, & Morosan, 1986). This may have been caused either by the nature of the tasks used in pretest and posttest, or by the participants' high pretest levels. Firstly, the tasks used may have directed the listeners' attention too much towards the acoustic differences between the stimuli by testing auditory instead of phonemic perception. However, we expected to find a combination of these listening levels in our results (Gerrits, 2001; Pisoni, & Lazarus, 1974). Secondly, most participants performed already quite well in the pretest (see, for example, figure 2). The room left for improvement as a result of training was thereby restricted and may have made such improvement difficult to distinguish from improvement by task repetition. We also think that the pretest was a training in itself due to its length, which helped control listeners to improve their scores in the posttest. Most earlier studies, however, did not involve a control group (e.g. Strange, & Dittmann, 1984; Logan, Lively, & Pisoni, 1991) and could therefore only report the test group's progress.

An explanation that may account for a portion of the errors made by the participants in reporting which words they had heard, is an effect similar to 'verbal transformation' (Warren, 1961). In Warren's study, listeners heard uninterrupted repetitions of a word or phrase. They reported hearing words that were not present in the speech signal. In our study, comparable misperceptions were reported. Even though there were silent intervals between subsequent presentations of word forms, perceptual disturbances may still have occurred. Schouten & van Hessen (1998), for example, reported that their participants 'hallucinated' as a result of prolonged exposure to speech sounds taken from a phoneme continuum. The fact that control listeners made more errors than trained listeners may be explained by the differences in speaker variation the two listener groups had been exposed to. During the course of the experiment, trained listeners heard the word forms spoken by six different voices, whereas control listeners heard only the test speaker's voice.

A sub-question that was addressed in the present study was whether the availability of knowledge of the language you are learning influences perceptual learning. We found that participants who knew they were listening to English, did not benefit from this knowledge. So either listeners in both the English and the Foreign Language conditions used their knowledge of English equally, or neither of the groups accessed this knowledge.

5 Conclusion

Dutch listeners improved their perception of British-English /θ-s/ during training. Trained listeners performed better in the posttest than in the pretest and in several respects they also did better than the control group. Their improved performance excluded Acquired Similarity, but did not strongly support Acquired Distinctiveness either. This lack of effect was thought to be due to both the high pretest performances of our participants and the nature of the tests used in pretest and posttest. Furthermore, control listeners, who received no training, also improved by simply performing the tests in pretest and posttest twice. These results show that it is important to include a control group into the design of a phoneme training study, which has often not been the case. Finally, listeners who knew that they were listening to British-English did not benefit from this knowledge opposed to listeners who were told they were listening to a foreign language.

References

Best, C.T., Traill, A., Carter, A., Harrison, K.D. & Faber, A. (2003). !Xóõ click perception by English, Isizulu, and Sesotho listeners. In M. J. Solé, D. Recasens & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences; Barcelona, 3-9 Augustus 2003* (pp. 853-856).

Collins, B.S. & Mees, I.M. (1999). *The phonetics of English and Dutch*. Leiden: Brill.

Ganong, W.F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Phsychology: Human Perception and Performance, 6*, 110-125.

Gerrits, E. (2001). *The categorisation of speech sounds by adults and children*. Doctoral Dissertation, Utrecht University.

Hessen, A.J. van (1992). *Discrimination of familiar and unfamiliar speech sounds.* Doctoral Dissertation, Utrecht University.

Jamieson, D.G. & Morosan, D.E. (1986). Training non-native speech contrasts in adults: acquisition of the English /ð/-/θ/ contrast by francophones. *Perception & Psychophysics*, *40*, 205-215.

Logan, J.S., Lively, S.E., & Pisoni, D.B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, *89*, 874-886.

Liberman, A.M., Harris, K.S., Kinney, J.A. & Lane, H. (1961) The discrimination of relative onset-time of the components of certain speech and nonspeech patterns, *Journal of Experimental Psychology*, *61*, 379-388.

Lively, S.E., Logan, J.S., & Pisoni, D.B. (1993). Training Japanese listeners to identify English /r/ and /l/. II The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America*, *94*, 1242-1255.

Pisoni, D.B., and Lazarus, J.H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America*, 55, 328-333.

Pisoni, D.B. (1991). Modes of processing speech and nonspeech signals. In I.G. Mattingly & M. Studdert-Kennedy (Eds.): *Modularity and the motor theory of speech perception* (pp. 225-238). Hillsdale NJ: Lawrence Erlbaum Associates.

Schouten, M.E.H., & Hessen, A.J. van (1998). Response distributions in intensity resolution and speech discrimination. *Journal of the Acoustical Society of America*, *104*, 2980-2990.

Strange, W. & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics*, *36*, 131-145.

Strange, W. & Jenkins, J.J. (1978). Role of linguistic experience in the perception of speech. In R.D. Walk & H.L. Pick (Eds.): *Perception and experience* (pp. 125-169). New York: Plenum Press.

Warren, R.M. (1961). Illusory changes of distinct speech upon repetition – the verbal transformation effect. *British Journal of Psychology*, *52*, 249-258.

Werker, J.F. & Tees, R.C. (1984). Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*, 49-63.

Planning in speech melody: production and perception of downstep in Dutch

Vincent J. van Heuven

Universiteit Leiden

Abstract

This paper studies the planning of downsteps in Dutch enumerations of two to six items long. Is it true that the available pitch range is subdivided into smaller downsteps as the number of accented items in the enumeration is larger, and how precisely do speakers programme the stepsize? Do listeners use the size of the (first) downstep to project the length of the enumeration? The results show that the size of the first downstep (when scaled in ERB) is exactly proportional to the number of items in the enumeration – indicating a high degree of planning on the part of the speaker – but that listeners are largely insensitive to the stepsize as when asked to predict the length of the utterance.

1 Introduction

It is obvious that some form of planning takes place during speech production. One line of evidence comes from research on speech errors. Nooteboom & Cohen (1974) show that speakers program their segments some seven syllables ahead relative to the moment of speaking. At the same time they programme meaningful units some seven morphemes (or words) ahead. The estimated size of the look-ahead window is derived from the distance (measured in linguistic units, i.e. syllables or words) that intervenes between the source and the target unit involved in the anticipation type of speech error. Someone who says *to colerate quite a lot* instead of *tolerate quite a lot* has mistakenly replaced the first phoneme of *tolerate* by the first phoneme of *quite*. This implies the words *tolerate* and *quite* were simultaneously present in the speaker's mind, which amounts to a sequence of four syllables.

Planning in speech production is not restricted to segmental phonetics. There is a wealth of evidence that considerable planning occurs in prosody as well. Nooteboom & Cohen (1974) reasoned that the so-called flat hat in Dutch intonation exists by virtue of planning ahead. The flat hat is the most frequently used intonation pattern that links the last and the second-but-last accent in a sentence. The pitch rises from the low declination level to the high level on the pre-final accent, remains high until the final accent, and only then drops down to the low declination line, to be followed by a low boundary tone. It follows from this description that the speaker must know at the point in time where the pre-final rise is executed, that only one more accent is to follow before the end of the sentence. Nooteboom & Cohen (1974) show that the distance between the rise and the fall accent that are linked by a flat hat, again spans some seven syllables. Planning in intonation is also visible in the course of the declination, i.e. the imaginary line that links the low pivot points in the pitch curve that can be measured in a spoken sentence. The longer the sentence, the higher the starting pitch but the slower the rate of descent, such that the speaker always reaches the same low pitch at the end of the utterance. Note that it may not be the planning of the onset pitch per se that is at issue here. Probably, when the speaker inhales, he has a rough idea of how much material he is going to speak until the next inhalation pause – which typically coincides with a deep prosodic

boundary such as the end of an utterance. This is what Liberman & Pierrehumbert (1984:220) have called 'soft' pre-planning.[1] Before longer sentences, then, the speaker will take a deeper breath than before short sentences, so it is the volume of air trapped inside the lungs that primarily determines the high onset pitch rather than some complex computational act the speaker performs on the pitch contour (which would be 'hard' pre-planning in Liberman & Pierrehumbert's terms). Whether hard or soft, the reflexes of this prosodic planning enable the listener to project the end of the sentence with substantial accuracy (Grosjean, 1983, for English; Leroy, 1984, for Dutch). My present contribution targets a similar phenomenon, the perceptual effects of which, however, have received only scant attention in the literature.

't Hart, Collier & Cohen (1990) describe the intonation of Dutch as a sequence of rises and falls between a lower and an upper declination line. There are five types of rise (called 1 through 5) and an equal number of falls (called A through E). Fall E is the movement of crucial importance to this chapter. It is described as a half fall, i.e., it does not drop over the full interval between the high and low declination – nominally 6 semitones – but only covers part of the interval. Fall E seems to be used in two different functions in the intonation system of Dutch. When it is the only half fall, it is sentence final, and signals an intention on the part of the speaker expressing 'what I am saying here is actually superfluous' (Keijsper, 1985; van Heuven & Kirsner, 1999). In this function, fall E indeed ends more of less midway between the high and the low declination line, or as Nooteboom and Cohen (1984:159) say: 'Fall E […] is typical for the "street call" intonation, and often creates the impression of a musical interval, a minor third' – which would place it exactly halfway between upper en lower declination in Dutch. However, fall E may also occur recursively on successive intonation phrases. In Nooteboom & Cohen (1984:161) we find seven grammatical stylised pitch contours on the sentence *Wij proberen de SPRAAK te beGRIJpen en te beHEERsen* 'We attempt the speech to understand and to control'.[2] The seventh contour shows a full rise on the accent on *SPRAAK*, and falls E on the syllables *GRIJ* and *HEER*. Since the two incomplete falls together span the distance between upper and lower declination, the stylised contour has been drawn such that each fall E spans half the distance. This pattern is the shortest exemplar of the 'terrace pattern' ('t Hart et al., 1990:166). Now, one might wonder what would happen if the utterance comprises a larger number of intonation phrases, say three, four or five, each of which would end in a partial fall E.[3] As a first approximation we would expect the speaker to divide the range between upper and lower declination into as many equal-sized intervals as are needed to make the required number of steps down. This situation typically arises in extended enumerations of the type *Ik wil een salade met MANgo, DRUIven, AARDbeien, meLOEN, DAdels en BRAmen* 'I want a salad with mango, grapes, strawberries, melon, dates and raspberries'. There would be a full rise on *MAN*, and incomplete falls E on each of the following items in the enumeration. Three incomplete falls would then require each one-third of the nominal 6-semitone span between upper and lower declination, i.e. 2 semitones (st). By the same token, a five-item enumeration would be realised as a full rise followed by four one-quarter falls E, and so on for even longer lists. If we find this specific behaviour on the part of the speaker, this would be a convincing case of so-called 'hard' pre-planning in intonation. It would indicate that the speaker knows how many items after the first item in his enumeration are to follow until the end, and then subdivides the available pitch range into the required

[1] Soft preplanning is tantamount to behavioural common sense, as opposed to 'hard' preplanning, which would involve right-to-left computation of the contour (van den Berg, Gussenhoven & Rietveld, 1992: 354-355).

[2] Throughout this chapter small caps denote pitch-accented syllables.

[3] The fall on the very last IP in a sequence of IPs is transcribed as A rather than E by 't Hart et al. (1990).

number of equal-sized steps. Moreover, it may be the case that the speaker anticipates on the length of his enumeration by making the rise on the first member of the enumeration larger as the enumeration is longer. This paper examines the speech behaviour of a sample of Dutch speakers reading aloud a set of enumerations of variable length (two to six items). We also wish to determine the communicative effects of the prosody of enumerations by asking listeners to predict how many items an enumeration will contain, when the original utterance is electronically truncated after the first downstep.

In a more recent autosegmental account of Dutch intonation (Gussenhoven, Rietveld & Terken, 1999) the terrace pattern is analysed as one possible surface realisation of a sequence of downstepped high tones (symbolised as '!H'). A downstepped high tone has a high target but less high than the immediately preceding high tone. A sequence like %L H* !H* !H* L% would then represent a terrace contour with one full rise and two incomplete falls, each spanning half the distance between upper and lower declination. Normally, however, the accents would be realised as (!)H*L configurations, so that an alternative pattern would be %L H*L !H*L !H*L L%. The two patterns, with and without the L tones, sound very much alike. It makes sense, therefore, to derive the terrace pattern from the sequence of rise-fall accents by optional deletion of the L and subsequent spreading of the high tone until the onset of the next !H, as illustrated in Figure 1.[4]

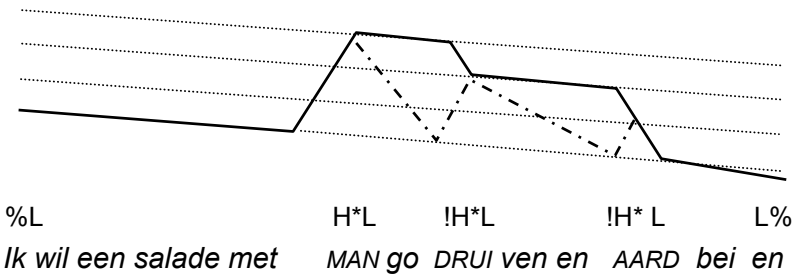


Figure 1. Underlying downstepped H*L contours (dashed) and surface terrace contours (solid).

The IPO model simply predicts that the 6-st span between upper and lower declination is bridged in two equal steps of 3 st. The autosegmental account computes the downstep values by applying a constant 'downstep ratio', such that a downstepped !H* target has a pitch (when expressed in hertz) that is a fixed percentage lower than the immediately preceding (!)H* target. Presumably the downstep ratio is smaller as the number of downsteps to be executed is larger.[5] A 3-st downstep is tantamount to a 0.8 ratio; there seems to be no principal difference here between the IPO and the autosegmental account of the phenomena. However, the autosegmental model claims that the very last downstep is considerably larger than all earlier ones – which detail is not accounted for in the IPO model.[6]

---

[4] Van den Berg et al. (1992: 341) suggest that the L is not deleted but squeezed so tightly onto the next !H* that it has no phonetic realisation anymore; in this latter view the H* spreading is the cause rather than the effect of the (virtual) L deletion. The autosegmental model predicts that the first downstep takes up one-third of the span between upper and lower declination (see figure 1), whilst the IPO account predicts a step down of half the span.

[5] Van den Berg et al. (1992: 354-355), however, show that (speaker-individual) constant downstep ratios across a range of enumeration lengths (two to five items) yield prediction errors that are only marginally poorer than those obtained for length-dependent downstep ratios.

[6] Liberman & Pierrehumbert (1984) include a so-called final downstep constant to account for this phenomenon in English. The magnitude of the lowering constant seems to be a language-specific parameter, and may in fact be equal to zero.

The standard version of the IPO grammar of Dutch intonation assumes that the upper and lower declination run parallel, predicting that [+full] accent-lending movements have the same excursion size irrespective of their position in the utterance. Although the onset pitch of a sentence is higher as it is longer, the excursion size of the accents remains unaffected. We cannot exclude the possibility that sentences with enumerations have a higher first accent as the enumeration is longer. If this is so, then the size of the first accent-lending rise might be a perceptual cue to the length of the enumeration (see also note 5). A second(ary) cue would then be the size of the first downstep, which would either be manifested as the interval of the fall after the first terrace, or the interval between the H*L and the first !H*L accents. The smaller the pitch interval between these two accents, the longer the enumeration should be.

2 Production experiment

2.1 Methods

Nine adult native speakers of Dutch (five male, four female) read the following ten sentences in different random orders each:

A2. Ik wil een salade met MANgo *en* DRUIven.
A3. Ik wil een salade met MANgo, DRUIven *en* AARDbeien.
A4. Ik wil een salade met MANgo, DRUIven, AARDbeien *en* meLOEN.
A5. Ik wil een salade met MANgo, DRUIven, AARDbeien, meLOEN *en* DAdels.
A6. Ik wil een salade met MANgo, DRUIven, AARDbeien, meLOEN, DAdels *en* BRAmen.
'I want a salad with mango, grapes, strawberries, melon, dates and raspberries'

B2. Op mijn lijstje staan BLOEM *en* BOter.
B3. Op mijn lijstje staan BLOEM, BOter *en* EIeren.
B4. Op mijn lijstje staan BLOEM, BOter, EIeren *en* ROOM.
B5. Op mijn lijstje staan BLOEM, BOter, EIeren, ROOM *en* MELK.
B6. Op mijn lijstje staan BLOEM, BOter, EIeren, ROOM, MELK *en* roZIJnen.
'On my list are flour, butter, eggs, cream, milk and raisins'

Before reading the stimuli speakers heard and repeated exemplars of a sentence with a (lexically different) three-item enumeration, with stylized resynthesized terrace contours in order to improve the odds that the speakers would then continue to use this intonation pattern when reading the actual stimuli.

Recordings were made in a sound-insulated booth through a Sennheiser MKH 416 uni-directional condenser microphone onto computer disk (16 bit, 16 kHz). Fundamental frequency curves were measured using the autocorrelation method implemented in Praat (Boersma & Weenink, 1996), interactively corrected when the algorithm had made an error, and psychophysically scaled in ERB units (see Nooteboom, 1997, and references therein).

2.2 Results

As a first approximation I present a rather abstract representation of the pitch contours measured in the materials. For each utterance, Figure 2 displays the F0 value (in ERB) for the onset pitch (representing the low boundary tone %L), the low (L) immediately preceding the accent-lending rise on the first accented syllable in the enumeration, and the terminal F0 (L%). Between L and L% there is an array of measurement point representing the F0 maxima measured on the accented syllables in the successive items of the enumerations. The maxima are either true peak values (in the case of rise-fall accents) or the beginnings of rather level tones (when a terrace-type of contour was realised). The measurements are collapsed separately over the five male and four female speakers, and over the two lexically different

series of enumerations, but are broken down by enumerations of different lengths. Note that although the contours all share the same beginning of the utterance (the %L and L measurement points), they are lined up on the *end* of the contour, such that, for instance, the last downstepped H* ('H-0') accent peaks are always at the same position along the time axis. This alignment of the measurement points affords the clearest view on what is going on in the data. The X-axis represents a continuous time axis only for the longest contours (with length = 6). For all other contours, the time axis is discontinuous after the low onset of the accent-lending rise on the first item in the enumeration (indicated by in interruption of the lines for length 2 through 5 in figure 2). Note that the X-axis in this figure abstracts away from physical duration: all pitch pivot points are drawn at equal 'time' intervals.

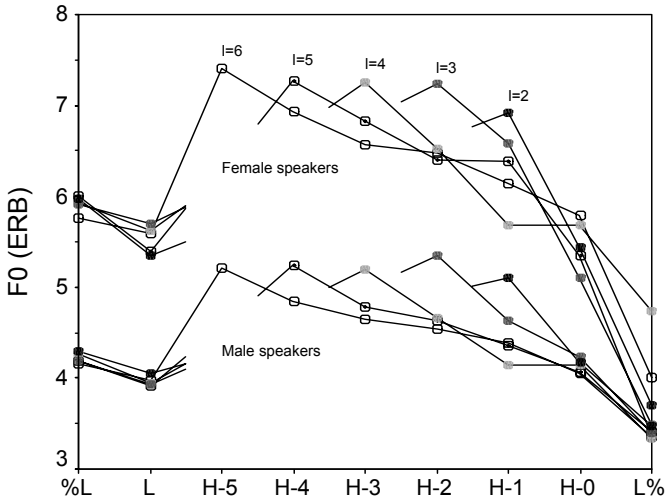


Figure 2. Fundamental frequency (F0 in ERB) of selected pivot points (see text) in contours produced by five male and four female Dutch speakers on enumerations of two to six items long. The time axis is normalized by inter-pivot distance, and discontinuous for all contours except for those with length = 6.

Three-way analyses of variance were done on the measured frequencies (in ERB) for each of the selected pivot points, with length of enumeration and sex of speaker as fixed factors and with the two lexical sentence types as a random factor. Table 1 lists the results for the effect of length of enumeration and for the interaction between length and sex of speaker. Of course, the effect of sex on the measured F0 values was always highly significant; given that the repetition rate of female voices is typically twice that of male voices, this is a predictable if not trivial effect, which we have not included in Table 1. The ANOVA results confirm the visual impression from Figure 2. The mean pitch values at the first two and the last two pivot points do not vary as a function of the length of the enumeration. The stretch of low declination in the precursor phrase is the same for all enumerations, and so is the pitch of the sentence-final downstepped H* accent (H_0) and the terminal pitch of the sentence (L%).

Counter to 't Hart et al. (1990), then, longer utterances do not start on a higher pitch than shorter utterances, at least not in this type of material. It is also apparent from Figure 2 that the F0 peak at the rise-fall accent on the first item of the enumeration is roughly constant across the entire range of enumerations. A one-way ANOVA on the peak F0 of the first accent shows no effect of the number of items to follow in the enumeration, $F(4,82)<1$. By and large, then, it seems that the first H* and the last downstepped !H* targets are constants,

irrespective of the number of downsteps that should be executed in between these two. It follows from this characterisation that the mean downstep size is smaller as the number of downsteps in the enumeration is larger. However, for an enumeration of a specific length the downsteps are equal, at least when F0 is expressed in ERB. That is, the downstep size is a linear function of the number of items in the enumeration. There seems to be no need, then, for a special lowering constant on the last downstep, as was advocated for English (Liberman & Pierrehumbert, 1984) and Mexican Spanish (Prieto, Shih & Nibert, 1996).

Table 1. Results of three-way ANOVA. Effect of length of enumeration and interaction of length × sex on nine selected pitch pivot points in intonation contours.

Pivot point	Length of enumeration				Length × sex			
	df_1=df_2	*F*	*p*	sign.	df_1=df_2	*F*	*p*	sign.
%L	4	3.9	.108		4	.4	.799	
L	4	.6	.684		4	2.0	.254	
H_{-5}								
H_{-4}	1	20.2	.139		1	1.9	.404	
H_{-3}	2	47.2	.023	*	2	.2	.822	
H_{-2}	3	17.7	.021	*	3	.2	.896	
H_{-1}	4	7.4	.039	*	4	2.2	.235	
H_0	4	.5	.767		4	3.9	.108	
L%	4	9.4	.026	*	4	6.1	.054	

Note: No effect or interaction can be measured at pivot point H_{-5}, since length of enumeration is a constant there.

As a final point in the analysis of the production data, let us consider the effect of length of the enumeration, i.e. the number of downsteps to be executed between the rise on the first item and the end of the sentence, on the size of the first downstep, i.e. the downstep on the second item in the enumeration. This is the first and the largest downstep in the sequence, and as such it potentially contains the earliest and clearest pitch cue that might allow the listener to project the length of the enumeration. Figure 3 presents the size of the first downstep as a function of the number of items in the enumeration for male and female speakers.

Figure 3 shows quite clearly that the size of the first downstep gets smaller as the number of items in the enumeration is larger. Since a constant range has to be divided into two, three, four etc. steps, depending on the length of the enumeration, we expect a reciprocal function of the type $\hat{y} = b_0 + b_1/x$ to optimally capture the relationship. Indeed the reciprocal function yielded the best fit to the data (better than logarithmic, power, exponential or linear fits). The relationship is more regular for the male speakers than for the females. Interestingly, the female downsteps tend to be larger than those of the men for the larger downsteps in shorter enumerations but asymptote to the same size across sexes for the longer enumerations.[7]

[7] The overall effect ties in with earlier observations that Dutch female speakers have larger pitch movements than their male counterparts, even when pitch is psychophysically scaled in ERB (Haan & van Heuven, 1999). Although this might also be taken as a cue that the ERB scale could be invalid for cross-sex comparison of pitch, I now believe that Dutch women genuinely have larger pitch movements than men. This belief is based on the fact that it is precisely the reciprocal function that captures the regularity in the downstep size most successfully. If the pitches had not been scaled in ERB, this result would not have been obtained.

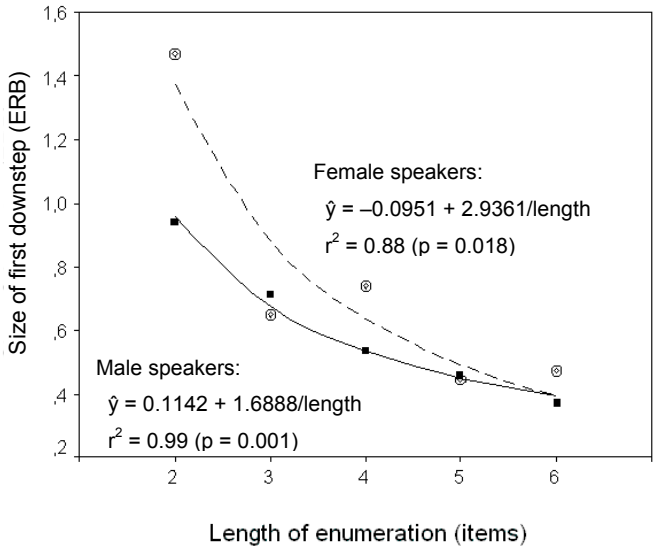


Figure 3. Size of first downstep (ERB) as a function of number of items in the enumeration for male and female speakers. Reciprocal functions have been fitted to the data points.

2.3 Conclusions

It seems that some planning mechanism has to be assumed in order to account for the results of the production study presented above. However, the speakers limit the burden of tonal planning to a minimum. For instance, they do not anticipate on longer utterances by starting off at a higher pitch nor do they execute a larger accent-lending rise on the first item in the enumeration. Rather, speakers realise all utterances, irrespective of their length, in the same register, bounded by a constant low onset pitch, a constant maximum on the first accent, a constant maximum on the last accent and a constant terminal pitch. The only planning that is brought to the task, is that the pitch range between the maximum on the first and the last accent is subdivided into a number of roughly equal-sized steps (when expressed in ERB); the step size then only depends on the number of steps that have to be walked down the terrace, i.e. on the number of items in the enumeration. So the speaker must know how many items his enumeration contains in order to select the required size of downstep.

For a perceptual follow up, we predict, then, that the only type of cue that allows the listener to project the length of the enumeration, is the size of the first downstep. So, in a gating experiment in which a larger initial portion of the utterance is made audible on successive presentations, the listener will have to wait until he has heard the portion of the utterance up to and including the second item in the enumeration; this is the point in time where the maxima of the first and second accent can be compared so that the size of the downstep can be computed. The size of the downstep should then be divided into the size of the accent-lending rise on the first item, and the listener will know how many downsteps will follow. These predictions will be tested in the next section.

3 Perception experiment

3.1 Introduction

In the perception study we targeted several potential cues that might help the listener project the length of an enumeration. As explained in section 1, the speaker might have chosen to execute a larger pitch rise on the first accent as the enumeration contains more items. Even

though our speakers did not exhibit this behaviour, it is still be possible that the listener might use such a cue when it is contained in the speech stimulus. Therefore we artificially decreased the size of the first accent (i.e. the peak F0 value of the accent) relative to the natural F0 used by the speaker. Also, we know that speaking rate is faster as the sentence is longer. Consequently, when hearing a fast initial portion of an utterance, the listener will be more likely to project a long enumeration than when hearing a relatively slow onset portion. In order to be able to determine the relative contribution of the three factors identified here, i.e. (i) speaking rate of utterance, (ii) size of first accent-lending rise and (iii) size of first downstep, these factors were systematically varied in the stimulus materials.

3.2 Methods

The basic materials for this experiment were the five sentences A2 through A6 as spoken by one of the five male speakers (the present author) in the production experiment. The sentences were divided into the precursor *Ik wil een salade met* 'I want a salad with' and the enumeration part. Precursors were taken from the utterances with enumerations containing two, four and six items; these had durations of 990, 910 and 810 ms, respectively. In order to keep the experiment reasonably short, no precursors originally spoken before three and five-item enumerations were used. The three precursors, with their natural speaking rate, were cross-spliced onto each of the five enumeration portions, yielding a set of three original utterances and twelve hybrids. The declination of the precursor to the four-item enumeration was stylised and used for all the stimuli. The enumeration portions were also stylised until the end of the second item, using a straight-line approximation of the accent-lending rise H* and the following terraces separated by the first downstep. The size of the downstep was chosen to be roughly in line with the values actually found in the five natural original utterances. The largest downstep, as found for the two-item enumeration, was set at 20 Hz; the smallest downstep, found for the six-item enumeration was given a value of 10 Hz. Intermediate downsteps were interpolated at 12.5, 15 and 17,5 Hz.[8] In order not to give away lexical clues, the conjunctive *en* 'and', which regularly precedes the last item of an enumeration in Dutch, was removed using the Praat waveform editor.[9] The accent-lending rise on the first item was either given its natural value (i.e., as produced by the speaker) or 0.25 ERB less or 0.50 ERB less. When a manipulated rise was used, the terrace patterns were shifted down in frequency so as to be precisely linked to the F0 peak of the rise on the first item. All stimuli were truncated immediately after the accented syllable of the second item in the enumeration, i.e. after /ˈdrœy/ in the word *DRUIven*.

In all 45 stimuli resulted from these manipulations: 3 (precursor rates) × 3 (rises on first item) × 5 (downsteps after first item). These were presented twice in different random orders to thirteen adult native Dutch listeners, in individual interactive sessions, over headphones in a quiet room. Subjects were instructed to indicate, with forced choice, for each utterance they heard, which of the five enumerations A2 through A6 they thought had most likely just been played to them. During the experiment the five sentences were displayed on a monitor in front

[8] The size of the downsteps was manipulated in the hertz domain for reasons of convenience. Within the very restricted range of F0 we needed for the stimulus manipulations, the ERB scale and the hertz scale are practically interchangeable. Note also that no systematic differences in temporal organisation were found for the enumeration parts of the utterances. Although there may still be subtle cues remaining in the segmental or temporal make-up of the five different enumeration parts, we take the view that the perceptual results are predominantly due to the manipulation of the downstep.

[9] Strictly speaking, the resulting stimulus is ungrammatical. Nevertheless, our listeners proved perfectly capable of performing their task (see results).

of the listener, who was instructed to click on one of five radio buttons which preceded the five sentences on the screen. After each response there was a 2-second pause before the presentation of the next stimulus.

3.3 Results

An analysis of variance on the estimated length of the enumeration (expressed in number of items) with precursor rate, rise and downstep as fixed factors revealed that precursor rate [$F(2,1125)=13.1$, $p<.001$] and downstep [$F(4,1125)=53.0$, $p<.001$] yielded significant main effects. There was no effect of the size of the accent-lending rise on the first item of the enumeration, nor were any of the interactions among the three factors significant ($F<1$ in all cases). The effects of precursor rate and of downstep on the length of the enumeration as estimated by the listeners, are displayed in Figure 4.

Figure 4 shows that the mean estimated length of the enumeration went up as the speaking rate of the precursor phrase was slower. A mean length of 3.68 items was estimated for the slowest precursor, 3.79 items for the intermediate rate and the highest estimate (4.11) was found for the fastest precursor. A post-hoc test for contrast (Scheffé, $p<.05$) shows that the fastest rate differs from the two slower rates, which do not differ significantly from one another.

A post-hoc test on the effect of downstep shows that the largest downstep (20 Hz) leads to the prediction of the shortest enumeration (2.84 items). The four smaller downsteps (10, 12.5, 15 and 17.5 Hz) yield estimations of enumeration lengths between 3.97 and 4.23 items) but these downstep sizes do not differ significantly from each other. It would seem, then, that the largest downstep (20 Hz) was in fact so large that it prompted our listeners to assume that the speaker would lower his pitch all the way down to the low declination line, which of course would then be an unmistakable cue for finality, i.e. that no more items were to follow after the word *DRUIven*. When no finality cue could be picked up, the estimates of the remaining number of items in the enumeration were roughly in the middle of the possible range, i.e. enumerations between 3 and 6 items long.

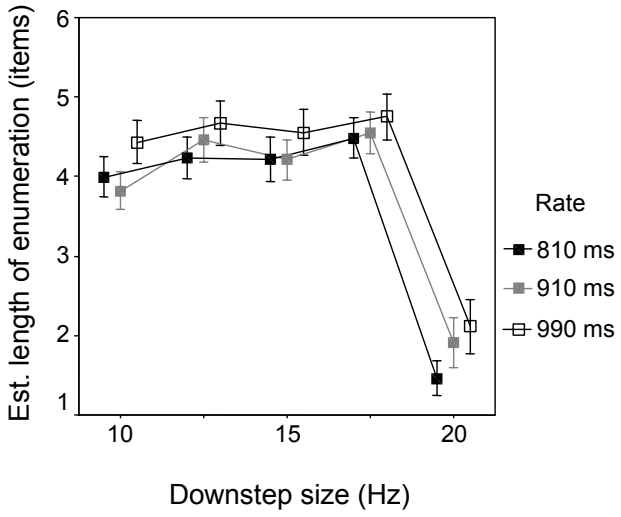


Figure 4. Estimated length of enumeration as a function of speaking rate of precursor phrase (duration in ms) and of size of downstep (Hz). Error bars represent ±2 SE.

We conclude, therefore, that neither the excursion size of the first accent-lending rise, nor the size of the first downstep provide any perceptually useful cue that allows the listener to

project the length of an enumeration. The successful prediction of a two-item enumeration for the largest downstep is not a downstep cue but would rather seem to be an experimental artefact: the pitch at the moment of truncation has dropped so low that the listener takes it as a cue for finality. The only perceptually useful element of preplanning that remains in the data, is to be found in the speaking rate of the precursor phrase to the enumeration.

4 Conclusion and discussion

Our production data confirm that pre-planning plays a role in the production of speech melody. In fact, it would seem tempting to claim that some form of 'hard' pre-planning is at stake, because the details of the phonetic specification of the first downstep in the enumeration reflect a mathematical subdivision of the available pitch span. However, this may be overstating the case. The results presented in Figure 3 have been collapsed over five male and four female speakers, and there was considerable between-speaker variability in the details of the downstep sizes. So, on aggregate, the more realistic conclusion must be that the production data do not afford a stronger conclusion than that there is 'soft' pre-planning in the production of speech melody in Dutch.

The perceptual effect of varying the size of the first downstep is a simple dichotomy. Only the difference between a fairly large downstep, indicating that the speaker is aiming for a two-item enumeration, versus a whole range of smaller downsteps, cueing a longer than a two-item enumeration, is reliably picked up by the listener. Moreover, given that the speaking rate in the onset portion of the sentence provides a more successful cue to the length of the enumeration, the later cue in the first downstep would seem to function as a confirmation rather than a refinement of the early cue. More definitive evidence on the temporal distribution of the cues would have to come from more elaborate gating experiments, similar to our studies of the perceptual cues underlying the difference between statement and (declarative) question in Dutch (van Heuven & Haan 2002).

Acknowledgement

The production and perception experiments reported here were run by my students Bart van Bezooijen and Mathieu Fannée, respectively, as a course requirement for the Fall 2003 edition of the Seminar in Experimental Phonetics taught as part of the Undergraduate Linguistics Program at Universiteit Leiden.

References

Berg, R. van den, Gussenhoven, C., & Rietveld, T. (1992). Downstep in Dutch: implications for a model. In G.J. Docherty & D.R. Ladd (Eds.), *Papers in Laboratory Phonology: Gesture, Segment, Prosody* (pp. 335-359). Cambridge: Cambridge University Press.

Boersma, P., & Weenink, D. (1996). *Praat, a system for doing phonetics by computer*. Report nr. 132. Amsterdam: Institute of Phonetic Sciences, University of Amsterdam.

Grosjean, F. (1983). How long is the sentence? Prediction and prosody in the on-line processing of language. *Linguistics, 21*, 501-529.

Gussenhoven, C., Rietveld, T. & Terken, J. (1999). *ToDI: Transcription of Dutch Intonation*. Available: http://www.lands.let.kun.nl/todi .

Haan, J., & Heuven, V. J. van (1999). Male vs. female pitch range in Dutch questions. In J.J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A.C. Bailey (Eds.), *Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, August 1-7, 1999* (pp. 1581-1584).

Hart, J. 't, Collier, R. & Cohen, A. (1990). *A perceptual study of intonation: An experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press. Cambridge Studies in Speech Science and Communication.

Heuven, V. J. van, & Haan, J. (2002). Temporal development of interrogativity cues in Dutch. In C. Gussenhoven & N. Warner (Eds.), *Papers in Laboratory Phonology VII* (pp. 61-86). Berlin: Mouton de Gruyter.

Heuven, V. J. van, & Kirsner, R. S. (1999). Interaction of grammatical form and intonation: Two experiments on Dutch imperatives. In R. Kager & R. van Bezooijen (Eds.), *Linguistics in the Netherlands 1999* (pp. 81-96). Amsterdam: John Benjamins.

Keijsper, C. E. (1984). Form and meaning of Dutch pitch contours [in Dutch]. *Forum der Letteren, 25*, 20-37, 113-126.

Leroy, L. (1984). The psychological reality of fundamental frequency declination. *Antwerp Papers in Linguistics*, *40*.

Liberman, M. & Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff & R. Oerhle (Eds.), *Language sound structure* (pp.157-233). Cambridge, MA: MIT Press.

Nooteboom, S. G. (1997). The prosody of speech: melody and rhythm. In W.J. Hardcastle & J. Laver (Eds.), *The handbook of phonetic sciences* (pp. 640-673). Oxford: Blackwell.

Nooteboom, S. G., & Cohen, A. (1974). Anticipation in speech production and its implication for perception. In A. Cohen & S.G. Nooteboom (Eds.), *Structure and process in speech perception* (pp. 124-142). Berlin: Springer. Communication and Cybernetics; 11.

Nooteboom, S. G., & Cohen, A. (1984). *Spreken en Verstaan: Een nieuwe inleiding tot de experimentele fonetiek* (2nd, revised ed., in Dutch). Assen: Van Gorcum.

Prieto, P., Shih, C. & Nibert, H. (1996). Pitch downtrend in Spanish. *Journal of Phonetics, 24*, 445-473.

(De-)accenting and discourse structure

Heleen Hoekstra

Utrecht University

Abstract

The research reported on in this paper is part of a larger study on sentence accent in (particularly) Dutch, in which a rather unconventional approach is followed. Instead of regarding accents chiefly as carriers of pragmatic information, the hypothesis is that it is not so much accent that carries information, but rather lack of accent in a position where the syntactic structure of the utterance 'predicts' one. The present paper discusses the question: given a definite NP that is a candidate for accent, in view of the syntactic structure, under what conditions is that NP left unaccented? The main conclusion is that de-accenting of definite NPs is determined by discourse grammar, which leads to the more general hypothesis that sentence accent in Dutch (and English) is essentially a matter of syntax: accenting is related to sentence syntax, de-accenting to discourse syntax.

1 Introduction

The research reported on in this paper is part of a larger study on sentence accent in (particularly) Dutch. Current theories on sentence accent share the view that accents are meaningful, in the sense that they indicate 'focus' (passim), 'activation' (Lambrecht, 1994), 'linking' (Vallduví, 1993), etc. A drawback of all these theories is that some accents are 'explained', but others are left unaccounted for. This problem, if acknowledged at all, is usually reasoned away by disposing of these accents as somehow 'secondary' or 'minor'. An example of such a 'secondary' or 'minor' accent is the one on *terribly* in the following example from Lambrecht (1994: 110-111):

> I heard something TERRIBLE last night. Remember MARK, the guy we went HIKING with, who's GAY? His LOVER just died of AIDS. […] Mark is *terribly* UPSET [my italics, HH].

The accent on *terribly* in the last sentence is not acknowledged, let alone accounted for, no doubt because it can neither count as a 'focal', nor as an 'activation' accent. It cannot be missed, though, any more than the one on *upset* can. I consider this an insurmountable drawback of this approach.

This is why I decided to approach the problem from the other end: instead of regarding accents chiefly as carriers of pragmatic information and, consequently, trying to unravel which accent contributes what to the interpretation of an utterance, the hypothesis is investigated that it is not so much accent that carries information, but rather lack of accent in a position where the syntactic structure of the utterance 'predicts' one. Stating the problem this way raises the following questions:

1. Where does the syntactic structure of an utterance 'predict' accents?
2. Given a constituent that is a candidate for accent, in view of the syntactic structure of the utterance it occurs in: under what conditions is that constituent left unaccented?

A tentative answer to the first question was given in Hoekstra (2000), in the form of an algorithm for the assignment of Sentence Accents in Dutch (SAiD). The present paper discusses

part of the second question, namely: how can we account for the well-formedness judgements as to when one can and cannot use an unaccented pronoun (or definite NP in general) to refer to some antecedent in a certain context?

2 Some examples

(1) (i) Phoebe is gisteravond naar een concert geweest.
'Phoebe went to a concert last night.'
(ii) PJ Harvey trad op.
'PJ Harvey played.'
(iii) Die is hier op tournee op het moment.
'She is on tour here at the moment.'

(2) (i) Phoebe is gisteravond naar een concert geweest.
'Phoebe went to a concert last night.'
(ii) PJ Harvey trad op.
'PJ Harvey played.'
(iii) Sindsdien neuriet ze aan een stuk door *Oh my lover*.
'She has been humming *Oh my lover* ever since.'

In the Dutch version of discourse (1), the pronoun *die* (meaning as much as 'the latter') can only refer to PJ Harvey. In (2), the pronoun *ze* ('she') can either refer to Phoebe or to PJ Harvey, though there is a slight preference for *Phoebe* as the antecedent, even apart from world knowledge, because *die* could have been used in case *PJ Harvey* had been the intended antecedent. As to the English equivalents: both in (1) and in (2) the pronoun *she* can either refer to Phoebe or to PJ Harvey. World knowledge makes us prefer the latter interpretation in (1) and the former in (2).

(3) (i) Chandler ging naar de film
'Chandler went to the movies'
(ii) en Monica naar een feest.
'and Monica to a party.'
(iii) Na afloop ging-ie een kop koffie halen.
'Afterwards, he went for a coffee.'

In (3), unlike in (1-2), there is only one possible antecedent for *ie/he*, namely *Chandler*, and yet, the discourse is not well-formed, as long as the pronoun is pronounced without accent.

Discourses (4-5) both start telling about Ross, Rachel and Phoebe, and then 'zoom in' on Phoebe.

(4) (i) Ross ging naar een lezing
'Ross went to a lecture'
(ii) en Rachel naar een feest.
'and Rachel to a party.'
(iii) Phoebe bleef thuis.
'Phoebe stayed at home.'
(iv) Ze/Die was met een liedje bezig
'She was working on a song'
(v) en bovendien moest er iemand oppassen.
'and besides, someone had to baby-sit.'
(vi) Toen Rachel tegen de ochtend thuiskwam,
'When by dawn Rachel came home,'
trof ze d'r slapend op de bank.

'she found her sleeping on the couch.'

(5) (i) Ross ging naar een lezing
'Ross went to a lecture'
(ii) en Rachel naar een feest.
'and Rachel to a party.'
(iii) Phoebe bleef thuis.
'Phoebe stayed at home.'
(iv) Ze/Die was met een liedje bezig
'She was working on a song'
(v) en bovendien moest er iemand oppassen.
'and besides, someone had to baby-sit.'
(vi) Ross en Rachel zetten de bloemetjes flink buiten.
'Ross and Rachel lived it up thoroughly.'
(vii) Toen Rachel tegen de ochtend thuiskwam,
'When by dawn Rachel came home,'
trof ze d'r slapend op de bank.
'she found her sleeping on the couch.'

The difference between (4) and (5) is that in (5) the 'substory' on Phoebe is interrupted by a statement about Ross and Rachel (5,vi), whereas in (4) it is not. The effect of the interruption is that (5), as opposed to (4), is ill-formed, in the sense that *d'r/her* in the last sentence cannot be used to refer to Phoebe. This is astonishing, as the context is such that *Phoebe* is the only possible antecedent for the pronoun in question. For the only other female person mentioned in the story is Rachel, and Rachel is not a possible referent for *d'r/her* in (5,vii) on syntactic grounds: the use of the non-reflexive pronoun enforces disjoint reference of *ze/she* and *d'r/her*, so if *ze/she* refers to Rachel, *d'r/her* cannot refer to Rachel and hence must refer to Phoebe.

Discourses (6-7) tell a little story about Chandler, containing some digressions involving two other people.

(6) (i) Chandler wou naar de film.
'Chandler wanted to go to the movies.'
(ii) Monica kon niet mee,
'Monica couldn't join him,'
(iii) want die moest nog schoonmaken.
'as she had to clean up.'
(iv) Joey kon ook niet:
'Joey couldn't come either:'
(v) die had een 'hot date'.
'he had a hot date.'
(vi) Dus ging-ie uiteindelijk maar alleen.
'So he ended up going alone.'

(7) (i) Chandler wou naar de film.
'Chandler wanted to go to the movies.'
(ii) Monica kon niet mee:
'Monica couldn't come:'
(iii) die had al een andere afspraak.
'she had another date.'
(iv) Dus ging-ie maar met Joey.
'So he went with Joey.'

(v) Toen-ie thuiskwam, stond ze op z'n antwoordapparaat
'When he came home, she was on his answering machine,'
om te vragen hoe de film geweest was.
'asking how the movie had been.'

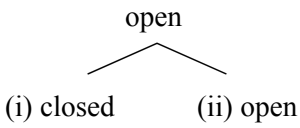
In neither of the two stories do we have a problem interpreting *ie/he* in (6,vi) and (7,iv), respectively, as referring to Chandler, in spite of the fact that in (6), (iv-v) were about Joey. Discourse (7), however, is ill-formed: in spite of the fact that Monica is the only female occurring in the story, one cannot use *ze/she* in (7,v) to refer to her.

What these examples show is that in some contexts, an unaccented pronoun can be used to refer to a certain antecedent, even though there is another candidate in between, whereas in other contexts, an unaccented pronoun can*not* be used to refer to a certain antecedent, even though it is both quite near and the only possible candidate.

3 Unaccented pronouns and discourse structure

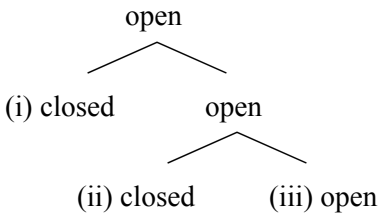
In the discourse grammar approach of Scha & Polanyi (1988), a discourse is parsed incrementally, from left to right, one clause at a time. At any point in the parsing process, only the right peripheral nodes (or vertices) in the parse tree are open for expansion, i.e. to forming new (sub)structures. For example, the parse trees representing the first two clauses (i-ii) of discourses (1-3) look essentially the same:

(1-3)



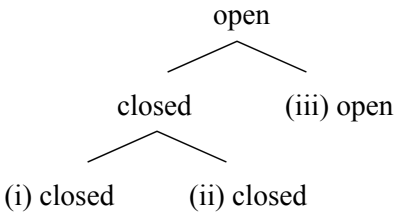
In (1), clause (iii) is a continuation of (ii) and hence is attached at (ii). The resulting structure is:

(1)



In (2) and (3), clause (iii) resumes clause (i) and hence is attached at the top node, resulting in:

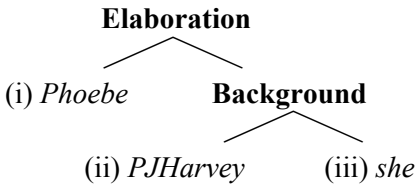
(2-3)



What do these representations reveal about the (im)possibilities of anaphora by means of an unaccented pronoun? Polanyi (1988: 616) states: "Pronominalization will not be permitted to elements in closed off constituents (…)." In (1), (ii) is an open node when (iii) is attached, so the rule permits a pronoun in (iii) referring to an antecedent in (ii). In (2) and (3), (i) has been closed off by the time (iii) is attached, so the rule forbids a pronoun in (iii) referring to an antecedent in (i). In other words, Polanyi's rule explains the acceptability of (1), as well as the unacceptability of (3), but not the acceptability of (2). This is where rhetorical relations come

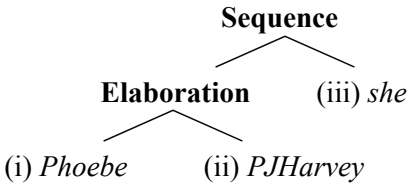
in. Scha & Polanyi (1988)[1] argue that subordinate and co-ordinate relations be treated differently: subordinations are binary structures in which one DCU (discourse constituent unit) is subordinate to the other and in which most of the relevant features (reference time, modal index, etc.) are inherited from the subordinating constituent; co-ordinations, on the other hand, are *n*-ary structures ($n \geq 2$) in which all elements have equal status. Among the information to be inherited from the subordinating constituent in a subordination is also a set of discourse referents, they claim. Let us consider the consequences of this assumption. (Of the rhetorical relations occurring in our examples, *Enumeration* and *Sequence* are co-ordinations and all the others subordinations.)[2]

(1)



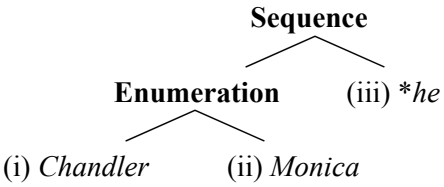
As to (1), the inheritance of discourse referents has no influence, as the node containing the antecedent *PJ Harvey* (ii) is itself an open node at the time the node containing the pronoun *she* (iii) is attached.

(2)



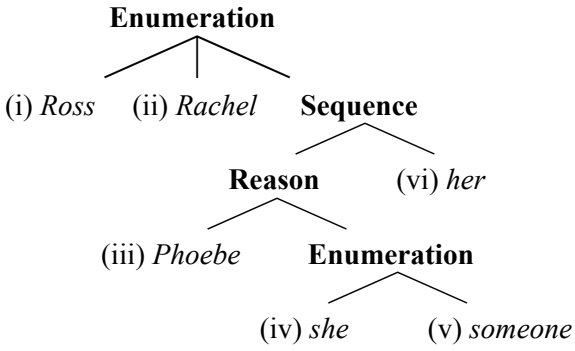
In (2), the Elaboration node inherits the discourse referent *Phoebe* from the subordinating DCU (i), hence *Phoebe* can now function as an antecedent for *ze/she* in (iii), as the Elaboration node is an open node at the time (iii) is attached.

(3)



In (3), no inheritance takes place (Enumeration being a co-ordination), hence the node containing the only possible antecedent for *ie/he* (i) has been closed off by the time (iii) should be attached, whence the ill-formedness of the discourse.

(4)

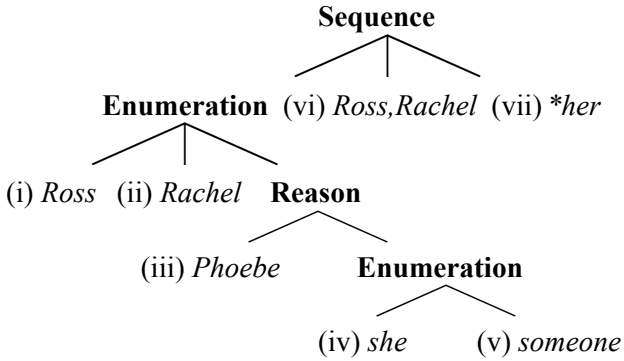


[1] A similar approach is taken in Asher (1993).

[2] The names of the rhetorical relations are taken from Hitzeman, Moens & Grover (1995).

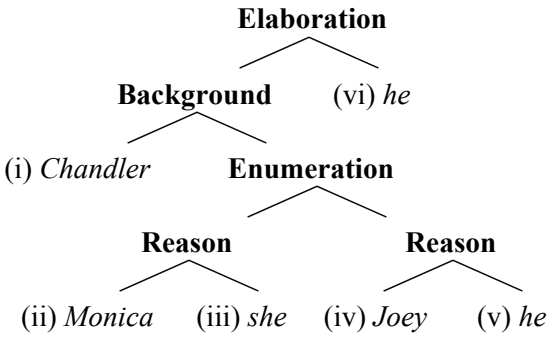
In (4), *Phoebe* in (iii) is inherited by the Reason node, and can now function as an antecedent for *d'r/her* in (vi).

(5)



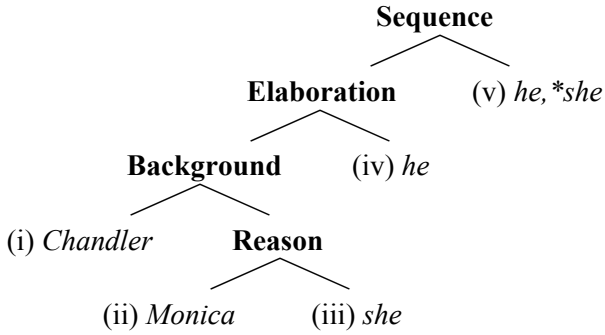
In (5), no discourse referents are inherited by the upper Enumeration node, but apart from that, (vii) could not be attached anyway, as the Enumeration node was closed off by the attachment of (vi).

(6)



In (6), the Background node inherits *Chandler* from (i), hence (vi) can be attached.

(7)



In (7), (v) cannot be attached: the Elaboration node inherits *Chandler* from (i), via the Background node, which is the subordinating constituent, so the *ie/he* is not a problem, but *Monica*, though she can pass the Reason node, cannot pass the Background node, as the Reason node is the subordinated node, hence the *ze/she* in (v) cannot be parsed.

The conclusion is that the well-formed discourses (1,2, 4, 6) are accepted and the ill-formed ones (3,5,7) are rejected.

As to the Dutch personal/anaphoric (as opposed to demonstrative/deictic) pronoun *die*: Polanyi's statement that "Pronominalization will not be permitted to elements in closed off constituents (…)", which we found to be too strict for pronouns in general, does in fact apply to anaphorically used *die*. In other words: anaphoric *die* can only be used if its antecedent is available directly, as opposed to as a result of inheritance.

4 Accented (and unaccented plural) pronouns

Unaccented pronouns only occur if plain co-reference is at stake. Accented pronouns, on the other hand, indicate, e.g., subsectional anaphora (Van Deemter, 1991).[3]

If we replace the ungrammatical unaccented pronouns in the ill-formed discourses (3,5,7) from sections 2-3 by accented ones, we get the following effects on the well-formedness judgements with respect to these discourses: (3) becomes well-formed, (5) and (7) remain ill-formed.

In (3), the DCUs containing *Chandler* and *Monica*, respectively, form an Enumeration node. Hence, *Chandler* cannot be referred to by means of an unaccented pronoun. It is possible, however, to refer to Chandler and Monica as a couple by the unaccented pronoun *they*, which can be explained by assuming a 'summation' operation like the one proposed by Kamp & Reyle (1993). Apparently, co-ordinations, though not capable of inheriting discourse referents from their members, can construct discourse referents of their own by e.g. summation, which can then serve both as antecedents for unaccented plural pronouns in a co-reference relation and – is my claim – as context sets (Westerståhl, 1985) for subsectional anaphors, like an accented pronoun: as Chandler is the only male in the context set consisting of Chandler and Monica, the accented pronoun *HIJ/HE* can indeed be used to refer to him.

What about (5) and (7): why do they not become well-formed if we replace the unaccented pronouns *d'r/her* in line (5,vii) and *ze/she* in line (7,v), respectively, by their accented counterparts *HAAR/HER* and *ZIJ/SHE*, respectively?

At the time (5,v) has been uttered, a discourse referent representing the set consisting of Ross, Rachel, and Phoebe can be made available at the Enumeration node. Accordingly, it is indeed possible at this point in the discourse to use the unaccented pronoun *they* to refer to the three of them. Instead of *Ross and Rachel lived it up thoroughly* (5,vi) one could have said something like *Four hours later they were sitting on the couch, telling each other about their evenings*. (Note that *they* can only refer to the threesome here, not to any subset of them.) With Kamp & Reyle (1993), I take it that such a discourse referent is only constructed if the occurrence of a plural pronoun – or a subsectional anaphor, I claim – requires one. By the time (5, vii) is uttered, however, the required context set can no longer be constructed, as the only discourse referents available at that point are Ross and Rachel (from (5,vi)), and hence even an accented pronoun cannot be used to refer to Phoebe. But even if the required set (one including Phoebe) could be made available, using a pronoun would be improper, because the interpretation of the pronoun *haar/her* requires that its referent (i.c. Phoebe) be the only female in the context set or, more precisely, that it constitute the intersection of the context set and the set of females in the model (Hoekstra, 1993: 57-58), a requirement that would not be met, as the intersection would consist of Phoebe and Rachel.

In (7) as well, the required context set, i.e. one containing Monica, can no longer be constructed by the time line (7,v) is uttered. If the required set had been available, however, the accented pronoun *ZIJ/SHE* could have been used to refer to Monica, as she would indeed have been the only female in the context set.

Let me summarize the possibilities and impossibilities of context set construction in discourse grammar as proposed in this section. At any node, be it a subordination or a co-ordination, a

---

[3] The term *subsectional anaphora* is used for a type of anaphora where the referent of the anaphor is interpreted as a subset of the referent of its antecedent, as in: *I saw a couple of children playing in the street. Two boys were kicking a can*. In this discourse, *two boys* can be analysed as anaphoric to *a couple of children*, in the sense that the set of two boys kicking a can is interpreted as a subset of the set of children playing in the street.

context set can be constructed out of the discourse referents of all (and only all) its daughter DCUs, i.e. subordinating and subordinated, or co-ordinated ones. However, this can only be done if the occurrence of a pronoun in the next utterance requires one. It is thus not allowed to assume the construction of a context set at a node A that is the subordinating DCU of a subordination B and to have B inherit this set from A in order to provide an antecedent for a pronoun in an utterance that occurs after B has been closed off.

We have seen what the effects are of replacing the ungrammatical unaccented pronouns in the ill-formed discourses (3,5,7) by their accented counterparts. What remains to be considered is what happens if we do the same with the unaccented pronouns in the well-formed discourses (1,2,4,6), that is, if we replace the (properly used) unaccented pronouns *die/she*, *ze/she*, *d'r/her*, and *ie/he* in (1,iii), (2,iii), (4,vi), and (6,vi), respectively, by their accented counterparts *DIE/SHE*, *ZIJ/SHE*, *HAAR/HER*, and *HIJ/HE*, respectively. All four resulting discourses sound weird when uttered out of further context. For instance, using *HAAR/HER* instead of *d'r/her* in (4,vi) suggests that Phoebe's sleeping on the couch is somehow contrasted with some other activity by someone else. It would be acceptable if, for example, discourse (4) would proceed: *whereas Emma* [the kid that Phoebe was supposed to look after] *was wide awake and playing with Hugsy*. In the extended discourse, the accented pronoun in fact functions as a subsectional anaphor, in much the same way as *HIJ/HE* in (3,iii) does.

Note the parallel between the behaviour of accented versus unaccented pronouns at discourse level on the one hand, and personal versus reflexive pronouns at sentence level on the other: roughly speaking, at sentence level, the accessibility constraints for reflexives are stricter than those for personal pronouns, but if the requirements for the use of a reflexive are met, then the use of a personal pronoun with the same antecedent is excluded; in much the same way, at discourse level, the constraints for unaccented use of a pronoun are stricter than those for accented use, but if the requirements for unaccented use are met, then accented use is excluded.

The findings from this and the previous section on pronouns can be summarized as follows: a (personal) pronoun, whether singular or plural, accented or unaccented, requires that its antecedent be available (either directly or as a result of inheritance or construction) at the node where the utterance containing the pronoun should be attached.[4] Once the antecedent is established, an unaccented pronoun requires plain co-reference, an accented pronoun calls for some linking operation to be carried out (subsectional relativization, bridging, …).[5]

5 Proper names

The next question to be discussed is: what are the effects of replacing the pronouns in (1-7) by, for example, proper names? Proper names are said to uniquely identify their referents, so it does not come as a surprise that replacing the unaccented pronouns *die/she* and *ze/she* in (1,iii) and (2,iii) by the de-accented proper names *PJ Harvey* and *Phoebe*, respectively, takes away the referential ambiguity from these examples, and that the same goes, mutatis mutandis, for the other well-formed discourses, (4) and (6). What *is* surprising, though, is that replacing the ungrammatical unaccented pronouns *ie/he*, *d'r/her*, and *ze/she* in the ill-formed discourses (3,5,7) by the de-accented proper names *Chandler*, *Phoebe*, and *Monica*, respectively, hardly seems to render these discourses any less unacceptable.

---

[4] For some pronouns, e.g. anaphorically used Dutch *die*, stricter conditions may apply (see section 3).

[5] The term *bridging* is used for a type of anaphora where the antecedent evokes a frame within which the anaphor finds its interpretation and reference. Examples are: *a car – the wheels*, *a book – the writer*, *a restaurant – the waiter*.

Replacing the unaccented pronouns in the well-formed discourses (1,2,4,6) by the corresponding *accented* proper names has the following effect: when uttered without further context, the discourses become weird with the unaccented pronouns replaced by the corresponding accented proper names, just as weird as with the unaccented pronouns replaced by their accented counterparts (see section 4). As for the ill-formed discourse (3), replacing the unaccented pronoun *ie/he* by the accented proper name *CHANDLER* renders the discourse well-formed, just like replacing the unaccented pronoun by its accented counterpart *HIJ/HE* did (see section 4). The ill-formed discourses (5,7) did not become well-formed by replacing the ungrammatical unaccented pronouns *d'r/her* and *ze/she* by their accented counterparts *HAAR/HER* and *ZIJ/SHE*, respectively. Replacing the pronouns by the corresponding accented proper names does render the discourses well-formed. Apparently, the use of accented proper names is not subject to the accessibility constraints that the use of (accented) pronouns is. Note, however, that it is hard to determine whether these proper names are used anaphorically or for (re-)introducing a discourse referent.

In summary, de-accented proper names must satisfy the same conditions as unaccented pronouns: the antecedent must be available (either directly or as a result of inheritance or construction) at the node where the utterance containing the proper name should be attached, and the relation between anaphor and antecedent must be one of plain co-reference. They only differ from pronouns in that they leave less room for referential ambiguity. If accented, proper names can be used much more freely than pronouns: they can be used to introduce new elements into the discourse, unlike pronouns.[6] Also, when used anaphorically, they do not seem to be subject to the accessibility constraints that pronouns are (though it is not always easy – or possible even – to decide whether an accented proper name is really used anaphorically, rather than for (re-)introduction). The only circumstance in which the use of an accented proper name is excluded seems to be when the accented proper name is used anaphorically and its relation with its antecedent is one of plain co-reference. Not many linguists will be surprised by the fact that pronouns on the one hand and proper names on the other show different behaviour. What *is* surprising is that the accessibility constraints for pronouns and proper names appear to be the same when they are left unaccented.

6 Conclusion

First, we have to assume that, within the discourse grammar approach, subordinations inherit a set of discourse referents from their subordinating DCU (in the spirit of Scha & Polanyi, 1988), and that at any node a context set can be constructed out of the discourse referents of all (and only all) its daughter DCUs, provided that the occurrence of a pronoun in the next utterance requires one (in the spirit of Kamp & Reyle, 1993). The accessibility conditions for pronouns in discourse can then be formulated as follows: a pronoun, whether singular or plural, accented or unaccented, requires that its antecedent be available (either directly or as a result of inheritance or construction) at the node where the utterance containing the pronoun should be attached. Once the antecedent is established, an unaccented pronoun requires plain co-reference, whereas an accented pronoun calls for some linking operation to be carried out (e.g. subsectional relativization, bridging). As for proper names: de-accented proper names must satisfy the same accessibility conditions as unaccented pronouns, whereas accented proper names are allowed in all contexts where de-accented ones are not.

---

[6] A counterexample to the claim that pronouns cannot be used for introduction is a stylistic trick of the type *And then SHE entered!*, where *she* is intended to refer to 'the woman of my dreams', or something of the sort.

The general conclusion is that de-accenting of both pronouns and proper names – of definite NPs in general, actually – is solely determined by discourse structure. In fact, as will be argued in Hoekstra (forthcoming), this holds for *all* NPs.

This conclusion in turn yields a more general hypothesis to investigate, viz. that de-accenting *in general* is determined by discourse structure, in much the same way as accenting is determined by sentence structure (Hoekstra, 2000). In other words, my hypothesis is that sentence accent in Dutch (and English) is essentially a matter of syntax: accenting is related to sentence syntax, de-accenting to discourse syntax.

References

Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Studies in Linguistics; 50. Dordrecht: Kluwer Academic.

Deemter, K. van (1991). *On the Composition of Meaning. Four Variations on the Theme of Compositionality in Natural Language Processing*. PhD thesis, University of Amsterdam.

Hitzeman, J., M. Moens, & C. Grover (1995). Algorithms for analysing the temporal structure of discourse. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 253-260). Dublin.

Hoekstra, H. (1993). Subsectional anaphora in DRT. In M. Everaert, B. Schouten, & W. Zonneveld (Eds.), *OTS Yearbook 1992* (pp. 53-62). Utrecht: OTS.

Hoekstra, H. (2000). An algorithm for the assignment of sentence accents in Dutch. In H. de Hoop, & T. van der Wouden (Eds.), *Linguistics in the Netherlands 2000* (pp. 105-118). (AVT Publications; 17). Amsterdam: John Benjamins.

Hoekstra, H. (forthcoming). *Sentence Accent without Focus: Evidence from Dutch and English* (provisional title). PhD thesis, Utrecht University.

Kamp, H., & U. Reyle (1993). *From Discourse to Logic*. Dordrecht: Kluwer.

Lambrecht, K. (1994). *Information Structure and Sentence Form. Topic, Focus and the Mental Representations of Discourse Referents*. Cambridge Studies in Linguistics; 71. Cambridge: Cambridge University Press.

Polanyi, L. (1988). A formal model of the structure of discourse. *Journal of Pragmatics, 12*, 601-638.

Scha, R., & L. Polanyi (1988). An augmented contextfree grammar for discourse. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING)* (pp. 22-27).

Vallduví, E. (1993). *Information Packaging: A Survey*. Research paper, Human Communication Research Centre, University of Edinburgh.

Westerståhl, D. (1985). Determiners and context sets. In J. van Benthem, & A. ter Meulen (Eds.), *Generalized Quantifiers in Natural Language* (pp. 45-71). Dordrecht: Foris.

On measuring multiple lexical activation using the cross-modal semantic priming technique

Esther Janse and Hugo Quené

Utrecht University

Abstract

Cross-modal semantic priming with partial auditory primes seems a good technique to assess spoken-word recognition, because it allows tracking the activation of multiple word candidates. However, previous research using this technique has found inconsistent results. First, a priming experiment is reported that addresses this technique's validity. Results show that semantic priming is not observed with partial auditory primes, but only with full primes. Secondly, Monte Carlo simulations are reported of a previous study that found partial priming effects; the simulations show that the particular design in that study yields a high risk of a Type-I error. In conclusion, the semantic priming technique cannot be used to investigate activation of multiple word candidates, and its use for that purpose should be discontinued.

1 Introduction

Among current models of spoken-word recognition, there is remarkable agreement about one basic stage or process, namely that of initial multiple activation of word candidates in the listeners' mental lexicon that roughly match the available auditory input. In a later stage, activated candidates compete and each candidate's activation increases or decreases when more auditory input becomes available, or when semantic context starts to influence the selection (Marslen-Wilson & Tyler, 1980; McClelland & Elman, 1986; Norris, 1994; McQueen & Cutler, 2000). Evidence for multiple activation of word candidates comes from several studies that employed the cross-modal semantic priming paradigm. This paradigm in its most common form was first introduced by Swinney (1979). The technique is based on spreading of activation from one lexical element to other semantically or associatively related elements (Collins & Loftus, 1975). Semantically related items in the mental lexicon are interconnected via facilitating links: an increase in the activation of one item leads to an automatic increase in the activation of related items.

If listeners are required to make a lexical decision on a visually presented target word (e.g., MONEY) after hearing an auditory prime word (e.g., *salary*), then they react faster if the visual target word and the auditory prime word are semantically or associatively related, than if these words are not related. In *partial priming*, the auditory prime words are cut off before their acoustic offset (e.g., *sala*-…), at a point where the acoustic information is not sufficient to identify the intended word uniquely. At that cut-off point, multiple word candidates are supposed to be still active. By presenting a visual target related to one of these candidates, immediately following the prime fragment's offset, one can measure the activation of that word candidate. This implies that, even before spoken words are completely recognised, they have already sent a detectable amount of activation to their semantic associates.

Zwitserlood (1989) used the cross-modal semantic priming task to investigate the activation of multiple word candidates, using partial priming. Her study was set up to investigate during which stage context affects the activation of lexical candidates. The various models of

auditory word recognition make different predictions with respect to the relative weights of sentence context and bottom-up acoustic information during word processing (cf. Forster, 1976; 1979; McClelland & Elman, 1986). The results of the Zwitserlood (1989) study showed two important things. First, multiple lexical candidates are accessed on the basis of partial auditory information: even when only a fragment of an auditory prime word is presented (e.g., *sala-*…), activation was found for both compatible word candidates *salaris* ('salary') and *salami* ('salami')[1]. Second, sentence context affects the activation of lexical candidates, thus providing evidence for a hybrid model of word recognition. However, after the Zwitserlood (1989) study, semantic priming studies have yielded inconsistent results, if any. The effects are small and inconsistent, especially in sentence context (Gaskell & Marslen-Wilson, 1996; Jongenburger, 1996). Chwilla (1996) found no partial priming effects, even though part of her material was identical to that of Zwitserlood (1989). Zwitserlood & Schriefers (1995) found that a short prime fragment only yielded a priming effect when extra processing time was available (between prime fragment offset and presentation of the visual target). Hence the "selection" stage of the recognition process may have been already concluded by the time the listener responded. Using auditory stimuli that were phonetically ambiguous with respect to the voicing value of the initial consonant (e.g., between *dip* and *tip)*, Connine, Blasko, & Wang (1994) observed multiple activation effects before the isolation point. The reported lexical decision times are relatively long, however, which again raises questions about the on-line nature of these effects. Moss, McCormick, & Tyler (1997) reported only a weak semantic priming effect of 10 ms at the isolation point. Note that in two of these studies (Zwitserlood & Schriefers, 1995; Moss, McCormick, & Tyler, 1997), activation was only measured for the actual prime word, and not for other candidates competing with the actual prime word. Consequently, these studies bear only little evidence on *multiple* activation.

So, whereas robust semantic priming effects have been reported with full primes (Meyer & Schvaneveldt, 1971; Neely, 1977; Swinney, 1979), it is important to know whether this technique also gives reliable and robust results when *partial* primes are presented. Importantly, the Distributed Cohort Model (Gaskell & Marslen-Wilson, 1997; 1999; 2002) provides an explanation why semantic priming effects obtained with partial primes may not be as robust as some earlier results suggest. In the Distributed Cohort (henceforth DC) model, the process of speech perception is modelled as a recurrent neural network. Lexical units are points in a multidimensional space, represented by vectors of phonological and semantic output nodes. The speech input maps directly and continuously onto this lexical knowledge. As more bottom-up information becomes available, the network moves towards the word under consideration. Activation of a word candidate is then inversely related to the distance between the output of the network (a point in the multidimensional space), and the word representation in this space. In connectionist models, multiple representations must interfere with each other if they are active simultaneously. This was also modelled in two older models, TRACE (McClelland & Elman, 1986) and Shortlist (Norris, 1984): lateral inhibition between activated word candidates is employed to reduce multiple activations. Before the uniqueness point, semantic activation depends strongly on the number of candidates that match the input sofar and their relative frequency.

Gaskell & Marslen-Wilson (1999) explain why the effects of phonological priming are generally much stronger than those of semantic priming if partial primes are presented. In phonological priming, the relation between the prime-target pair is such that the visual target

[1] Priming effects were found for both related visual targets, GELD ('MONEY') and WORST ('SAUSAGE').

(e.g., PORT or PORK) is fully or partially identical to the auditory prime (e.g., *por*-…) in terms of its acoustic phonetic form. In the DC model, priming for a certain candidate occurs if a lexical representation is more similar to the target representation than to an unrelated baseline. Phonologically, the word candidates are obviously coherent, but the semantic representations of the different candidates often have no meaning overlap at all. "In repetition priming, the target lexical representation is related to the prime representation in all dimensions, so recognition of te target can take advantage of overlap on both semantic and phonological nodes (…). By contrast, semantic priming relies on overlap in the semantic nodes alone" (Gaskell & Marslen-Wilson, 1999, p. 452). Empirical results (Gaskell & Marslen-Wilson, 2002) support their claim that the phonological effects of partial priming are much stronger than the semantic effects. In their experiment, primes were presented either complete, or in two cut-off conditions. Semantic priming occurred only after the moment that the prime has become unambiguous onwards. By contrast, significant phonological priming effects were found at all cut-off points. In summary, partial priming (activating multiple candidates) necessarily leads to weak activation of the candidates' semantic associates. While multiple candidates are active, their disparate semantic properties provide only weak semantic priming of other words, if any. With full priming, only a single candidate remains, and its coherent semantic properties provide strong semantic priming of its related words.

Our understanding of spoken-word perception crucially depends on the concepts of automatic spreading of activation to semantic associates and of multiple activation of word candidates, as well as the time courses of these processes. The present study therefore attempts to clarify this discrepancy among semantic priming studies, in two ways. Its first aim is to provide decisive empirical evidence about partial semantic priming. This is done by means of a modified replication of the cross-modal semantic priming study by Zwitserlood (1989), which has provided the strongest evidence in favour of partial priming. In our "heteromethod" replication (Campbell, 1969), the original design and stimulus materials were slightly modified, to improve the chances of detecting partial priming effects. If the effect of semantic priming after the presentation of partial primes is robust enough, then we should also be able to find it with a design that is different from the original study. Due to space limitation, the design and results of this replication experiment will be discussed only very briefly. The second, methodological aim of this study is to illustrate the use of Monte Carlo simulations for post-hoc evaluation of experimental designs. Such simulations provide realistic estimates of the chances of Type I and Type II errors, even for complicated repeated-measures designs. Hence, they provide a relatively easy alternative to formal power analyses. Simulation results indicate that in the original study by Zwitserlood (1989), the chance of a Type I error was far greater than the stated level of significance.

2 Experiment

Since Zwitserlood's sentences and test words have yielded the clearest partial priming effects to date, her (Dutch) materials were used here whenever possible. The main difference here is in our experimental design. The present experiment ignores the sentence context factor, such that a within-subjects design is possible. For a more detailed description of the materials, design, procedure, and for more detail on the results, the reader is referred to Janse (2003).

The results of the replication experiment can be summarised as follows: the presentation of the partial prime fragment (prime cut off at isolation point) did not yield semantic facilitation, not for the target related to the actual prime word, nor for the target related to the competitor. The presentation of the full actual prime word yielded a significant 32 ms priming effect for the prime word's related target. The reaction time data were fed into repeated measures

ANOVAs, with Relatedness (related vs. unrelated/control condition), Candidate (target related to either prime word or closest competitor), and Prime Length (partial/full) as fixed factors. With increasing auditory information, visual targets related to the intended auditory prime are predicted to show priming effects, whereas visual targets that are either unrelated or related to the prime's competitor should show no such effects. This three-way interaction was in fact not significant [$F_1(1,59)=2.8$, n.s.; $F_2(1,23)=2.8$, n.s.]. The ANOVAs were also carried out for Actual Primes and Competitors separately. The Relatedness × Prime Length interaction was significant in the sub-analysis for intended auditory primes [$F_1(1,59)=9.3$, $p=0.003$; $F_2(1,23)=9.4$, $p=0.005$], but not for the prime's competitor [$F_1(1,59)<1$; $F_2(1,23)<1$].

The absence of partial priming is an important finding in this study. But could this perhaps be due to low power in detecting such an effect? This is quite improbable, for two reasons. First, let us inspect one very robust priming effect, viz. priming of the visual target by a semantically related auditory prime that is fully audible. This priming effect is also present in our data (32 ms). Post-hoc power analyses for this separate contrast indicate that the present experiment had ample power in detecting this contrast, viz. .973 (for ANOVA by subjects) and .988 (for ANOVA by items). Second, post-hoc power analyses indicated that the relevant two-way interaction of Candidate × Prime Length (for the sub-analyses for intended auditory primes, see above) was detected with adequate power, viz. .558 (by subjects) and .644 (by items). Taken together, these analyses indicate that our amendments in design and stimulus materials by Zwitserlood (1989) have not reduced the power of our study.

3 Discussion

First, reliable semantic priming is observed in this experiment, if full auditory primes are presented. This agrees with previous research (Swinney, 1979; Chwilla, 1996), which lends credibility to the present results. Second, no priming was observed when partial primes are presented. Third, RTs were shorter for full primes than for partial primes: not only for intended auditory primes, but also for their competitors, *and* for unrelated control conditions as well. Hence, this decrease in RT as auditory information increases is not due to priming, but can be explained as an effect of lexical competition in general; responses to any target can be faster if there is less competition between lexical candidates. A similar pattern of results was found by Mattys & Clark (2002) using a pause detection task. In their study, RTs for early-unique words were shorter than for late-unique words, for which lexical competition persists longer. Hence, the absence of a Relatedness × Prime Length interaction, and of the three-way interaction effect, are in agreement with their explanation. This pattern of results suggests that there is indeed lexical competition in 'partial' conditions, even though this does not lead to semantic priming effects.

But how can we explain the discrepancy between these replication results and those of the original study? Our answer is that the original study (Zwitserlood, 1989) suffered from an incomplete between-subjects design (Cochran & Cox, 1957), which was necessary because of its mis-proportion of number of conditions (32) and number of items (24). In such a design, an interaction effect between listeners and conditions cannot be separated from the main effect of conditions (Cox, 1958, Chapter 11; Bailey, 1982). To rely on the assumption that any effect of interest would be equal across participants is in fact quite dangerous. Participants vary in their cognitive behaviour, and these differences obviously persist when they process spoken or written language (e.g. Connine, Blasko, & Wang, 1994; Plaut & Booth, 2000). Thus, it is entirely conceivable that priming effects show up in *some* listeners, perhaps the 'fast' or verbally proficient ones, but not in others. These inter-listener

differences amount to an interaction between listeners and conditions. In an incomplete between-subjects design, such interactions between listeners and conditions are pooled with the main effect of conditions, thus inflating the significance of the latter. Also note that comparisons between conditions must of necessity assume sphericity, that is, differences among conditions are assumed to have equal variances across listeners. This assumption is nowadays often regarded as dangerous and not warranted (O'Brien & Kaiser, 1985; Max & Onghena, 1999). Even though the original Zwitserlood (1989) data are no longer available for investigation, it is possible to address this design issue by means of Monte Carlo simulations.

4 Monte Carlo simulations

In a so-called Monte Carlo simulation (Hammersley & Handscomb, 1964), a data set from an imaginary experiment is generated at random, with statistical properties programmed in the simulation. The appropriate test statistic, e.g. an *F* ratio, is calculated from each simulated data set. This process is then repeated a large number of times. For example, we could generate many sets of RT data, in which there are no "true" differences between related test and unrelated control conditions (i.e. in accordance with H_0 that priming effects are absent). In realistic simulations, of course, there are also random variance components associated with items, with listeners, and perhaps with other random sources of variance.

In our simulations, each data set was generated as follows. First, the difference between auditory prime words and their competitors was ignored; all simulations were run for actually intended prime words only. This yields 2 × 2 treatment conditions, defined by the Relatedness and Prime Length factors. All four treatment effects were set to zero. Each observation was generated by adding an arbitrary grand mean, plus the treatment effect, plus several random effects. Values for these random effects were drawn from separate gaussian distributions having mean zero, and having the following standard deviations: experimental lists (or listener group) 30 ms, listeners 100 ms, items 100 ms, and within-cell observations 100 ms. These standard deviations correspond roughly to the random variance components observed in other cross-modal priming studies in our laboratory.

For each observation, the appropriate values of these random effects were used, in order to simulate the actual design of the experiment presented above. Crucially, RT observations were generated both with and without the controversial listener-by-condition interaction as an additional random effect. This random variance component was drawn from a gaussian distribution, with mean zero and *SD* 75 ms. The latter is a conservative estimate, based on the observed variance due to the corresponding interaction of Relatedness × Prime Length × Experimental List in the experiment above (*MS*=8007, *s*=89 ms). Each data set, or simulated experiment, consisted of 24 test items, and 4 groups of 15 listeners each, as in the actual experiment. For each design (with and without interaction), 5000 simulations were performed.

Each of these data sets was fed into two repeated measures univariate ANOVAs. Results of these Monte Carlo simulations take the form of 5000 F_1 ratios (by listeners) and 5000 F_2 ratios (by items) corresponding to the two-way interaction of Relatedness × Prime Length. Most relevant for our purposes is the proportion of these *F* ratios exceeding the appropriate critical *F* value. This corresponds to the probability of a 'positive outcome', i.e. of rejecting H_0, and of concluding that a priming effect exists. Since a priming effect is known to be absent, rejecting H_0 amounts to a Type I error here. Figures 1 and 2 give the distributions of F_1 and F_2 of this interaction, respectively, along with this probability of a positive outcome. The Monte Carlo results for F_1 and F_2 show similar tendencies, and will be discussed together.

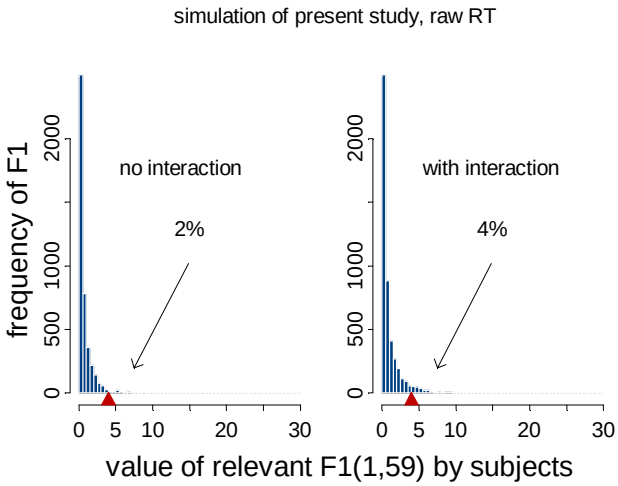


Figure 1. Distributions of 5000 F_1 ratios. Simulations were done with a listener × condition interaction effect either absent (left) or present (right) in the data sets. The critical F_1 value (α=.05) is marked along the abscissa, and the percentage of 'positive outcomes' is given with each distribution.

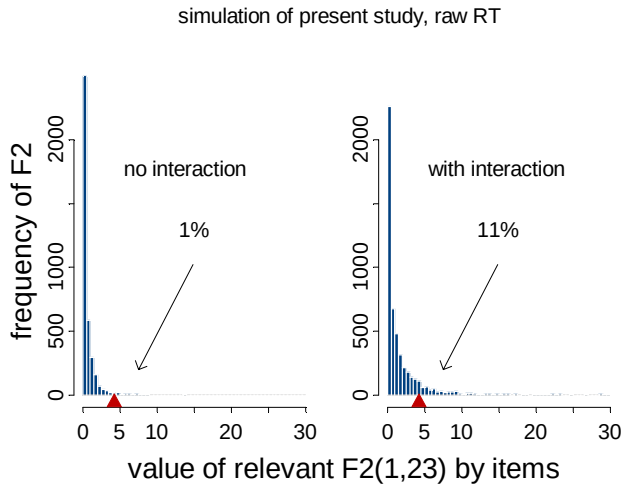


Figure 2. Distributions of 5000 F_2 ratios.

First, if an interaction between listeners and conditions is present, then the chance of a Type I error is inflated somewhat (right) relative to the no-interaction case (left), even in this within-subject design. Note that there is a very low probability of Type I error, i.e. of finding spuriously significant effects, in case the priming effect is indeed absent.

Next, let us compare these findings with Monte Carlo simulations of the original experiment by Zwitserlood (1989). To this end, we changed the experimental design in the simulations from a complete within-subjects design to an incomplete between-subject design. In the original study, each listener participated in 24 out of 32 conditions. In our simulation of that study, each listener participated in 3 out of 4 conditions, corresponding to 3 out of 4 treatment conditions as defined above. Listeners and items were rotated evenly across conditions. Four listener groups were used, each consisting of 6 listeners (cf. Zwitserlood, 1989). Again, the Candidate factor (actual prime vs. competitor) was ignored for practical purposes. Further

details of these Monte Carlo simulations were identical to those above. The incomplete between-subjects design precludes any ANOVA by listeners, because a single listener only participated in 3 out of 4 treatment conditions. This Monte Carlo simulation therefore only yields F_2 ratios, which are given in Figure 3 below.

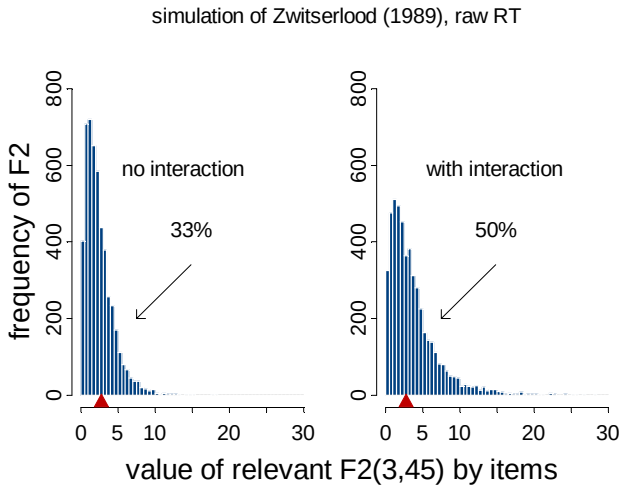


Figure 3. Distributions of 5000 F_2 ratios.

First, a Type I error is indeed highly probable, if an interaction component between listeners and conditions is forced to be present in the data (right panel). If listeners vary in their susceptibility to priming, this would indeed increase the chance of finding a 'significant' effect. The priming effect, although absent in the data, is incorrectly reported to be 'significant' in about half of the simulations. Second, even without this disputed interaction (left panel), the chance of finding a spurious significant effect, a Type I error, is dangerously high. This itself indicates that the reported effect may well have been spurious.

Zwitserlood (1989) also realised that listeners and conditions were confounded in her design. Differences among test conditions are 'contaminated' by differences among listeners' averages. A normalisation procedure removes this contamination part, but analysis of the resulting data still requires the assumption that there is no interaction between conditions and listeners. The listener's average RT was subtracted from each observation (and then the grand mean was added). The resulting normalised RTs are still contributed by different listeners in different conditions. Thus, individual priming differences are still likely to exist.

In order to investigate the effect of this normalisation, the Monte Carlo simulations of the original experiment were repeated, with Zwitserlood's normalisation procedure inserted after random generation of the data sets, before statistical analysis. In all other respects, the simulations were equal to those on the raw RT data. The resulting F_2 ratios are summarised in Figure 4.

First, we see that the presence of interaction between listeners and conditions in the normalised data inflates the probability of a Type I error to 62% (right panel). This is the probability of reporting a significant priming effect, if in fact such an effect is forced to be absent, and if individual differences in priming susceptibility are forced to vary among listeners. Obviously, this high probability of a Type I error raises strong doubts about the validity of the experiment that is simulated here. As before, even without this disputed interaction component in the data (left panel), the chance of finding a spurious significant effect, a Type I error, is apparently inflated by the normalisation procedure. In summary,

these Monte Carlo simulations suggest that the original experiment by Zwitserlood (1989) does not warrant valid and reliable conclusions.

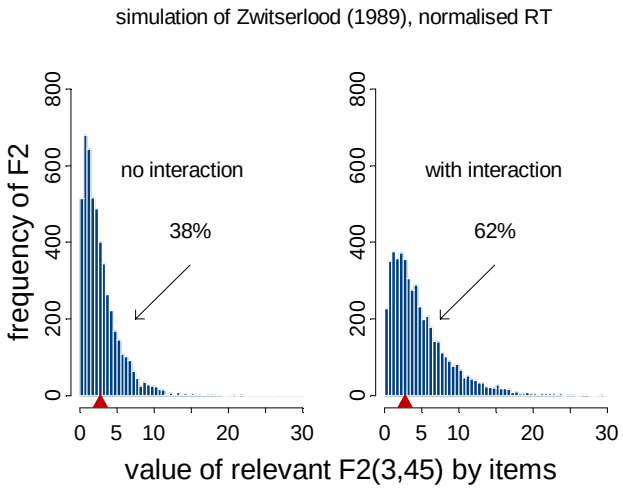


Figure 4. Distributions of 5000 F_2 ratios, after normalisation of each data set.

5 Discussion and conclusion

An absence of reliable partial semantic priming effects does not imply that multiple word candidates have not been activated. There is a large body of evidence supporting the concept of early multiple activation of lexical candidates (with subsequent competition among these candidates). This evidence has been collected using various experimental tasks: phonological priming (Slowiaczek, McQueen, Soltano, & Lynch, 2000), word identification (Luce, Pisoni, & Goldinger, 1990), word spotting (Cutler & Norris, 1988; Norris, McQueen, & Cutler, 1995), phoneme classification (Borsky, Tuller, & Shapiro, 1998), phoneme monitoring (Gaskell & Marslen-Wilson, 1998; Vroomen & de Gelder, 1999), tracking of eye movement (Dahan, Magnuson, Tanenhaus, & Hogan, 2001), and pause detection (Mattys & Clark, 2002). Thus, spoken-word processing indeed involves activation of multiple word candidates, but cross-modal semantic priming is not suitable for tapping into these multiple activations.

It seems that the activation of a word candidate may have to surpass a certain threshold level before it passes on to semantically related items, or before it has *detectably* affected the activation of semantic associates. Only when one candidate remains, activation of that candidate will be high enough to spread activation detectably to its semantic relatives.

Within the DC model (Gaskell & Marslen-Wilson, 1997, 1999, 2002), as in all connectionist networks, multiple representations must interfere with each other if they are active simultaneously. As long as multiple candidates are active, the lack of overlap in the semantic nodes translates into small or inconsistent semantic priming effects, if any (cf. Gaskell & Marslen-Wilson, 2002). Perhaps future research can shed more light on how activation spreads to related items. At this point it is impossible to choose between the original spreading-of-activation account (Collins & Loftus, 1975) and the DC model. However, both accounts predict that tapping into multiple activation via semantic priming is inherently difficult because the effects are small. This prediction is verified by several other failures to find partial-priming effects (Chwilla, 1996; Jongenburger, 1996).

A recent article has investigated the psychological reality of the recognition point in spoken-word processing. The Cohort model (Marslen-Wilson, 1993) proposes a recognition point as

the point where the word diverges from the other members of its word-initial cohort; the Shortlist model (Norris, 1994) however does not predict when a word presented in isolation will be recognised. The uni-modal repetition priming study by Bölte & Uhe (2004) investigated the influence of sensory information following the recognition point of the prime. Repetition priming effects were studied at the recognition (RP cut-off) point, at a later cut-off point (RP-plus), and at the offset of the complete prime. Bölte & Uhe found that the priming effect at the RP-plus condition was slightly larger than at the recognition point, but that the priming effect in the complete-prime condition was significantly larger than at the two cut-off conditions. These results provide counterevidence against a strong formulation of the recognition point, in which lexical activation does not increase any further from the recognition point onwards. Bölte & Uhe (2004: 145) argue that the recognition point is the "moment at which the word recognition system makes a commitment to a certain lexical representation. Further information is used (1) to distinguish between, for instance, morphological alternatives and (2) to raise the lexical activation of matching lexical representations rather gradually. Still, a word is not selected at this moment." Importantly, at the recognition point, even the phonological priming effect has not reached its maximum. It is therefore no surprise that semantic priming effects show up only after the recognition point.

In conclusion, we have employed Monte Carlo simulations to investigate the chances of Type I and Type II errors in one key study. These simulations indicate that this study had a high chance of finding spurious effects of partial priming (of a Type I error). Partial priming effects were not observed in the replication experiment, nor in several similar studies. However, there is overwhelming experimental evidence that multiple word candidates are activated during spoken-word recognition. We have to conclude that the cross-modal semantic priming technique does not provide valid and reliable insight into this multiple activation. It is therefore advisable to discontinue its use for that purpose.

References

Bailey, R. A. (1982). Confounding. In S. Kotz (Ed.), *Encyclopedia of Statistical Sciences* (Vol. 2, pp. 128-134). New York: Wiley.

Bölte, J. & Uhe, M. (2004). When is all understood and done? The psychological reality of the recognition point. *Brain and Language*, *88* (1), 133-147.

Borsky, S., Tuller, B., & Shapiro, L. P. (1998). "How to milk a coat": The effect of acoustic and semantic information on phoneme categorization. *J. Acoustical Society of America*, *103*, 2670-2676.

Campbell, D. T. (1969). Prospective: Artifact and control. In R. L. Rosnow (Ed.), *Artifact in Behavioral Research* (pp. 351-382). New York: Academic Press.

Chwilla, D. J. (1996). *Electrophysiology of word processing: the lexical processing nature of the N400 priming effect*. Unpublished doctoral dissertation, University of Nijmegen, Nijmegen.

Cochran, W. G., & Cox, G. M. (1957). *Experimental Designs*. New York: Wiley.

Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review*, *83*, 407-428.

Connine, C. M., Blasko, D. G., & Wang, J. (1994). Vertical similarity in spoken word recognition: Multiple lexical activation, individual differences, and the role of sentence context. *Perception and Psychophysics*, *56*, 624-636.

Cox, D. R. (1958). *Planning of Experiments*. New York: Wiley.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *J. Experimental Psychology: Human Perception and Performance*, *14*, 113-121.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, *16*, 507-534.

Forster, K.I. (1976). Accessing the mental lexicon. In: R.J. Wales & E.C.T. Walker (Eds.), *New Approaches to Language mechanisms: a cross section of of psycholinguistic studies* (pp. 257-287). Amsterdam: North Holland. North Holland Linguistics series; 30.

Forster, K.I. (1979). Levels of processing and the structure of the language processor. In: W.E. Cooper, & E.C.T. Walker (eds.). *Sentence processing: Psycholinguistic studies presented to Merill Garrett* (pp. 27-85). Cambridge, MA: MIT Press.

Gaskell, M. G., & Marslen-Wilson, W. D. (1996). Phonological variation and inference in lexical access. *J. Experimental Psychology: Human Perception and Performance*, *22*, 144-158.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: a distributed model of speech perception. *Language and Cognitive Processes*, *12* (5/6), 613-656.

Gaskell, M. G., & Marslen-Wilson, W. D. (1998). Mechanisms of phonological inference in speech perception. *J. Experimental Psychology: Human Perception and Performance*, *24* (2), 380-396.

Gaskell, M. G., & Marslen-Wilson, W. D. (1999). Ambiguity, competition and blending in spoken word recognition. *Cognitive Science*, *23* (4), 439-462.

Gaskell, M. G., & Marslen-Wilson, W. D. (2002). Representation and competition in the perception of spoken words. *Cognitive Psychology*, *45*, 220-266.

Hammersley, J. M., & Handscomb, D. C. (1964). *Monte Carlo Methods*. London: Methuen.

Janse, E. (2003). *Production and perception of fast speech*. Doctoral dissertation, Utrecht University, Utrecht, pp.175-197.

Jongenburger, W. (1996). *The role of lexical stress during spoken word processing*. Doctoral dissertation, Universiteit Leiden.

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (1990). Similarity neighborhoods of spoken words. In G. T. M. Altmann (Ed.), *Cognitive Models of Speech Processing*. Cambridge, MA: MIT Press.

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken-language understanding. *Cognition*, *8*, 1-71.

Mattys, S.L., & Clark, J.H. (2002). Lexical activity in speech processing: evidence from pause detection. *J. Memory and Language*, *47*, 343-359.

Max, L., & Onghena, P. (1999). Some issues in the statistical analysis of completely randomized and repeated measures designs for speech, language, and hearing research. *J. Speech, Language and Hearing Research*, *42*, 261-270.

Maxwell, S.E., & Delaney, H.D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.), Mahwah, NJ: Lawrence Erlbaum Associates.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1-86.

Meyer, D., & Schvaneveldt, R.W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *J. Experimental Psychology*, *90*, 227-234.

Moss, H. E., McCormick, S. F., & Tyler, L. K. (1997). The time course of activation of semantic information during spoken word recognition. *Language and Cognitive Processes*, *12* (5/6), 695-731.

Neely, J.H. (1977). Semantic priming and retrieval from lexical memory: roles of inhibitionless spreading of activation and limited-capacity attention. *J. Experimental Psychology: General*, *106*, 226-254.

Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition*, *52*, 189-234.

Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and Segmentation in Spoken-Word Recognition. *J. Experimental Psychology: Learning Memory and Cognition*, *21* (5), 1209-1228.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, *23* (3), 299-370.

O'Brien, R.G., & Kaiser, M.K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, *97* (2), 316-333.

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review*, *107* (4), 786-823.

Slowiaczek, L. M., McQueen, J. M., Soltano, E. G., & Lynch, M. (2000). Phonological representations in prelexical speech processing: Evidence from form-based priming. *J. Memory and Language*, *43* (3), 530-560.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *J. Verbal Learning and Verbal Behaviour*, *18*, 645-659.

Vroomen, J., & de Gelder, B. (1999). Lexical access of resyllabified words: Evidence from phoneme monitoring. *Memory & Cognition*, *27*(3), 413-421.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, *32*, 25-64.

Zwitserlood, P., & Schriefers, H. (1995). Effects of sensory information and processing time in spoken word recognition. *Language and Cognitive Processes*, *10*, 121-136.

The Integrated Language Database, with an aside on the Spoken Dutch Corpus *

Truus Kruyt

Institute for Dutch Lexicology INL

Abstract

One of the current projects of the Institute for Dutch Lexicology (INL) is the Integrated Language Database of 8th–21st-Century Dutch (ILD). The aim is to create a flexible linguistic research instrument by linking electronic dictionaries, a balanced diachronic text corpus and lexicons of historical and present-day Dutch. We aim to link part of our data with data collections stored at other institutes, creating a supra-institutional research instrument. The present paper gives an overview of the project, with an aside on the Spoken Dutch Corpus.

1 Introduction

The Institute for Dutch Lexicology (INL) has a long-standing tradition in corpus-based lexicography. As a result, the INL now has electronic scholarly dictionaries of the Dutch language covering the vocabulary from 1200 up to 1976, and text corpora covering mainly (Early) Middle Dutch and present-day Dutch. The Dutch PAROLE corpus and the Dutch PAROLE/SIMPLE lexicon were developed in a European context[2].

Three linguistically annotated corpora of present-day Dutch have been widely used for various research purposes in the fields of linguistics and social studies, for lexicography and lexicon building, for academic teaching, and for the delivery of customized data, since they became Internet-accessible in 1994 (Kruyt, 1998). The Dutch PAROLE corpus will soon be accessible for similar purposes (Van der Kamp & Kruyt, 2004). The follow-up is a bi-national, long-term INL project: the Integrated Language Database of 8th–21st-Century Dutch (ILD). Our aim is to provide a flexible instrument for a wide range of synchronic and diachronic research into the Dutch language (and culture) throughout the centuries. For the purpose of flexible retrieval and navigation, various data types within the ILD will be linked. We also intend to link part of our data with data stored at other centres, creating a supra-institutional research instrument. See for projects with common features: Gellerstam, Cederholm & Rasmark (2000), Fournier (2001), Ruus (2002).

This paper reports on the overall ILD design (§2). Then the user's perspective is considered (§3). A prototype will function as a demonstration model to verify and assess user needs. In the current phase of the project, it functions to test the design empirically for its applicability to 'real data', to develop efficient procedures, and to obtain figures on workload for future planning (§4). The paper concludes with an aside on the Spoken Dutch Corpus (CGN).

---

* This paper is adapted from Kruyt (2004).

[2] See URL http://www.inl.nl/eng/europe/projects.htm

2 The Overall ILD Design

2.1 Contents

The ILD will have two dimensions. One is the diachronic dimension; data cover 8th- to 21st-century Dutch. The other is the linguistic dimension; for each time period, various types of linguistic data are available: encoded dictionary data, linguistically annotated texts, and lexicon data.

The ILD will consist of three mutually linked components: a dictionary component, a balanced diachronic text corpus component, and a component with lexicons of historical and present-day Dutch. The dictionary component will comprise the Dictionary of Early Middle Dutch VMNW (four printed volumes), the Dictionary of Middle Dutch MNW (ten volumes) and the Dictionary of the Dutch Language WNT (43 volumes), and in the longer term the dictionaries of Old Dutch and present-day Dutch (ongoing INL projects). These dictionaries are the most comprehensive dictionaries of the Dutch language, compiled according to scholarly principles, and eventually covering the Dutch vocabulary from the 8th up to the 21st century. For these reasons, they are considered a separate component of the ILD (along with some smaller supplementary dictionaries). They are available in machine-readable form, albeit with a different extent of encoding.

The diachronic text corpus should support a wide range of user needs (cf. §1). It will therefore cover many varieties of Dutch written language, dating from the 8th–21st century. As no existing corpus design turned out to be applicable to texts from so many centuries, we developed a new one (Van Dalen-Oskam, Geirnaert & Kruyt, 2002), which, after several empirical tests, has been applied in the prototype. The leading principle is 'the primary aim of a text', with two major divisions that more or less correspond with fiction and non-fiction: a creative ('imagination') vs. a factual ('information') representation of knowledge and information.[3] Within this framework, twenty-two text types have been distinguished. Criteria for text selection have also been developed. Apart from other texts, texts quoted in the dictionaries of the dictionary component will be selected so as to be able to link corpus and dictionary data. For the acquisition of digital texts, see §4.

The lexicon component will consist of a rather restricted, well-motivated selection of present-day and historical lexica. The main criterion for the selection of lexica is that they provide information that is not at all, or much less, elaborated in the dictionaries of the dictionary component. For present-day Dutch, the PAROLE/SIMPLE lexicon, although period specific, is relevant. Historical lexica will contain headwords (and their paradigmatic word forms) found in dictionary quotations or in corpus texts, but not covered as an entry by the dictionaries in the dictionary component. Such a lexicon not only fills a gap in the lexicographical description of Dutch, but is also needed for the annotation of historical Dutch texts.

2.2 Annotation

For linking and retrieval purposes, the data will be encoded according to the TEI standard. The dictionaries in the dictionary component will be encoded for the major information types within the entries (headword, etymology, quotation, meaning description, etc.). For the dictionaries MNW and WNT, this requires a substantial extension of the current encoding and an improvement of the dictionary files, due to inconsistency and lexicographical practices (cf. Kruyt & Van der Voort van der Kleij, 1992-93). Furthermore, present-day

[3] Especially for old texts, the distinction between fiction and non-fiction cannot be drawn sharply.

Dutch headwords are added to the (historical) headwords of all dictionaries for easy retrieval and linking; this work has been finished for the VMNW dictionary and for about 90,000 headwords of the WNT.

The texts in the diachronic text corpus will be encoded at several levels. At the text level, the text type and other metadata (still to be specified) will be encoded, as parameters for the selection of a user-defined subcorpus. Within the texts, the text structure, the typography and some other textual elements will receive basic encoding geared to retrieval purposes. The design is ready (Depuydt & Dutilh, 2002) and is now being tested with prototype texts. We have adopted a 'database view' on text, which implies, among other things, a clear distinction between the actual text and its medium (such as manuscript, printed book, electronic file); see further §4.

Work on how we should tag the words ('tokens') of the texts for part of speech (PoS) from a diachronic perspective is in progress. In a first, maximal approach, we used a slightly different version of the Dutch EAGLES/PAROLE tag set and we manually tagged the tokens of three historical prototype texts from different periods with both a 'lexical' and a 'functional' tag when applicable (cf. Dutilh & Kruyt, 2002). Decorte, Dutilh-Ruitenberg & Kruyt (2004), however, conclude that this approach is not feasible, mainly due to lack of consensus among linguists on how to handle linguistic phenomena such as transcategorisation, lexicalisation and grammaticalisation. See §4.4 for the follow-up. Apart from PoS, all tokens will be lemmatized with a present-day Dutch headword, or an etymologically reconstructed one when there is no modern equivalent.

Lexica need no annotation, because all information is explicit and unambiguous.

2.3 Linking

For user-friendly navigation, links will be established between data within a source and between data of different sources, including external sources. The linking functionality implies that a mouse-click leads the user from a particular point in a query result to related data elsewhere, within or outside the ILD. We will implement direct and indirect links, the latter offering the user several destinations to choose from. Links foreseen include a link from a dictionary entry to its corresponding entry in another dictionary (through the present-day headwords; §2.2); from a dictionary quotation to its equivalent in the original text in the corpus component (for more context); from a corpus word to corresponding entries in the dictionary component and, vice versa, from a dictionary headword to corpus tokens (through the present-day headwords); from a corpus text to metadata; from an arbitrary word in the ILD to other occurrences in the ILD, or to the same word stored at some external centre.

In the longer term, different dictionary headwords (also with different PoS) will be linkable at word-sense level by using the SIMPLE lexicon and its ontology with semantic types and qualia roles (Pustejovsky, 1998).

3 The user's perspective

The ILD data will be accessible by means of a retrieval system that will offer its users many more facilities than our present corpus systems, due to the various data types and the diachronic dimension within the ILD, and due to more advanced means of retrieval and navigation. An information-technological concept is now being developed by our IT department. The PAROLE interface (Van der Kamp & Kruyt, 2004) functions as a model for the corpus component. In the EC-funded ELAN project, we participated in building a

prototype retrieval system with access to geographically distributed data through one user interface.

To be geared to a broad user group, historical data will be accessible by use of a present-day Dutch headword. Etymologically reconstructed headwords will be presented to the user together with morphologically or semantically related modern Dutch headwords (e.g. reconstructed *aanvaardigen* with modern *aanvaarden*, 'accept'). Of course, specialists in historical Dutch can have access via historical forms as well.

Provided that the data are sufficiently annotated and linked, such a retrieval system will offer users many research facilities. Here follow some examples. A researcher who is interested in the history of words may ask the system: for the present-day Dutch word X, give me the corresponding headwords with their form variants and etymology sections from the dictionaries WNT, MNW and VMNW. A researcher can ask for more usages of a headword in the corpus texts if the quotations in the dictionaries are not satisfactory. Someone interested in spelling may ask: list all variant forms of the headword Y with their text source and geographical location. A researcher interested in loan words may ask: list loan words from French attested in the dictionary WNT and in 18th-century narrative texts. And if relations from the SIMPLE lexicon can be used, a researcher interested in the vocabulary of the Industrial Revolution may ask: find words with word senses belonging to the semantic class of 'instrument' attested in 19th-century texts about science. If specific information is not available in the ILD, the researcher can navigate to an external database. The list of potential research options that make use of the annotated and linked data is virtually endless. That is still in the future. The first step now is the ILD prototype (§4).

4 The ILD prototype

4.1 Introduction

When the corpus design was ready, we started building an ILD prototype, a small-scale model of the contents and the retrieval functionalities of the ILD, including links from some French-Flemish dialect headwords in the VMNW and MNW dictionaries to a dialect centre in Belgium.

In the current phase of the project, it is used to empirically test the soundness and applicability of the conceptual ideas, to develop efficient procedures, and to measure workloads in view of future planning of intermediate products. These functions have turned out to be extremely useful, as particularly historical texts and their information carriers have many unforeseen characteristics requiring solutions. The prototype can therefore be considered an indispensable pilot for the ILD. We started with the prototype corpus component. Below follows a description of the results so far.

4.2 Text selection and acquisition

In principle, 224 text fragments of about five pages (carefully selected from front, body and back) were planned and selected according to the corpus design, covering the 8th to 20st century represented by eight periods. The proportion is 33% 'imagination' and 66% 'information' (cf. §2.2). In 31 cases, suitable texts could not (yet) be found, almost all of them for the period before the 15th century, due the general problem that only few old texts have survived. Forty-three texts were acquired from digital repositories elsewhere, 150 text fragments were digitized by in-house scanning and correction. For text editions, we applied the criteria for measuring the editorial quality (Van Dalen-Oskam, Geirnaert & Kruyt, 2002), in order to choose the best one if more than one was available. For all texts, bibliographic and

other metadata are available in a rather simple Access database; for the ILD, we foresee a more sophisticated database.

We started with instructions for digitizing that were aimed at a rather detailed representation of textual characteristics. This was common practice in many other projects using TEI and, due to organisational factors, we did not yet know at the time what degree of detail would be necessary for our TEI encoding of text structure and typography. The experience we gained from digitizing and encoding so many historical texts, with so many unexpected peculiarities, has changed our view on future digitizing for the ILD, which will be less detailed (in line with our current database view; §2.2), and focused on the actual text rather than on the characteristics of the text medium, such as certain decorative features. Furthermore, due to the knowledge of TEI encoding acquired through the prototype, it will become possible, to a large extent, to merge the processes of digitizing and TEI encoding. This will lead to a much more efficient procedure in the future.

4.3 Encoding of text structure and typography

There are two major issues relevant to the encoding of text structure and typography: our database view on text (§2.2) and the notion of what we consider 'the text to be encoded'.

The database view implies that we will abstract from the original typography and font, and display equal structural text elements in a uniform rendering on screen. We still need to define what rendering we will use. Due to the detailed method of digitizing, we will have to remove the encoding that has become superfluous according to the database view.

As for the notion of 'the text to be encoded', we give priority to the original text selected according to the corpus design, irrespective of its publication in a text edition or as part of a larger entity (an anthology, for example). Consequently, when applicable, the text is isolated from the text around it and the encoding does not account for the place of the text in the overall structure of the complete publication (whether a comprehensive printed work or an electronic file). We only retain the editor's transcription method and the editorial notes, which offer essential information to the user. As a practical consequence, we do not need to digitize more text than intended for our purpose.

We nearly finished the TEI encoding of the 150 in-house digitized prototype text fragments. After some automatic conversion and validation procedures, 'pre-TEI' tagged XML files of the texts have been encoded manually with the aid of a purpose-built editorial tool. So far, the encoding design has only needed some minor adaptations, though some issues are still to be decided on. For example, we consider extending the form-based type specification of particular div's[4] (e.g. letters), in view of a refined retrieval or subcorpus selection. The application of the design presented us with three major practical problems, as 'real' texts show much more variety than TEI accounts for. One was that TEI sometimes does not provide satisfying solutions, resulting in rather contrived encoding. The second was the choice of a suitable TEI tag when a structural text element approaches more than one TEI definition. The third was the development of criteria for consistent and transparent solutions when more than one solution is TEI-acceptable.

As for the files derived from external repositories, we investigated their characteristics and differences with the in-house digitized files, and we started to encode them, adhering to the principle that all files will receive basic encoding according to the design, if feasible.

---

[4] Subdivisions within a text; for an exact definition see URL http://www.tei-c.org/P4X/REFTAG.html

4.4 PoS tagging and lemmatizing

After our first experience with PoS tagging (§2.2), we elaborated a more modest approach, starting from a reduced tag set and applying a lexical tag method only. We will investigate to what extent we can compensate for the less refined tagging by offering predefined complex queries in the interface, which can be customized by the user, i.e. a functionality similar to the one called 'patterns' in the PAROLE corpus retrieval system (Van der Kamp & Kruyt, 2004). We need a substantial amount of data to be able to define such patterns. As PoS tagging and lemmatisation are related issues, we recently started tagging and lemmatising prototype text fragments from all periods. A tool was built to make this manual work as efficient as possible. Our approach to tagging and lemmatising probably needs to be customized gradually, depending on the empirical results. The outcome will be a linguistically annotated prototype corpus and a prototype historical lexicon. As a separate activity, we are currently developing a historical lexicon of Middle Dutch by automatically matching the MNW headwords with their paradigmatic word forms attested in the quotations. In spite of all spelling variants, our program matches over 92% for the alphabetic sections A to K.

4.5 The next steps

All design decisions have been accounted for in reports and discussed with the User Committee and Advisory Board connected to our project. After completion of the prototype, a comprehensive report and the prototype will be presented not only to the members of these committees, but also to our present corpus users and other interested parties. They will be requested to give feedback. This may result in revisions of aspects of the design and/or the retrieval functionalities.

After the prototype, we will develop and make available subsystems of the ILD as intermediate products, rather than wait until the complete ILD has been realized. We will investigate whether it is feasible to give users access already in the development phase.

5 The ILD (and other INL corpora) and the Spoken Dutch Corpus (CGN)

Given its context, a *liber amicorum* for Sieb Nooteboom, this paper would be incomplete if it would not touch on a question related to Sieb's field: to what extent can an INL corpus of written Dutch (the ILD or another corpus) and the Spoken Dutch Corpus (CGN) be used for comparative research into written and spoken language? There are several reasons to be optimistic about the answer. The CGN is unique in the sense that it is designed for use in a number of widely different fields of interests (Oostdijk, 2000; Oostdijk, Goedertier, van Eynde, Boves, Martens, Moortgat & Baayen, 2002). This also applies to many written corpora, including the already available INL corpora (cf. §1) and the ILD design. Both the CGN and the INL corpora distinguish text types for the selection of corpus data, which are assumed to cover many varieties of linguistic usage; the text types are different, but there is an overlap. Both the CGN and the INL corpora contain Dutch and Flemish. Both in the CGN and in the INL corpora, metadata enable the user to query a subcorpus rather than the whole corpus. Both in the CGN and in the INL corpora, the corpus data are annotated with lemma and part of speech; the tag set and the method of tag assignment are different but nevertheless compatible to a reasonable extent (Dutilh-Ruitenberg, 2002). Both the CGN and the INL corpora apply international standards, such as EAGLES and TEI/CES. And, finally, the corpus data of both the CGN and the INL corpora are accessible through a retrieval system.

Although the differences between the corpora should not be ignored, a tentative conclusion is that the CGN and the INL corpora have enough common features to warrant comparative

research into written and spoken language. Of course, a firm answer to the question cannot be given yet. But supposing the retrieval systems can be connected in some way (as we managed to do with two other participants in the ELAN project; cf. §3), then researchers could confirm our preliminary conclusion. By virtue of the CGN, the research communities of written and spoken Dutch might get closer to each other.

References

Decorte, S., Dutilh-Ruitenberg, T., & Kruyt, T. (2004). Language change and linguistic annotation in the Integrated Language Database of 8th-to21st-Century Dutch. Manuscript submitted for publication in *Proceedings 2. Freiburger Arbeitstagung zur Romanistischen Korpuslinguistik: Korpuslinguistik und Historische Sprachwissenschaft. Sektion A. Korpusprojekte, Sprachdatenverwaltung und Analyse-werkzeuge*. Available: http://www.inl.nl/eng/pub/decorte_freiburg_eng.pdf

Depuydt, K., & Dutilh-Ruitenberg, T. (2002). TEI encoding for the Integrated Language Database of 8th–21st-Century Dutch. In Anna Braasch & Claus Povlsen (Eds*.), Proceedings of the Tenth Euralex International Congress, EURALEX 2002* (Vol. II, pp. 683-688). Copenhagen: Center for Sprogteknologi. Available: http://www.inl.nl/eng/pub/tei.htm

Dutilh-Ruitenberg, T. (2002). *Verschillen tussen CGN en PAROLE: de tagmethode en tagset.* Unpublished INL paper.

Dutilh, T. & Kruyt, T. (2002). Implementation and Evaluation of PAROLE PoS in a National Context. In Manuel González Rodríguez & Carmen Paz Suarez Araujo (Eds*.), Proceedings of the third International Conference on Language Resources and Evaluation* (pp. 1615-1621). Paris: ELRA. Available: http://www.inl.nl/eng/pub/lrec2002.pdf

Fournier, J. (2001). New directions in Middle High German Lexicography: Dictionaries Interlinked Electronically. *Literary and Linguistic Computing*, 16(1), 99-111.

Gellerstam, M., Cederholm, Y., & Rasmark, T. (2000). The bank of Swedish. In Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S., & Stainhaouer, G. (Eds.), *Proceedings Second International Conference on Language Resources & Evaluation* (pp. 329-333). Paris: ELRA.

Kruyt, J.G., & Van der Voort van der Kleij, J.J. (1992-93). Towards a computerized historical dictionary of Dutch. In *Acta Linguistica Hungarica* 41(1-4) (pp. 159-174). Budapest: Hungarian Academy of Sciences.

Kruyt, J.G. (1998). Dutch written language resources, their users and uses. In Rubio, A. Gallardo, N., Castro, R. & Tejada, A. (Eds.), *Proceedings of the First International Conference on Language Resources & Evaluation* (pp. 959-963). Paris: ELRA. Available: http://www.inl.nl/eng/pub/grancon.htm

Kruyt, J.G. (2004). The Integrated Language Database of 8th – 21st-Century Dutch In Lino, M.T, Xavier, M. F., Ferreira, F., Costa, R., Silva, R. (Eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation* (pp. 1751-1754). Paris: ELRA. Available: http://www.inl.nl/eng/pub/LREC2004_kruyt_eng.pdf

Oostdijk, N. (2000). The Spoken Dutch Corpus. Overview and first Evaluation. In Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S., & Stainhaouer, G. (Eds*.), Proceedings Second International Conference on Language Resources & Evaluation* (pp. 887-893). Paris: ELRA.

Oostdijk, N., Goedertier, W., Van Eynde, F., Boves, L., Martens, J-P., Moortgat, M., & Baayen, H. (2002). Experiences from the Spoken Dutch Corpus project. In Manuel González Rodríguez & Carmen Paz Suarez Araujo (Eds.), *Proceedings of the third International Conference on Language Resources and Evaluation* (pp. 340-347). Paris: ELRA.

Pijnenburg, W.J.J., Van Dalen-Oskam, K.H., Depuydt, K.A.C., & Schoonheim, T.H. (2000). *Vroegmiddelnederlands Woordenboek. Woordenboek van het Nederlands van de dertiende eeuw in hoofdzaak op basis van het Corpus-Gysseling.* Leiden: Instituut voor Nederlandse Lexicologie.

Pustejovsky, J. (1998) *The generative lexicon.* Cambridge, MA: MIT Press.

Ruus, H. (2002). A Corpus-based Electronic Dictionary for (Re)search. In Anna Braasch & Claus Povlsen (Eds.), *Proceedings of the Tenth Euralex International Congress*, *Euralex 2002* (Vol. I, pp. 175-185). Copenhagen: Center for Sprogteknologi.

Van Dalen-Oskam, K., Geirnaert, D., & Kruyt, T. (2002). Text Typology and Selection Criteria for a balanced Corpus: The Integrated Language Database of 8th–21st-century Dutch. In Anna Braasch & Claus Povlsen (Eds.), *Proceedings of the Tenth Euralex International Congress Euralex 2002* (Vol. II, pp. 401-406). Copenhagen: Center for Sprogteknologi. Available: http://www.inl.nl/eng/pub/euralex_dalen_eng.htm

Van der Kamp, P., & Kruyt, T. (2004). Putting the Dutch PAROLE Corpus to Work. In Lino, M.T, Xavier, M. F., Ferreira, F., Costa, R., Silva, R. (Eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation* (pp. 1767-1770). Paris: ELRA. Available: http://www.inl.nl/eng/pub/ LREC2004_kamp_kruyt_eng.pdf

Verwijs, E., & Verdam, J. (1885-1929). *Middelnederlandsch Woordenboek*. 's-Gravenhage: Martinus Nijhoff.

Woordenboek der Nederlandsche Nederlandsche Taal (1882-1998). Leiden: Martinus Nijhoff. 's Gravenhage: SDU.

Segmental anchoring of pitch movements: autosegmental phonology or speech production?

D. R. Ladd

Edinburgh University

Abstract

Arvaniti, Ladd & Mennen (1998) reported a phenomenon of "segmental anchoring": the beginning and end of a pitch movement are anchored to specific locations in segmental structure, while the slope and duration of the pitch movement vary according to the segmental material with which it is associated. This finding has more recently been replicated and extended in several languages. In a superficially attractive autosegmental interpretation of the phenomenon, the pitch movement is analysed as a sequence of tones which independently undergo "secondary association" to specific points in structure. However, recent empirical findings make such a phonological account less plausible, and suggest that some aspects of segmental anchoring need to be explained in terms of quantitative language-specific phonetic detail. Analogies to diphthong dynamics and to voice onset time, meanwhile, provide an impetus for detailed studies of the production and perception of pitch movements that may advance our understanding of segmental anchoring in general.

1 Segmental anchoring: the basic finding

In some intuitively clear way, F0 features such as tone and accent belong with specific elements of the segmental string: Chinese tones go with syllables (or possibly syllable rhymes), English pitch accents go with stressed syllables, Japanese word accents go with a specific mora, etc. This loose belonging together is known in the autosegmental phonological literature as "association". However, it has been clear for some time that the precise temporal coordination or alignment of F0 events with segmental events is quite complex, and does not follow straightforwardly from the mere fact of association. For example, the peak of a pitch accent may be aligned outside (usually after) the stressed syllable with which it is intuitively associated. Various studies in the early 1990s (notably Silverman & Pierrehumbert, 1990, on English and Prieto, van Santen & Hirschberg, 1995, on Spanish) attempted to identify the interacting factors that contribute to the extent of this phenomenon of "peak delay". Their principal finding was that an upcoming word boundary and/or an immediately following pitch accent ("stress clash") can cause the peak to occur earlier; in effect, the peak can be "delayed" when the distance to the next prosodically relevant event is increased. About the same time, Caspers & van Heuven (1993) investigated the effect of "time pressure" on pitch movements in Dutch. They studied the alignment of rising and falling pitch movements with the associated accented syllable as they manipulated speech rate, phonological vowel length, and distance to the following pitch movement. They found a complex set of effects, apparently different for rises and falls, but did not arrive at any general principles governing the alignment of pitch movements with the associated segmental material. However, one of their key findings is relevant for this paper, namely that irrespective of the "time pressure" manipulations, the beginning of an accentual pitch rise in Dutch is aligned with the beginning of the accented syllable. The same was subsequently reported for Spanish by Prieto et al. (1995).

All of the studies just cited were based on the assumption that the duration of a pitch movement would be affected by the amount of time available for its completion, and for at least certain conditions they showed that this assumption is valid. However, it has since become clear that "time pressure" does not provide a comprehensive explanation of peak delay, and this is the topic of this paper. The line of research sketched here began with a study of the alignment of prenuclear rising pitch accents in Modern Greek (Arvaniti, Ladd & Mennen, 1998 [henceforth ALM]). ALM originally accepted the "time pressure" assumption, and already knew (Arvaniti & Ladd, 1995) that the beginning of the prenuclear rising pitch accent in Greek is aligned with the beginning of the syllable – as in Dutch and Spanish. In their first experiment they therefore varied the amount of time between one prenuclear rise and the next, assuming that this would affect peak delay. As it happened, however, their experimental manipulations did not seem to have any significant or consistent effect on peak delay, and they came to suspect that "time pressure" affects alignment only when it is extreme, e.g. when one pitch movement is *immediately* adjacent to another pitch movement, or to a prosodic boundary. Given enough distance – a couple of syllables – between pitch movements, "peak delay" does not seem to be controlled by the kinds of factors investigated in the studies just cited. In their subsequent experiments, ALM showed that if there is no strong "time pressure", alignment is controlled by what we may call *segmental anchoring***.** Specifically, they discovered that *both* the beginning and the end of pitch movements are consistently aligned in time with identifiable landmarks in the segmental string, such as specific segment boundaries or syllable boundaries. In the case of Greek, prenuclear rising accents begin at the end of the pretonic syllable and end early in the posttonic vowel, regardless of the distance between those points. This underlying alignment is what is modified by extreme "time pressure".

Figure 1. Idealized diagrams of the segmental anchoring of the Greek prenuclear rise with different segmental material. In the left-hand panel the rise spans the short-duration stretch [ðit] and in the right-hand panel it spans the longer sequence [ɾɛmv], but the starting and ending F0 levels are the same in both cases.

Among other things, ALM's finding means that the duration of a pitch rise is almost entirely a function of the time interval between the two segmental landmarks – the end of the pretonic syllable and the beginning of the posttonic vowel. Moreover, the "scaling" (F0 level) of the beginning and the end of the rise is unaffected by its duration; longer rises do not have greater overall pitch excursions. This means that the slope of the rise also depends on the time interval between the two segmental landmarks. These findings are summarized in the idealized diagrams in Figure 1. As ALM point out, these findings are difficult to reconcile with earlier phonetic models of pitch contours (e.g. Fujisaki, 1983, 't Hart, Collier & Cohen, 1990, even Taylor, 2000, to some extent), which assume that slope and/or duration are the most appropriate ways of characterizing accent types. The existence of segmental anchoring clearly suggests that slope and duration are not the identifying characteristics of pitch movements, but rather that slope and duration depend on the scaling and alignment of "tonal targets". The finding of segmental anchoring thus appears highly consistent with the

autosegmental interpretation of pitch movements as sequences of tones (L+H*, etc.): on this view, the individual tones would be anchored to specific places in structure. I return to this interpretation in section 4.

2 Further developments and related findings

Surprising though it was to many people, ALM's finding of segmental anchoring has since been replicated for other languages and extended in various ways. These developments are summarized here; points are numbered for subsequent reference:

i. Segmental anchoring under changes of speech rate in English (Ladd, Faulkner, Faulkner & Schepman, 1999). English rising prenuclear accents remain anchored to segmental landmarks regardless of speech rate: slope and duration are adjusted to keep the beginning and end of accentual F0 rises aligned with their respective segmental anchors as segment durations decrease or increase with rate. Caspers & van Heuven found a similar effect for Dutch in their study, though of course they did not interpret it in exactly these terms.

ii. Effects of phonological vowel length (tenseness) on segmental anchoring in Dutch and English (Ladd, Mennen & Schepman, 2000). The beginning of prenuclear rising accents in both Dutch and English is aligned as in Greek, but the alignment of the end (i.e. peak) of the rise depends on whether the vowel is phonologically long or short (tense or lax): the peak accompanying a long vowel is late in the vowel, but that accompanying a short vowel is late in the following consonant. This is not merely a "time pressure" effect of segment duration, as was assumed by Caspers & van Heuven: Dutch "long" /i/ and "short" /ɪ/ are essentially identical in phonetic duration and differ only in vowel quality; nevertheless, a difference in alignment is still found. Similar effects are found in English (Ladd, Schepman & White, work in progress), although the definition of phonological vowel length is less clear in many varieties of English than it is in Dutch.

iii. Consistent alignment of between-accent F0 valleys in English (Ladd & Schepman, 2003; Dilley, Ladd & Schepman, in press). The F0 valley between accents is aligned with the beginning of the second accented syllable. This means that in potentially ambiguous phrases like *Norma Nelson* and *Norman Elson,* the alignment of the F0 valley is affected by the syllable membership of the ambiguous consonant. However, in keeping with idea that the valley and the peak are aligned independently, there is no significant effect of the consonant's syllable membership on the alignment of the following accentual peak; accentual rises are shorter and steeper in syllables that begin with a vowel (*Norman Elson*) than in those that begin with a consonant *(Norma Nelson*).

iv. Regularities in the alignment of Chinese lexical tone contours with syllables (e.g. Xu 1998, 1999). Xu's extensive body of work on alignment in Chinese, which is entirely independent of our own, has yielded a number of findings consistent with the work that builds directly on ALM. The clearest example is his finding that the end of the rising contour for Mandarin second tone is closely coordinated with the end of the syllable, regardless of speech rate and syllable composition (specifically the presence or absence of a nasal coda).

v. Phonological factors in the alignment of Japanese accentual H (Ishihara 2003). The end (peak) of the F0 rise signalling a word-initial accent in Tokyo Japanese is consistently aligned with the "moraic" part (in the sense of Hayes 1989) of the *second* mora, regardless of whether the second mora is a separate CV syllable or only the second half of a long (CVN or CVV) syllable. However, adjustments to segment durations in the different types of syllable mean that it is more difficult to distinguish between structural and "time pressure" explanations than in the case of Dutch long and short vowels (point *ii* above).

vi. Stress effects on alignment of "phrase accents" in a variety of languages (Grice, Ladd & Arvaniti, 2000; Lickley, Schepman & Ladd, submitted). The F0 minimum in a falling-rising nuclear accent in Dutch and German is aligned with the most prominent postnuclear syllable (e.g. secondary stress; lexical stress of a post-nuclear content word; etc.).

vii. Consistent differences between nuclear and prenuclear accents (Schepman, Lickley & Ladd, submitted; Nibert 2000). In both Dutch and Spanish F0 peaks are aligned earlier in nuclear accents than in prenuclear accents, while the leading F0 valley is unaffected. Caspers & van Heuven found essentially the same effect, if we equate "prenuclear" with their "Type 1 rise" occurring on its own and "nuclear" with their sequence of "Type 1 rise" and "Type A fall". If this equation is valid, Caspers & van Heuven interpret the earlier alignment of nuclear peaks as a case of "time pressure": the peak of the Type 1 rise is pushed earlier by the immediately following Type A fall. A very similar explanation, with "L phrase accent" (Grice et al., 2000) in place of "Type 1 fall", is proposed by Schepman et al.

viii. Consistent small differences of alignment between languages and between varieties of the same language (Atterer & Ladd, in press). In German, rising prenuclear accents are aligned consistently later than those in English and Dutch, and within German, such accents are aligned consistently later in Southern varieties than in Northern varieties. The effects are small but significant, and apply to both the beginning and the end of the rise. Among other things, this means that the consistent alignment of the beginning of the rise with the beginning of the syllable found in Spanish, Dutch, English, Greek and Japanese is not universal. Preliminary data on English (Ladd, Schepman & White, work in progress) similarly suggest that Scottish speakers align pitch accents slightly later than Southern English speakers. More generally, the fact that both the beginning and the end of the rise are affected lends support to Xu's idea that the rise is, at some level of description, a unitary gesture, the alignment of which is specified *as a whole.* This in turn suggests that one possible type of language-specific difference in the alignment of pitch movements may be a matter of what we may call "phasing": we are not so much aligning specific targets at specific places in structure, but aligning whole movements relative to whole syllables. This possibility has important implications for the interpretation of segmental anchoring, as we shall see presently.

3 Segmental anchoring: empirical summary so far

The work reviewed in section 2 can be summarized in two principal points:

1. In all languages studied so far, the duration of a pitch movement is strongly correlated with the duration of the associated segmental material, while the amount of F0 change (the pitch excursion) is unaffected by such differences. This applies whether the differences in duration are brought about by changes in rate, by intrinsic segment durations, or by different syllable structures. In some important sense, the beginnings and ends of pitch movements represent production targets, in a two-dimensional space defined by pitch level and alignment with the segmental string.
2. The details of the correlation in point 1 – the precise alignment of F0 changes with the associated stressed syllable – can vary from language to language and from variety to variety within the same language. These differences can be simple differences of what we just called "phasing" (e.g. "the same" F0 change can be aligned earlier or later, as in Northern German vs. Southern German, point *viii* above) or differences of basic duration (e.g. the F0 change can have longer or shorter duration relative to "the same" syllable structure: in Dutch (point *ii* above) and Greek (section 1) the beginning of a

rising prenuclear accent is aligned in the same way, but the end of the rise is later in Greek than in Dutch).

In my view these two points can be regarded as established empirical findings, which need to be taken into account in any attempt to explain the details of "peak delay" and other aspects of the way pitch features are aligned with the segmental string.

4 Interpreting segmental anchoring

On the face of it, the findings just summarized readily lend themselves to an "autosegmental" interpretation, in which a pitch movement is the phonetic manifestation of a sequence of "tones" aligned with the segmental string in well-defined ways. The alignment of one tone can be at least partially independent of another, as can be seen from the clear language-specific differences in the duration of the same basic kind of pitch movement. The relative independence of the two targets makes problems for the model proposed in Xu's work (point *iv* above), in which whole F0 movements are aligned with whole syllables. However, as the full picture emerges of what is involved in segmental anchoring – in particular, when we consider the implications of the "phasing" differences between languages and language varieties – the autosegmental interpretation becomes less attractive. The problem arises because of the way the autosegmental analysis tries to deal with "alignment" differences.

In an autosegmental model, the obvious way to think about specific patterns of alignment is in terms of the "secondary association" of tones (Pierrehumbert & Beckman, 1988; Gussenhoven, 2000; Grice, Ladd & Arvaniti, 2000). For example, in the case of the Greek prenuclear accents studied by ALM, the basic association is between the L+H accent *as a whole* and the accented *syllable*, but if we want to express the details of the independent alignment of the L and the H in our autosegmental representation, we might say that the L is *secondarily* associated with the left edge of the syllable, and the H is secondarily associated with the following vowel. This is shown in Figure 2. Secondary association seems an especially appropriate way to account for Lickley et al.'s findings on the alignment of the F0 minimum in Dutch falling-rising questions (point *vi*); this whole issue is discussed in detail by Grice et al. (2000).

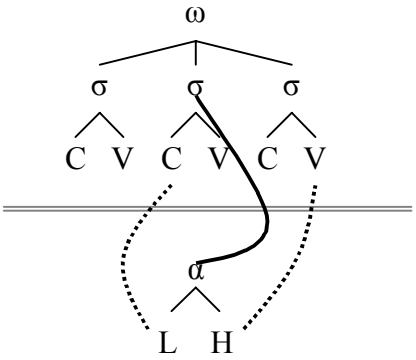


Figure 2. Hypothetical autosegmental representation of the primary association (solid line) of an accent α to a syllable σ and the secondary association (dashed line) of the individual tones L and H to individual segments. See Pierrehumbert & Beckman (1988: 128) for a similar representation of the secondary association of phrasal tones to specific moras in Japanese.

However, using secondary association to represent the phonetic detail of alignment leads to a rapid proliferation of distinct phonological representations for subtly different variations of phonetic detail between languages or between language varieties. For example, to express the

differences among English, Northern German, and Southern German demonstrated by Atterer & Ladd (point *viii*), we would have to associate the initial L tone of the rise with the left edge of the accented syllable (English), the left edge of the accented syllable nucleus (Southern German), and the onset of the accented syllable (Northern German), and posit similar differences for the association of the H tone. Moreover, even if we ignore the fundamental implausibility of assigning different phonological representations to closely related (and phonetically similar) phenomena in closely related languages, we must acknowledge that such an analysis fails to express two important facts: first, that there appears to be a *continuum* of alignment possibilities from English to Southern German; and second, that differences in the alignment of the L and the H are *not independent,* but that the earlier or later alignment somehow applies to the pitch movement as a whole. The first point calls to mind any number of recent demonstrations that language-specific phonetic detail needs to be described quantitatively, not symbolically. The second point suggests that the pitch movement as a whole does have some kind of phonetic unity, as maintained by Xu and as assumed by earlier models such as the one developed at IPO.

With regard to the issue of language-specific phonetic detail, a useful analogy can be drawn to voice onset time (VOT). (Note that VOT, like F0/segment alignment, is phonetically a matter of coordination between laryngeal and supralaryngeal gestures.) Several features of VOT are relevant to our theoretical understanding of alignment. First, we know that languages can differ in the number of categories they distinguish by VOT: many languages have a single contrast between earlier and later VOT (usually described as voiced / voiceless), but some have no contrast (e.g. many Australian languages), and others have a three-member contrast of early, mid, and late VOT (usually described as voiced / voiceless unaspirated / voiceless aspirated). In the same way, some languages clearly have contrasts of F0/segment alignment, while others seem not to. Second, we also know that the phonetic detail of any given language is poorly predicted by the phonology: in two languages with early-late VOT contrasts, the same VOT value may manifest the "early" category in one language and the "late" category in the other. Even in languages with no VOT contrasts we can still make language-specific phonetic generalizations about VOT. Thus it should come as no surprise that, for example, prenuclear H peaks can be aligned differently in Greek and in Dutch. Finally, we know that there do not appear to be favored phonetic "slots" for VOT (say, +15 ms for voiceless unaspirated and +60 ms for voiceless aspirated); rather, the very careful cross-linguistic study of VOT by Cho & Ladefoged (1999) makes clear that average VOT for a given category in a given language can take on any of a continuum of values. We do not yet have comparable data for F0/segment alignment, but the data we do have shows no evidence that there are favored patterns. All of these considerations point to the conclusion that the fine phonetic detail of segmental anchoring is not a matter of secondary association after all, but of quantitative language-specific phonetic detail in the realization of phonological categories.

As for the issue of the "unity" of pitch movements, it is usefully compared to the unity of a diphthong. The phonology of diphthongs (like other complex segments such as affricates and prenasalized stops) is a perennial conundrum, but at some level of description we are dealing with a specified movement from one set of formant values to another. These sets of values can be seen as phonetic targets, and the movement from one target to another in any specific context will be affected by speech rate, prosodic structure, surrounding consonants, and so on. The details are beyond the scope of this short paper, but two things are clear: first, the unitary nature of the articulatory gestures involved in a diphthong does not make the starting and ending targets irrelevant to the diphthong's phonological characterization; and second,

there is no reason to assume that the time-course of the diphthong gesture is somehow universally locked to some other specific articulatory event or events – in fact, some recent work (notably Geumann & Hiller, 1996; Scobbie, Turk & Hewlett, 1999) suggests that the temporal details of the formant movements in a diphthong can have language-specific phonological significance.

If we draw an analogy between pitch movements and diphthong gestures, we can see two useful things more clearly. First, we can see that there is no contradiction between treating the pitch movement as a phonetic gesture and describing it phonologically in terms of its endpoints (e.g. L+H). This is important for reconciling the work of e.g. the IPO tradition with work based on autosegmental phonology. Second, we can acknowledge the phonetic unity of pitch movements without also assuming (as Xu and others have tended to do) that pitch movements are coordinated with their associated syllables in a universal deterministic way. In his work, Xu has started with the indisputable fact that segmental syllables and pitch movements are *coordinated* in time, and has elaborated it into a model in which the beginnings and ends of pitch movements and syllables are expected in many cases to be *simultaneous*. This requires him, among other things, to gloss over differences like those between Greek and Dutch prenuclear accents (e.g. Xu & Wang, 2001, which emphasizes that "tones are tightly aligned with host syllables despite the variations" (p.328) and appears to suggest that variations from simultaneity are all to be explained in biomechanical terms). But there is no reason to think that the language-specific differences summarized in Section 2 are spurious, so we must find a way of accommodating them in our theoretical understanding, just as we must find a way to understand findings about diphthongs like those of Geumann & Hiller or Scobbie et al.

In the end, though, clear analogies will only get us so far. I think I am expressing a view appropriate to a Nooteboom Festschrift when I say that further studies of speech production and perception are going to be crucial for advancing our understanding of segmental anchoring. The key problem for all models of F0/segment alignment is that *segmental anchoring implies lookahead.* In a simple model of pitch movements like the one incorporated into the IPO theory ('t Hart et al., 1990: 73), the production mechanism only needs an anchor point and a specification of the pitch excursion (e.g. "starting now, raise F0 for 120 ms at a rate of 50 semitones per second"), and Xu's model seems to aim at a similar conceptual simplicity. But a more complicated model will be required if we are going to deal with segmental anchoring, because the specification of the pitch excursion will be something like "raise F0 by 6 semitones, starting now and finishing at the onset of the next vowel". In order to know how fast to raise the pitch, the model needs to anticipate how long it will take to get to the specified finishing point. This is the kind of problem that models of segmental articulation – including diphthongs – have long had to deal with. Cross-fertilization from that work can only help in developing our understanding of pitch movements.

Author's note

Although it may not be obvious to the reader, there is a connection between Sieb Nooteboom and the work reported here. My very first public presentation of the findings that form the basis of this paper – the findings that were later published as Arvaniti, Ladd & Mennen (1998) – was in Utrecht in early 1996. Sieb had introduced my talk and was at the front of the audience. I recall watching the growing astonishment on his face as I presented our experimental evidence that the duration of pitch movements is highly dependent on the segmental material that they accompany. For someone like Sieb who was raised on the unitary "prominence-lending pitch movements" of the classic IPO approach to intonation, the

autosegmental interpretation of pitch movements was hard to accept, yet Sieb readily admitted that the Arvaniti et al. results were unexpected, and certainly looked like support for the autosegmental view. So it seems appropriate to come back, nearly a decade later, and give, as my contribution to Sieb's Festschrift, a summary of what followed those first findings. Somewhat ruefully, I note that it is appropriate to round off the summary with an acknowledgement that the autosegmental interpretation was not as strong as it seemed in 1996, and that maybe the idea of unitary prominence-lending pitch movements has something to recommend it after all.

References

Arvaniti, A., & Ladd, D. R. (1995). Tonal alignment and the representation of accentual targets. In *Proceedings of the 13th International Congress of Phonetic Sciences, Stockholm* (vol. 4, pp. 220-223).

Arvaniti, A., Ladd, D. R., & Mennen, I. (1998). Stability of tonal alignment: the case of Greek prenuclear accents. *Journal of Phonetics, 26*, 3-25.

Atterer, M., & Ladd, D. R. (in press). On the phonetics and phonology of "segmental anchoring" of F0: evidence from German. To appear in *Journal of Phonetics,* 2004.

Cho, T. & Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *Journal of Phonetics, 27*, 207-229.

Dilley, L., Ladd, D. R., & Schepman, A. (in press). Alignment of L and H in bitonal pitch accents: testing two hypotheses. To appear in *Journal of Phonetics,* 2004.

Face, T. L. (2002). *Intonational marking of contrastive focus in Madrid Spanish.* Munich, Lincom Europa.

Fujisaki, H. (1983). Dynamic characteristics of voice fundamental frequency in speech and singing. In P. F. MacNeilage, (Ed.), *The Production of Speech* (pp. 39-55). New York: Springer.

Geumann, A. & Hiller, M. (1996). Diphthong dynamics in Swabian. *Journal of the Acoustical Society of America, 100*, 2687.

Grice, M., Ladd, D. R., & Arvaniti, A. (2000). On the place of "phrase accents" in intonational phonology. *Phonology, 17*, 143-186.

Gussenhoven, C. (2000). The boundary tones are coming: on the nonperipheral realization of boundary tones. In M. Broe & J. Pierrehumbert (Eds.), *Papers in Laboratory Phonology V* (pp. 132-151). Cambridge: Cambridge University Press.

't Hart, J., Collier, R., & Cohen, A. (1990). *A perceptual study of intonation: An experimental-phonetic approach*. Cambridge : Cambridge University Press.

Hayes, B. (1989). Compensatory Lengthening in Moraic Phonology. *Linguistic Inquiry, 20*, 253-306.

Ishihara, T. (2003). A phonological effect on tonal alignment in Tokyo Japanese. In M. J. Solé, D. Recasens & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences; Barcelona, 3-9 Augustus 2003* (vol. 1, pp. 615-618).

Ladd, D. R. (1996). *Intonational Phonology*. Cambridge: Cambridge University Press.

Ladd, D. R., Faulkner, D., Faulkner, H., & Schepman, A. (1999). Constant "segmental anchoring" of F0 movements under changes in speech rate. *Journal of the Acoustical Society of America, 106*, 1543-1554.

Ladd, D. R., Mennen, I., & Schepman, A. (2000). Phonological conditioning of peak alignment of rising pitch accents in Dutch. *Journal of the Acoustical Society of America, 107*, 2685-2696.

Ladd, D. R., & Schepman, A. (2003). "Sagging transitions" between high pitch accents in English: experimental evidence. *Journal of Phonetics, 31*, 81-112.

Lickley, R., Schepman, A., & Ladd, D. R. (submitted). Lab speech is real speech: the case of Dutch falling-rising questions. Submitted to *Language and Speech*, October 2003. Under revision.

Nibert, H. (2000). *Phonetic and Phonological Evidence for Intermediate Phrasing in Spanish Intonation.* PhD Dissertation, University of Illinois.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. PhD Dissertation, MIT.

Pierrehumbert, J., & Beckman, M. E. (1988). *Japanese Tone Structure.* Cambridge MA: MIT Press.

Prieto, P., van Santen, J., & Hirschberg, J. (1995). Tonal alignment patterns in Spanish. *Journal of Phonetics, 23*, 429-451.

Schepman, A., Lickley, R., & Ladd., D. R. (submitted). Effects of vowel length and "right context" on the alignment of Dutch nuclear accents. Submitted to *Journal of Phonetics,* November 2003.

Scobbie, J. M., Turk, A., & Hewlett, N. (1999) Morphemes, phonetics, and lexical items: The case of the Scottish Vowel Length Rule. In J.J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A.C. Bailey

(Eds.), *Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, August 1-7, 1999* (vol. 2, pp. 1617-1620).

Silverman, K. E. A. & Pierrehumbert, J. (1990). The timing of prenuclear high accents in English. In J. Kingston and M. Beckman (Eds*.) Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech* (pp. 71-106). Cambridge: Cambridge University Press.

Taylor, P. A. (2000) Analysis and synthesis of intonation using the Tilt model. *Journal of the Acoustical Society of America, 107*, 1697-1714.

Xu, Y. (1998) Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica, 55*, 179-203.

Xu, Y. (1999) Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics, 27*, 55-105.

Xu, Y. & Wang, Q. E. (2001) Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication, 33*, 319-337.

Phonetics and Phonology then, and then, and now *

John J. Ohala

University of California, Berkeley

Phonetics attempts to describe and understand how speech is produced and perceived; phonology attempts to understand the patterning — in general, the behavior — of speech sounds in particular languages and in all languages. Is phonetics part of phonology? This straightforward question has received various answers at different points in the history of linguistics. In this paper I attempt to document that for the two centuries starting approximately with the eighteenth century, phonetics was well integrated into linguistics but that around the start of the 20th century phonetics and phonology were estranged, at least in some cases. During the second half of the 20th century there began a trend, continuing today, to re-integrate phonetics and phonology.

The history I give is admittedly selective, interpretive, and possibly biased. Most if not all histories are like this. Whoever may disagree with this history is free to — indeed, obliged to — present and document their own interpretive history.

1 The two 'phonetics': taxonomic & scientific

I believe it is possible to distinguish two forms of phonetics, *taxonomic* and *scientific*, and historically their place in phonology has been different. Taxonomic phonetics provides two basic tools for the dealing with speech sounds: first, uniformity in naming and classifying speech sounds, and, second, transcribing them. Although the attempt to arrive at a uniform system to accomplish these essential functions dates back many centuries (Kemp, 1994), as far as the transcription is concerned, a widely adopted standardization was achieved in the late 19th century with the rise of the International Phonetic Association, founded in 1886 by Paul Passy and with the eventual codification of the International Phonetic Alphabet (IPA) (MacMahon, 1994). Although it has seen additions and modifications, it has remained in essence unchanged since then. Of course, if it is to remain useful in providing a lingua franca when describing and classifying speech sounds of the world's languages, it should not change in radical ways over time. In spite of some imperfections, the International Phonetic Alphabet (IPA) has been maintained in essentially the same form since its introduction and this has provided the basis for a vocabulary and a system with which to describe individual speech sounds as well as phonetic and phonological universals.

The other form of phonetics, which I call 'scientific phonetics' seeks to understand how speech works at all levels from the brain of the speaker to the brain of the hearer. This has a long history but unlike taxonomic phonetics, is continually in a state of flux, constantly taking in new data, new methods, new theories, and discarding older ones found to be less competitive. Although ideally there is a search for some convergence among theories, before that is achieved there is a healthy debate between adherents of competing theories (Fowler, 1996; Ohala, 1996). If the arguments against a particular theory are overwhelming, there is

---

* This paper is an adaptation of Ohala (1991).

no hesitation in abandoning it completely [e.g., see Ladefoged (1967) arguing against Stetson's (1928) hypothesized "breath pulses"; van den Berg (1958) arguing against Husson's (1950) neurochronactic theory of vocal cord vibrations; on the latter controversy, see also Abramson (1972)].

In any case, it is in scientific phonetics where most of the work of phonetics lies. This is where theories are formulated, statistical analysis of results performed as well as controlled observations, calibrations and all the other characteristics of traditional scientific procedures. This is what one finds presented at phonetics conferences and congresses and in the phonetics journals; it is where most of the "action" is.

Besides these typical external differences between taxonomic and scientific phonetics there is a profound philosophical difference. The 'scientific' approach implies, as do all other sciences since the Renaissance that any given theory, including whatever one believes most fondly, may be erroneous but that by gathering data in a rigorous way such error may be minimized or avoided. In contrast, taxonomic phonetics thrives through conformity.

Phonology unquestionably embraces taxonomic phonetics — at the very least it provides the vocabulary for stating phonological generalizations. The question is to what extent it incorporates scientific phonetics. In the following sections I explore this question by briefly reviewing cases of what may be regarded as phonological studies over the centuries, and finally looking for long-range trends.

2 Scientific phonetics in the 17th, 18th and 19th centuries: phonetics integrated with phonology

I give examples where traditional phonological (linguistic) questions are offered phonetically-based answers *or* where the same individual is equally productive in scientific phonetics and phonology in general.

Johan Conrad Amman (1669-1724) was a Swiss physician practicing in the Netherlands. He wrote of his attempt to teach speech to a deaf person, *The talking deaf man: or, a method proposed whereby he who is born deaf may learn to speak* (Amman, 1694), and *Dissertatio de loquela* (Amman, 1700). He made some original observations about speech articulations including a characterization of how laterals are produced (namely, that the lateral channel may lie in the buccal sulcus, not necessarily in the space enclosed by the teeth). He noticed nasals assimilating to the place of following stops in connected speech and he proposed a "natural" hierarchical classification of the features of speech, e.g., in assimilation and in pathological speech, substitution of one sound for another involves features at the lowest strata of the hierarchy, e.g., place, not at the highest such as manner; thus substitutions of consonant for vowel, or nasal for fricative do not occur.

There were several others in the 17th c. who claimed to be able to teach the deaf to speak or who recommended procedures by which this might be done, e.g., van Helmont (1667), Holder (1669), and Wallis (1653). Wallis, in particular, exemplified an admirable union of a scientific approach to description of speech sounds with what would be regarded as phonological observations.

Wolfgang von Kempelen (1734-1804), was a lawyer, physicist, engineer, and student of language in the Austro-Hungarian empire. He conceived of and built what is regarded as the first mechanical speech synthesizer capable of producing connected speech (even though one of the 'components' of the system was the hand of the operator which helped to shape the resonating cavity that produced the different vowels). He published a detailed description of his device and his experience with it in 1791 (*Mechanismus der menschlichen Sprache*,

Vienna). His work was influential for more than a century following that and was an inspiration to, among others, Alexander Graham Bell.

The first few chapters are devoted to a review of the existing literature on speech production and to the phonology of various languages, esp. Hungarian. He showed how speech sound *behave* in languages, i.e., the *phonology* of Hungarian.

Erasmus Darwin (1731-1802), product of the Enlightenment, was a scientist, philosopher, promoter of liberal values (including the education of women). He constructed an elementary speech synthesizer (not unlike von Kempelen's, but much simpler) (Darwin, 1803). He proposed a system of 13 unary features by which to describe any and all speech sounds including, notably, the voiceless lateral [l̥] of Welsh. He also conducted what may be the first instrumental phonetic study on a live, intact speaker: he inserted cylinders of tin foil into his mouth to determine by the indentations made on them by the tongue where the different vowels were articulated.

Robert Willis (1800-1875), was a Cambridge professor of mechanics (engineering we would call it today). In his work "On the vowel sounds" (1830) he specified quantitatively the vocal tract resonances of vowels—a single resonant frequency for each vowel—and claimed that the major determinant of these resonances was vocal tract length. He demonstrated this with a uniform, cylindrical, tube whose resonating space was varied in length by putting the sound source (the excitation) on a piston-like structure that could move up and down within the cylindrical resonator. He suggested that with some refinement of his method, it should be possible to provide "philologists with a correct measure for that shade of differences in the pronunciation of the vowels by different nations."

Today we might fault his claims that there is a *single* characteristic resonance for each vowel sound (as opposed to being differentiated by *multiple* — at least 3 — different resonances (F1, F2, F3). But his single resonance may correspond to the F2' of Fant & Risberg (1963) and Chistovich & Lublinskaja (1979), or to the most characteristic resonance of the vowels by those concerned with the auditory transform of vowels.

But the philologists (at least one) paid heed to Willis! **T. Hewitt Key** (1799-1875), trained in medicine and mathematics, who became the first professor of Latin and then the professor of comparative philology at London University (now University College), published a paper in the Transactions of the Philological Society (of London) in 1852 entitled "On vowel assimilation, especially in relation to Professor Willis' experiment on vowel sounds." He tries to explain vowel harmony and umlaut by invoking Willis' notion that vocal tract length is the main determinant of vowels' characteristic resonance. This explanation might not be accepted today but it is the willingness to apply physical phonetics to philological questions that is admirable.

Hermann Grassmann (1809-1877) is one of the few linguists included in the *Dictionary of Scientific Biography,* but his inclusion is mainly due to his contributions to mathematics (namely his general calculus for vectors). However, within linguistics (or philology) he is primarily honored as the discoverer of "Grassmann's Law" (dissimilation of aspirated consonants in Greek and Sanskrit), thus accounting for a major set of exceptions to Grimm's Law. Less well known is that he also determined the resonant frequencies of spoken vowels using purely auditory analysis (1854; thus anticipating Helmholtz's results by 9 years).

Karl Verner (1846-1896) is famous for his resolution in 1875 of, until then, one of the thorniest exceptions to Grimm's Law, the variation in voicing of the medial obstruents in Gothic. He showed that the different reflexes were the effects of different accent on the adjacent vowels.

In his later years studied accent phonetically. He obtained an early Edison cylinder phonograph and devised on his own an elaborate system for magnifying the curves and projecting them onto a wall where they could be traced on paper. He did not obtain any results he thought worthy of publication. We know of this work only through the posthumous publication of his correspondence with the Finnish phonetician Hugo Pipping, to which Eli Fischer-Jørgensen has called attention (Fischer-Jørgensen, 1967).

Paul Passy (1859-1940) was founder of the International Phonetic Association (in 1886). This laid the foundation for today's taxonomic phonetics. Nevertheless his dissertation of 1890 offered phonetic explanations for sound change including, e.g., a cogent account of the aerodynamic factors favoring voiceless in obstruents.

Abbé J.-P. Rousselot (1846-1924) is widely regarded as the 'father of experimental phonetics'. With Rosapelly (1876) he pioneered and refined the use of the kymograph for study of speech articulations. His dissertation (1891) was a philological survey of the sound changes that gave rise to the contemporary pronunciation and an instrumental phonetic study of the factors that may have caused them.

Charles Rosapelly was an early recruit in the laboratory of E. J. Marey, who did pioneering research on time-varying physiological events: heart beat, walking, birds' flying, etc. Rosapelly (1876) described early work with the kymograph and commented that

> L'importance de ces études semble grande au point de vue des linguists, dont la science chaque jour pour precise tend à prendre pour point de depart une étude expérimentale. L'étude comparée des différentes langues et celle des transformations successive que chacune d'elles a subies dans sa formation ont permis, en effet, de saisir certaines lois qu'on pourrait appeler physiologiques et qui ont preside à l'évolution du langage.

3 Phonetics from ca. 1900 to ca. 1950: the estrangement of phonetics and phonology

Just when phonetics was starting to make significant advances in the understanding of the physical nature of speech, there is evidence that traditional phonology and linguistics started to distance itself from phonetics.

Henry Sweet (1845-1912) was the founder of the British School of Phonetics. He was the inspiration (in part) for G. B. Shaw's "Henry Higgins" in his play *Pygmalion*, from which the musical *My Fair Lady* was adapted. Sweet raised the standards of phonetic description. With Passy and Viëtor he helped to establish the taxonomic system (the descriptive and classificatory framework) used today in phonetics, as well as the IPA for transcription. But he had a harsh assessment of instrumental phonetics (Sweet, 1910):

> The claims of instrumental phonetics have been so prominently brought forth of late years that they can no longer be ignored, even by the most conservative of the older generation of phoneticians. But it is possible to go too far the other way. Some of the younger generation seem to think that the instrumental methods superseded the natural ones in the same way as the Arabic superseded the Roman numerals. This assumption has had disastrous results. It cannot be too often repeated that instrumental phonetics is, strictly speaking, not phonetics at all. It is only a help; it only supplies materials which are useless till they have been tested and accepted from the linguistic phonetician's point of view. The final arbiter in all phonetic questions is the trained ear of a practical phonetician: differences which cannot be perceived must—or at least may be—ignored; what contradicts the trained ear cannot be accepted.

Rousselot (1897:1) also noticed the "distance" between experimental phonetics and linguistics:

> … les procédés des sciences expérimentale sont assez étrangers aux linguists. Une sorte de terreur superstitieuse s'empare d'eux dès qu'il s'agit de toucher au méchanisme le plus simple. Il fallait donc leur montrer que la difficulté est moindre qu'ils ne se la figurant et leur faire entrevoir le champ immense que l'expérimentation ouvre devant eux.

In the translator's preface to Holger Pedersen's *Sprogvidenskaben i det nittende aarhundrede*, [*Linguistic science in the nineteenth century*], Spargo (1931:viii) writes:

> … one important feature of the work which should be mentioned is the striking role assigned to the study of phonetics in increasing our knowledge of linguistics. It is shown clearly that every important advance during the last century and a quarter was made by a scholar who attacked his problem from the phonetic side.[2] Surely this fact has its importance for the future of linguistic study, and suggests that the indifference to phonetics in many of the graduate schools in the United States is an evil presage for future progress.

One suspects this distance between instrumental phonetics and linguistics arose out of misunderstandings and misgivings that the linguistically- and philologically-trained researchers had towards a methodology that was unfamiliar to them.

Similarly, many traditionally-trained anthropologists had misgivings about modern work determining the hominid "family tree" structure using the techniques of microbiology, i.e., measuring the degree of similarity between DNA molecules and other biologically important molecules.

But remember: what defines a field are its questions, not its methods; one uses whatever methods get us the answers; there is no glory to researchers who get less-than-satisfactory answers to questions because they had a distaste for the methods that would yield the answers.

But perhaps the greatest wedge between experimental phonetics and linguistics was driven by structuralism: the focus not on the substance of speech but on the relations, the contrast between speech sounds. This was brought about by the Prague School which had great influence within phonology. In effect they banished experimental phonetics from linguistics.[3]

Trubetzkoy (1933) wrote,

> La phonétique actuelle se propose d'étudier les facteurs matériels des sons de la parole humaine: soit les vibrations de l'air qui leur correspondent, soit les positions et les mouvements des organes qui les produisant. … Le phonéticien est nécessairement atomiste our individualiste … Chaque son de la parole humaine ne peut être étudie qu'isolement, hors de tout rapport avec les autres sons de la même langue.

A similar stereotype applied to astronomy would characterize it as merely finding and cataloguing stars. But this would ignore cosmology, astrophysics and, in general, any attempt to generalize about the birth, development, and death of stars, the formation of galaxies, the origin of the universe. Trubetzkoy commits the fallacy of equating the immediate, visible, object of study as the ultimate object of study.

Phonetics, then and now, studied the physical (and psychological) aspects of speech sounds in order to understand how speech works, including the contrastive aspect that Trubetzkoy focused on.

---

[2] He was probably referring of the phonetic decomposition of sounds as exemplified in the diachronic work of, among others, Rask, Grimm, von Raumer, Grassmann, Verner, Brugmann, and Saussure (on the IE "laryngeals").

[3] This is not to say that all members of the Prague school endorsed this view; see Laziczius (1966); nor that some phonetics research addressing traditional linguistic questions was prevented from being done; and it is also not claimed that other schools and individuals outside the Prague School didn't express similar views to them.

To give a more balanced history, it must be recognized that there were also phoneticians at this time whose approach might best be characterized as "positivist", for whom the physical aspect of the speech sounds tended to be the dominant focus of study, e.g., E. W. Scripture and Guilo Panconcelli-Calzia (Kohler, in press).

4 Phonetics since then: the present day: phonetics again becoming integrated with phonology

With a few exceptions, the distance between scientific phonetics and phonology continued up to approximately the mid-20th century. Pivotal developments contributing to reducing the distance were:

- The synthesis of speech from phonemic or other phonological input (Klatt, 1987; Maxey, 2002). This included the Haskins' claim that they had "found" the invariance of phonemes underlying their phonetic variants. (A claim subsequently qualified or even retracted.)
- The collaboration between Jakobson, Fant & Halle in proposing the acoustically-defined "distinctive features" (in 1952).

In spite of their short-lived popularity, the Jakobson-Fant features demonstrated that some of the linguistic functions of speech sounds, e.g., their contrastiveness and some certain phonological behavior (e.g., phonotactics), could be explained by invoking their acoustic-auditory nature — as discovered by experimental phonetics.

Since then, at least, courses in phonetics and phonetics laboratories have had a relatively secure home within departments of linguistics.[4]

Today "Linguistic Phonetics", "Experimental Phonology", "Laboratory Phonology" and similar movements are represented in the literature and have regular conferences. This success is based not on a fad but on 'existence proofs' — demonstrations of the relevance of physical and psychological aspects of speech for explaining sound patterns in language, the traditional concern of phonology.

5 Sieb Nooteboom's place in this history

Sieb Nooteboom has furthered the rapprochement between experimental phonetics and rest of linguistics. This has been done by providing *existence proofs* of the benefits of phonetic and psycholinguistic studies for answering linguistic questions.

In such diverse research areas as speech production and perception, speech technology, prosody, psycho-phonology, speech errors, addressing such fundamental problems as the nature of the units of speech, the role of feedback in speech production and many others. He has enlarged and enriched phonetics by demonstrating the utility of new methods and exploration of new research domains. He has a secure and honored place in the history of the integration of phonetics and phonology!

References

Abramson, A. S. (1972). Comments to: Lafon's "Essai sur la physiologie du son laryngé". In A. Rigault & R. Charbonneau (Eds.), *Proceedings 7th International Congress of Phonetic Sciences* (pp. 25-26). The Hague: Mouton.

[4] There are exceptions: certain leading universities in the U.S.A., famous for their linguistic work, do not require a course in phonetics for any of their academic degrees.

Amman, J. C. (1694). *The talking deaf man: or, a method proposed whereby he who is born deaf may learn to speak*. London: Tho. Hawkins.
Amman, J. C. (1700). *Dissertatio de loquela*. Amsterdam: J. Wolters.
van den Berg, J. W. (1958). Myoelastic-aerodynamic theory of voice production. *Journal of Speech and Hearing Research*, *1*, 227-244.
Chistovich, L.A., & Lublinskaja, V.V. (1979). The 'center of gravity' in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing Research, 1*, 185-195.
Darwin, E. (1803). *The temple of nature*. London: J. Johnson.
Fant, G. & Risberg, A. (1963). Auditory matching of vowels with two formant synthetic sounds. *STL-QPSR, 4*, 7-11.
Fischer-Jørgensen, E. (1967). A sketch of the history of phonetics in Denmark until the beginning of the 20th century. *Annual Report Institute of Phonetics, University of Copenhagen, 13*, 135-169.
Fowler, C. A. (1996). Listeners do hear sounds not tongues. *Journal of the Acoustical Society of America, 99,* 1730-1741.
Grassmann, H. (1854). Leitfaden der Akustik. Programm des Stettiner Gymnasiums.
Helmont, F. M. van (1667). *Alphabeti verè Naturalis Hebraici brevissima delineatio*. Sulzbach: Lichtenthaler.
Holder, W. (1669). Elements of speech: An essay of inquiry into the natural production of letters. London: J. Martyn.
Husson, R. (1950). Étude des phénomènes physiologiques et acoustiques fondamentaux de la voix chantée. *Éditions de La revue scientifique, A, 2334* (3206), 1-93.
Kemp, J.A. (1994). Phonetic transcription: History. In R. E. Asher & J. M. Y. Simpson (Eds.), *The encyclopedia of language and linguistics* (pp. 3040-3051). Oxford: Pergamon Press.
Kempelen, W. von (1791). *Mechanismus der menschlichen Sprache*. Vienna: J. B. Degen.
Key, T. Hewitt. (1852). On vowel assimilation, especially in relation to Professor Willis' experiment on vowel sounds. *Transactions of the Philological Society* [London].
Klatt, D. H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America, 82*, 737-793.
Kohler, K. (in press) Beyond Lab Phonology: The phonetics of speech communication.
Ladefoged, P. (1962). Subglottal activity during speech. In A. Sovijarvi and P. Aalto (Eds.), *Proceedings 4th International Congress of Phonetic Sciences, Helsinki, 1961* (pp. 73-91). The Hague: Mouton.
Ladefoged, P. (1967). *Three areas of experimental phonetics*. Oxford: Oxford University Press.
Laziczius, G. (1966). *Selected Writings of Gyula Laziczius* (edited by Thomas A. Sebeok). The Hague: Mouton.
MacMahon, M. K. C. (1994). International Phonetic Association. In R. E. Asher & J. M. Y. Simpson (Eds.), *The encyclopedia of language and linguistics* (pp. 1730-1731). Oxford: Pergamon Press.
Maxey, H. D. (2002). *Smithsonian Speech Synthesis History Project*. Available: http://www.mindspring.com/~ssshp/ssshp_cd/ss_home.htm
Ohala, J. J. (1991). The integration of phonetics and phonology. In *Proceedings of the XIIth International Congress of Phonetic Sciences, Aix-en-Provence, 19-24 Aug 1991* (Vol. 1, pp. 1-16).
Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America, 99*, 1718-1725.
Passy, P. (1890). *Étude sur les changements phonétiques*. Paris: Firmin-Didot.
Pedersen, Holger. (1924). *Sprogvidenskaben i det nittende aarhundrede*. Copenhagen: Gyldendalske Boghandel Nordisk Forlag.
Rosapelly, C. L. (1876). Inscriptions des mouvements phonétiques. *Physiologie Expérimentale; Travaux du Laboratoire de M. Marey, 2*, 109-131. (Paris: G. Masson.).
Rousselot, P.-J. (1891). *Les modifications phonétiques du language étudiées dans le patois d'une famille de Cellefrouin (Charante)*. Paris: H. Welter. (Extrait de la Revue de Patois Gallo-Romans, 1891.)
Rousselot, P.-J. (1897). *Principes de phonétique expérimentale*. Paris.
Spargo, J. W. (1931). Preface to translation from Danish of Holger Pedersen's *Linguistic science in the nineteenth century*. Cambridge: Harvard University Press.
Stetson, R. H. (1928). *Motor phonetics. A study of speech movements in action*. Archives néerlandaises de phonétique expérimentale; 3.
Sweet, H. (1910). Phonetics. In *Encyclopædia Britannica* (11th ed.).
Trubetzkoy, N. (1933). La phonologie actuelle. *Journal de Psychologie, 1-4*, 227-246. [Theme issue on Psychologie du Langage, in French].
Wallis, J. (1653). *Grammatica lingua Anglicanae* [with an appendix: De loquela, sive sonorum formatione, tractus grammatico-physicus.] London: Leon Lichfield.

Willis, R. (1830). On the vowel sounds, and on the reed organ-pipes. *Transactions of the Cambridge Philosophical Society*, *3*, 229-268.

Expanding Phonetics

Louis C.W. Pols

University of Amsterdam

Abstract

The *optimistic title* of my contribution to this Festschrift is supposed to underline the substantial growth in Phonetic Sciences as for instance reflected by comparing the program size and the number of participants of the 1st International Congress of Phonetic Sciences (ICPhS) organized in 1932 in Amsterdam with that of the 10th ICPhS that was again organized in the Netherlands, this time in Utrecht in August 1983, and then with that of the 15th and most recent one in September 2003 in Barcelona. It is also meant to be a somewhat *cynical title*, reflecting the fact that the Chair of Phonetics in Utrecht, that was so capably occupied by prof. Antonie Cohen and then by prof. Sieb Nooteboom, will not automatically be continued after Sieb's retirement. Worldwide, the multidisciplinary field of Phonetics is expanding in many directions, whereas in the Netherlands the number of students that choose this specialization is declining. However, optimists emphasize that perhaps only the name tag differs and that actually Phonetics is an indispensable element in the web of interdisciplinarity surrounding Linguistics, Experimental Psychology, Speech and Language Technology, Signal processing, ENT, Audiology and the like.

1 Introduction

The retirement of Sieb Nooteboom as full professor in Phonetic Sciences at Utrecht University is a suitable occasion to talk about growth and decline of Phonetics. As players in this field, we all see that worldwide there is a strong upsurge of activities in Phonetics and Speech Communication, as least if we consider the number of conferences and workshops, and the variety of participants in those events, as well as the variety and multitude of ongoing research projects, and the many scientific journals and books in our field, as proper indicators of that. However, it is also realistic to acknowledge that the actual situation in the Netherlands is not so promising at all. Retiring professors are not replaced, the independence of Phonetics as a discipline is under siege in the new bachelor-master system, and students study (General) Linguistics or Informatics, rather than Phonetics. Phonetics is at best just a (small) part of those bachelor programs. We can only hope to attract a sufficient number of students in the future to our master specializations like speech communication, speech and language technology, or speech and language pathology.

2 Growth in Phonetics according to ICPhS and otherwise

A strong indication of the liveliness of the Phonetics community, and actually of any scientific community, is the way its members meet and work together internationally, publish in the open literature, train their students and contribute to the needs of society. This is why I first of all will give some qualifications of the 1st, the 10th and the 15th International Congress of Phonetic Sciences (ICPhS).

2.1 First ICPhS in 1932 in Amsterdam

With Jac. van Ginneken as President and Louise Kaiser as Secretary, the first ICPhS was organized in July 1932 in Amsterdam. According to the 221-pages proceedings book, the topics supposed to be covered at that congress were:

- physiology of speech and voice (experimental phonetics in its strict meaning);
- study of the development of speech and voice in the individual; their evolution in the history of mankind; the influence of heredity;
- anthropology of speech and voice (racial differences in the articulation basis and the pitch of the voice in different peoples);
- phonology;
- linguistic psychology;
- pathology of speech and voice (clinical experimental phonetics);
- comparative physiology of the sounds of animals;
- musicology.

There were 136 participants from 16 different countries. The program contained 43 plenary papers and 24 demonstrations. Some of the famous people present there were Daniel Jones (London) 'The theory of phonemes, and its importance in Practical Linguistics', Sir Richard Paget (London) 'The Evolution of Speech in Men', R.H. Stetson (Oberlin) 'Breathing Movements in Speech', Prince N. Trubetzkoy (Wien) 'Charakter und Methode der systematischen phonologischen Darstellung einer gegeben Sprache', and E. Zwirner (Berlin-Buch) 'Phonetische Untersuchungen an Aphasischen and Amusischen' and 'Quantität, Lautdauerschätzung und Lautkurvenmessung (Theorie und Material)'.

2.2 10th ICPhS in 1983 in Utrecht

After 2. London (1935), 3. Ghent (1938), 4. Helsinki (1961), 5. Münster (1964), 6. Prague (1967), 7. Montreal (1971), 8. Leeds (1975), and 9. Copenhagen (1979), the 10th ICPhS came again to the Netherlands in 1983, this time to Utrecht, under the presidency of Antonie Cohen. The organization of this conference was a truly national endeavor with an organizing committee consisting, next to the president, of Marcel van den Broecke as secretary general, and Florien Koopmans-van Beinum, Sieb Nooteboom and Louis Pols as members. There were some 570 participants from 44 countries, with no fewer than 121 participants from the organizing country. At this 5½-day conference, with one day for excursions, 22 sections were distinguished, from 'Acoustics of Speech' and 'Speech Synthesis' to 'Perception of Phonemes' and 'Prosody'. Each section had one to four sessions during the week, with each time up to eight papers per session. During the conference up to 20 sessions could take place in parallel! People that participated still vividly remember the crowds moving up and down the escalators between sessions. There were also two working groups and six plenary morning sessions plus six symposia (from 'Semantics, Syntax, and Prosody' to 'Speech Recognition') and only one poster session with demonstrations over lunch. Eli Fischer-Jørgensen and Gunnar Fant gave the opening lectures. A book with all abstracts was already available well before the start of the congress. Full proceedings were not yet published, but only a book with a selection of individual contributions (van den Broecke & Cohen, 1984).

2.3 15th ICPhS in 2003 in Barcelona

The series continued with 11. Tallinn (1987), 12. Aix-en-Provence (1991), 13. Stockholm (1995), and 14. San Francisco (1999), whereas the 15th and most recent one took place in August 2003 in Barcelona. Number 16 will take place in Saarbrücken in 2007. In Barcelona the number of participants was about 950 from 51 different countries (36 participants from

the Netherlands). At this six-day conference, with still one day for excursions, there were five invited plenary lectures, 22 thematic symposia organized by convenors, and 55 oral (337 papers) and 18 poster (427 papers) sessions. On a typical conference day there were six time slots of one and a half hour each. At 9 o'clock the day started with a plenary lecture or five oral sessions in parallel, then after the coffee break again five oral sessions, then two big poster sessions in parallel over lunch. After lunch the program continued with five oral sessions or symposia, then again two big poster sessions, followed by five oral sessions or a plenary lecture. The growth in number of posters (427) is astounding. The actual number of papers was around 815, of which 31 had a first author from the Netherlands. The printed proceedings (three volumes) contained over 3200 pages. Fortunately, all this information is now also available on CD-ROM, and in a searchable form as well! This was probably the last ICPhS were (only) extended abstracts were reviewed, the Permanent Council of ICPhS decided that in Saarbrücken a full-and-final paper review procedure will be applied, as is customary nowadays with most major speech conferences. Since the biennial series of Eurospeech (odd years) and ICSLP (even years) conferences have started in 1989 and 1990, respectively, attention for speech technology has diminished at ICPhS.

It is my impression that the growth in terms of number of papers and number of participants at ICPhS has stabilized, also for practical reasons. Conferences with more than 1000 participants become unpractical and unmanageable, and together these participants will produce over 800 papers, which again is a lot to consume. Furthermore, specialized workshops become more and more popular, and many scientists prefer this style of discussing a specialized topic with a smaller and more homogeneous audience. This, of course, does not imply that the number of phoneticians and/or speech scientists has stabilized as well. The International Phonetic Association (IPA), which is now also responsible for organizing ICPhS had, according to the written report of the most recent General Meeting in Barcelona (JIPA 33(2), 2003, 275-277), by the end of 2003, slightly more than 1000 members. Also, the International Speech Communication Association (ISCA) presently has about 1350 members. I guess that the number of speech scientists worldwide is at least ten times larger. This does not yet make it a big community, but still a rather influential one. The number of serious books published in Phonetics and related fields must be at least 25 yearly. The membership list of the Netherlands Association of Phonetic Sciences contains about 140 names (including some from Belgium). In the Netherlands about seven doctoral theses in Phonetics and related fields are produced every year; the actual list for 2001 to 2003 is presented in Table 1.

3 Growth in Phonetics as a multidisciplinary science

In 1983 the Dutch Ministry of Education decided in the Academic Statute that only at Utrecht University there could be a Department of Phonetic Sciences. Nevertheless several regular chairs in Phonetics continued to exist in the Netherlands (Utrecht: Cohen and later Nooteboom; Leiden: first Nooteboom, then nobody, now Van Heuven; Nijmegen: Vieregge (who retired in 2000), later also Boves (Language Engineering) and Rietveld (Speech Pathology); Groningen: Graham Stuart for a while; Amsterdam: Pols and Hilgers (Speech Pathology). In inventive ways slightly different names were used in the curriculum, for instance in Amsterdam the study was called Speech Communication, in Nijmegen they had Speech Pathology and Speech Technology as specializations, and in Leiden there was (and still is) a Phonetics Laboratory as part of the Linguistics Department. At least since the introduction of the bachelor-master system, the situation is totally different again. Now Phonetics is part of (General) Linguistics everywhere, and specialization in Phonetics or Speech Technology is possible in the Master phase only, if at all. Furthermore, there is a

tendency not to maintain specializations, like Phonetics, everywhere, but only at a few places. Also at the Institute for Perception Research (IPO) in Eindhoven a lot of speech research was done until the institute was dismantled. In May 2004 the very last doctoral thesis in the IPO synthesis tradition was defended (van Herwijnen, 2004). For a while there were also individual phoneticians throughout the various Dutch universities working in the Departments of Modern Languages (e.g., Peeters, 1991; van Buuren, 1999). Also outside the Faculty of Arts, phonetic research may be found in Education Research (e.g., van Gelderen, 1992), in the Ear, Nose and Throat clinics (e.g., Verdonck-de Leeuw, 1998; Jansonius-Schultheiss, 1999; van As, 2001) and in the Departments of Audiology or Pediatrics Neurology (e.g., Nijland, 2003) of the academic hospitals. Also in psycholinguistics there is a certain interest in phonetics, most clearly reflected in some of the work done at MPI, Nijmegen (e.g., van Alphen, 2004; Kamps, 2004). Finally, speech and language technology gets some attention in Departments of Computer Linguistics, Informatics, Artificial Intelligence, or Technology Management, also at Technological Universities (e.g., Andringa, 2002; Dinther, 2003; Ordelmans, 2003). On the one hand, this short survey indicates a reduction in clearly labeled Phonetics settlements; on the other hand, it shows how strongly Phonetics and phoneticians are involved in many aspects of science and technology.

Table 1. Doctoral theses in Phonetics and related fields, defended in the Netherlands over the years 2001 to 2003.

Name	Univ.	Year	Promotor(es)	Topic
P. Adank	KUN	2003	van Hout	prevoicing in Dutch
T.C. Andringa	RuG	2002	Duifhuis	continuity preserving signal proc.
C. van As	UvA	2001	Pols	tracheoesophageal speech
O.A. Crasborn	UL	2001	Ewen/ van Heuven/ van der Hulst	sign language
J. van Dijk	UvA	2001	Pols	ear modeling
R. van Dinther	TUE	2003	Kohlrausch/ Liljencrants	voice source
A.M. Elgendy	UvA	2001	Pols	Arabic pharyngeals
E. Gerrits	UU	2001	Nooteboom	categorical perception
J. Haan	KUN	2002	van Heuven/ Gussenhoven	Dutch question intonation
S.R. Hamann	UU	2003	Hall (Leipzig) / Zonneveld	retroflexes
E. Janse	UU	2003	Nooteboom	fast speech
J.M. Kessens	KUN	2002	Boves	pronunciation variation in ASR
E. Konst	KUN	2002	Kuijpers-Jagtman	babbling with infant orthopaedics
E. Marsi	KUN	2001	Gussenhoven	intonation in spoken language generation
L. Nijland	KUN	2003	Kraaimaat/ Gabriëls	developmental apraxia
R.J.F. Ordelman	TUT	2003	de Jong	speech recognition
A.C.L. Remijsen	UL	2001	van Heuven/ Stokhof	word prosody
B.M. Streefkerk	UvA	2003	Pols	prominence
J.M. de Veth	KUN	2001	Boves	speech sound model accuracy
M. Wester	KUN	2002	Boves	pronunciation variation in ASR
F. de Wet	KUN	2003	Boves	ASR in adverse conditions
S. v Wijngaarden	VU	2003	Houtgast	intelligibility of non-native speech

4 State of affairs of Phonetics in the Netherlands

It does not happen very often that the Dutch chair holders in Phonetics individually or jointly express themselves about their vision of the field, apart from their inaugural addresses at the start of their professorship (Boves, 1993; van Heuven, 2002) or when they retire (Cohen, 1987; Vieregge, 2000; Nooteboom, 2004). Still, there was such an occasion in 1997 when the Netherlands Association of Phonetic Sciences celebrated its 65th anniversary in Amsterdam with a special theme meeting about the personal visions of all four professors in Phonetics at that time.

Boves presented himself as the representative of speech and language technology. He blamed phonetics and linguistics for not giving enough attention to testable hypotheses. He said to be an advocate of knowledge to be applicable. He preferred simple models and workable approaches, like the use of LPC and concatenative synthesis, rather than models that are too complex, like ARMA. The probabilistic approach in ASR may be simplistic, but it works. Still he would like to develop more robust recognition, for instance by taking into account pronunciation variation. In the discussion the audience questioned whether the probabilistic approach would still be our best option for more complex problems in the near future.

Nooteboom expressed an interest in both *matter* (physically measurable objects) and *mind* (mental processes). He asked attention for questions like 'How many words do we look ahead in preparing our spoken output?' (e.g., *ik wees wat ik lees* rather than *ik weet wat ik lees*), 'What are the units in the mental lexicon?' (morphemes, words, standing expressions), 'During which phases can previous context influence word recognition?' (lexical access, selection, integration), and 'Is recognizing fixed expressions similar to word recognition?'. He claimed that in his head there is a very big lexicon and only very few rules. In the discussion one wondered whether there would be only one (word) lexicon.

Vieregge defended the unification of perception and movement in speech production (Phonetischer Gestaltkreis; Quantal Nature of Speech). Feedback can be tactile/kinesthetic, auditory, and communicative. He distinguished the 'type of movement' versus the 'moving time'. Finding a proper balance between the canonical movement and the time constraints of speech production frequently seems to be a problem with synthetic speech, baby speech, speech of drunkards, foreign accents, and pathological speech. In the discussion it was questioned whether there was any linguistic relevance of the quantal nature of speech.

Pols summarized the work in his group in four main themes: (i) speech analysis and perception (both segmental and supra-segmental); (ii) evaluation of normal and pathological speech and of speech technology; (iii) adding specific knowledge to speech technological systems; and (iv) speech development and pathology. He then proposed to establish a (virtual) Dutch phonetics center for joint research. He also advocated the concept of computational phonetics (Pols, 1999, 2001).

Although these presentations reflect viewpoints from several years ago, they nevertheless quite clearly represent the key research areas of the four groups up to now.[1] Also the topics of doctoral theses completed under their supervision, as can partly be seen in table 1, are proof of that.

Actually, phonetics has quite a record of good cooperation in large joint projects, such as SPIN-ASSP (Van Heuven & Pols, 1993), a five-year strategic research project (1985-1990) towards high-quality text-to-speech generation, under the stimulating leadership of Cohen.

[1] The fifth group at Universiteit Leiden, headed by Van Heuven, specializes in linguistic phonetics and laboratory phonology.

This project was sponsored by the Stimulation Project Information technology Netherlands (SPIN) that operated under the jurisdiction of the Ministries of Economic Affairs and of Education and Science. Another large project (1995-2000) was the NWO Priority program 'Language and Speech Technology (TST)' (Strik, Russel, van den Heuvel, Cucchiarini & Boves, 1996), which recently got a follow-up in the NWO IMIX project (Interactive Multimodal Information eXtraction). Another example of joint cooperation is the existence since 1987 of SPEX, the speech processing expertise center that was located for many years in Leidschendam but is now part of the Center for Language and Speech Technology (CLST) at Nijmegen University. Until recently SPEX fell under the responsibility of SST, the Foundation for Speech Technology with Nooteboom as the chairman of the Board. Chances seem to be good for still another cooperative project BLARK (Basic Language Resources Kit, in Dutch BATAVO) (Strik, Daelemans, Binnenpoorte, Sturm, de Vriend & Cucchiarini, 2002). A Belgium-Netherlands bid to ISCA for organizing Eurospeech 2007 in Antwerp has recently been accepted. Over the years various workshops of ESCA, now ISCA, ELSNET and others have been organized in the Netherlands (Speech input/output assessment and speech databases, Noordwijkerhout, 20-23 September 1989; Elsnet in Wonderland, Soesterberg, March 1998; Modeling pronunciation variation for automatic speech recognition, Rolduc, 4-6 May 1998; Multi-lingual interoperability in speech technology, Leusden, 13-14 September 1999; Spoken word access processes, Nijmegen, 29-31 May, 2000; LabPhon7, Nijmegen, 29 June - 1 July 2000; Speech recognition as pattern classification, Nijmegen, 11-13 July 2001).

5 Conclusions

Despite all these activities in our field, phonetics as an independent discipline is endangered, especially in the new bachelor-master educational system. There is a serious threat that in the future not enough students will be properly trained in the theoretical and experimental aspects of phonetics. For the moment, the phonetic flavour in many projects seems to be flourishing, but this is only possible because students trained in the old curriculum are still available. Moreover, there is a tendency even now to attract Ph.D. students from abroad. The integration of Phonetics and Phonology is a matter of (historical) ups and downs. Ohala's presentation preceding the farewell lecture of Nooteboom (2004) gave a good overview of that process (Ohala, 2004; see also Cohen, 1987, and Boersma, 1998).

References

Adank, P.M. (2003). *Vowel normalization: A perceptual-acoustic study of Dutch vowels.* Ph.D. thesis, University of Nijmegen.

Alphen, P.M. van (2004). *Perceptual relevance of prevoicing in Dutch.* Ph.D. thesis, University of Nijmegen. MPI Series; 25.

Andringa, T.C. (2002). *Continuity preserving signal processing.* Ph.D. thesis, University of Groningen.

As, C.J. van (2001). *Tracheoesophageal speech. A multidimensional assessment of voice quality.* Ph.D. thesis, University of Amsterdam.

Boersma, P.P.G. (1998). *Functional phonology. Formalizing the interactions between articulatory and perceptual drives. Ph.D. thesis*, University of Amsterdam. LOT Dissertation Series; 11.

Boves, L. (1993). *Sprake van taal,* Oratie. Nijmegen: Katholieke Universiteit Nijmegen.

Buuren, L. van (1990). *The indispensable foundation of linguistic study. Ph.D. thesis*, University of Amsterdam.

Cohen, A. (1987). *Taal en spraak.* Afscheidsrede. Utrecht: Rijksuniversiteit Utrecht.

Crasborn, O.A. (2001). *Phonetic implementation of phonological categories in Sign Language of the Netherlands*. Ph.D. thesis, University of Leiden. LOT Dissertation Series; 48.

Dijk, J.S.C. van (2001). *Mechanical aspects of hearing. Ph.D. thesis*, University of Amsterdam.

Dinther, R. (2003). *Perceptual aspects of voice-source parameters. Ph.D. thesis*, Technological University Eindhoven.

Elgendy, A.M. (2001). *Aspects of pharyngeal coarticulation,* Ph.D thesis, University of Amsterdam. LOT Dissertation Series; 44.
Gelderen, A.J.S. van (1992). *De evaluatie van spreekvaardigheid in communicatieve situaties. Globale beoordeling en gedetailleerde analyse van spreekprestaties van 11- en 12-jarigen. Ph.D. thesis*, University of Amsterdam. SCO Report; 303.
Gerrits, E. (2001). *The categorisation of speech sounds by adults and children.* Ph.D. thesis, Utrecht University. LOT Dissertation Series; 42.
Haan, J. (2002). *Speaking of questions. An exploration of Dutch question intonation.* Ph.D. thesis, University of Nijmegen. LOT Dissertation Series; 52.
Hamann, S.R. (2003). *The phonetics and phonology of retroflexes.* Ph.D. thesis, Utrecht University. LOT Dissertation Series; 75.
Heuven, V.J. van (2002). *Boven de klanken / Beyond the segments,* Oratie, Universiteit Leiden. Amsterdam: Edita.
Heuven, V.J. van & Pols, L.C.W. (Eds.) (1993). *Analysis and synthesis of speech. Strategic research towards high-quality text-to-speech generation,* Berlin: Mouton de Gruyter.
Herwijnen, O.M. van (2004). *Weighted error minimization in assigning prosodic structure for synthetic speech*. Ph.D. thesis, Technological University Eindhoven.
IMIX, Interactive Multi-modal Information extraction, http://www.nwo.nl/imix/.
IPA, The International Phonetic Association, http://www.arts.gla.ac.uk/IPA/ipa.html.
ISCA, International Speech Communication Association, http://www.isca-speech.org/.
Janse, E. (2003). *Production and perception of fast speech.* Ph.D. thesis, Utrecht University. LOT Dissertation Series; 69.
Jansonius-Schultheiss, K. (1999). *Twee jaar spraak en taal bij schisis.* Ph.D. thesis, University of Amsterdam. LOT Dissertation Series; 17.
Kamps, R.J.J.K. (2004). *Morphology in auditory lexical processing. Sensitivity to fine phonetic detail and insensitivity to suffix reduction.* Ph.D. thesis, University of Nijmegen.
Kessens, J.M. (2002). *Making a difference. On automatic transcription and modeling of Dutch pronunciation variation for automatic speech recognition.* Ph.D. thesis, University of Nijmegen.
Konst, E. (2002). *The effects of infant orthopaedics on speech and language development in children with unilateral cleft lip and palate.* Ph.D. thesis, University of Nijmegen.
Marsi, E. (2001). *Intonation in spoken language generation.* Ph.D. thesis, University of Nijmegen. LOT Dissertation Series; 46.
Nijland, L. (2003). *Developmental apraxia of speech: Deficits in phonetic planning and motor programming.* Ph.D. thesis, University of Nijmegen.
Nooteboom, S.G. (2004). *Waar komen de letters in het alfabet vandaan?* Afscheidscollege. Utrecht: Universiteit Utrecht.
Ohala, J. (2004). Phonetics and Phonology then, and then, and now. In this volume.
Ordelmans, R.J.F. (2003). *Dutch speech recognition in multimedia information retrieval.* Ph.D. thesis, University of Twente, Centre for Telematics and Information Technology.
Peeters, W.J.M. (1991). *Diphthong dynamics. A crosslinguistic perceptual analysis of temporal patterns in Dutch, English, and German.* Ph.D. thesis, Utrecht University.
Pols, L.C.W. (1999). Flexible, robust, and efficient human speech processing versus present-day speech technology. In J.J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A.C. Bailey (Eds.), *Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, August 1-7, 1999* (Vol. 1, pp. 9-16).
Pols, L.C.W. (2001). Acquiring and implementing phonetic knowledge. In *Proceedings Eurospeech'01, Aalborg* (Vol. 1, pp. K3-K6).
Remijsen, A.C.L. (2001). *Word-prosodic systems of Raja Ampat languages,* Ph.D. thesis, University of Leiden. LOT Dissertation Series; 49.
SPEX, Speech Processing EXpertise Centre, http://www.spex.nl/.
Streefkerk, B.M. (2003). *Prominence. Acoustic and lexical/syntactic correlates. Ph.D. thesis*, University of Amsterdam. LOT Dissertation Series; 58.
Strik, H., Daelemans, W., Binnenpoorte, D., Sturm, J., De Vriend, F. & Cucchiarini, C. (2002). Dutch HLT resources: From BLARK to priority lists. In *Proceedings ICSLP, Denver* (Vol. 3, pp. 1549-1552).
Strik, H., Russel, A., Heuvel, H. van den, Cucchiarini, C. & Boves, L. (1996). Localizing an automatic inquiry system for public transport information. In *Proceedings ICSLP, Philadelphia* (Vol. 2, pp. 853-856).
Verdonck-de Leeuw, I.M. (1998). *Voice characteristics following radiotherapy: The development of a protocol,* Ph.D. thesis, University of Amsterdam. IFOTT Series; 33.
Veth, J.M. de (2001). *On speech sound model accuracy.* Ph.D. thesis, University of Nijmegen.

Vieregge, W.H. (2000). *Homo Loquens: Nog steeds een wonder.* Afscheidscollege, University of Nijmegen.
Wester, M. (2002). *Pronunciation variation modelling for Dutch automatic speech recognition.* Ph.D. thesis, University of Nijmegen.
Wet, F. de (2003). *Automatic speech recognition in adverse conditions.* Ph.D. thesis, University of Nijmegen.
Wijngaarden, S. van (2003). *The intelligibility of non-native speech.* Ph.D. thesis, Free University of Amsterdam.

What is the Just Noticeable Difference for tempo in speech?

Hugo Quené

Utrecht University

Abstract

Tempo (speaking rate) varies both between and within speakers. Such variations in tempo are easily noticeable. But what is the just noticeable difference for tempo in speech? As a first approximation, between-speaker tempo variation is quantified in a sample of similar interviews with 80 speakers of Dutch. Second, the JND is assessed using a somewhat unconventional method, viz. detection of tempo drift. This results in a JND of 15%; it is argued that this value is upwardly biased. Third, the JND is assessed using a conventional same~different pairwise comparison, yielding a JND of about 10%. Tempo variations between and within speakers typically exceed this JND, and are therefore potentially important in speech communication.

1 Introduction

Human speech is produced by moving the vocal organs and articulators. These movements result in an articulated speech signal, in which phonetic events occur at particular moments in time. The rate at which these speech events occur constitutes the tempo or speed or rate of speech. Many textbooks in phonetics state that speakers vary their speaking rate, in anticipation of the time listeners will need to process their words. Hence, important or unpredictable portions are spoken at a relatively slower rate (e.g., Zwaardemaker & Eijkman, 1928, p.304; Nooteboom & Cohen, 1984, p.165). This tendency follows from the adaptation principle or H&H principle (Lindblom, 1989): speakers adjust phonetic properties of their speech to ensure an optimal balance between economy of articulatory energy, and perceptual clarity for the listener. After all, speakers speak in order to be understood. This phonetic principle underlies the usual rhetoric advice to public speakers, to slow down when important information is conveyed (e.g., Humes, 2002). In most phonetic studies supporting these claims and recommendations, however, information value (newness, importance) and accentuation have been confounded. Eefting and Nooteboom (1993) (and Nooteboom & Eefting, 1994) show that speaking rate is *not* slower for new words, if accentuation is also taken into account (at least for the single professional standard speaker in their study). Their conclusion is based on the relatively small increment by about 4% in duration of a target word *John*, in the [+new, –accent] condition (1) relative to the [–new, –accent] condition (2).

(1) (What did you say?) John MILLER is ill.
(2) (What did you say about John Miller?) John Miller is ILL.

Hence, this evidence rests on the implicit assumption that the observed increment of 4% in duration (or decrement in speaking rate) is not relevant for speech communication[1]. If this 4% difference is above the difference limen for speaking rate, however, then this assumption

[1] By way of comparison: accenting the target word ('JOHN is ill') yields an increment in word duration and in syllable duration of about 25% (Eefting & Nooteboom, 1993).

is not warranted. But what exactly is the difference limen (DL) or just noticeable difference (JND) for speaking rate? If a speaker changes tempo, how large does the tempo change have to be in order to be perceptually relevant? To this date, a few studies have addressed this question (Benguerel & D'Arcy, 1986; Eefting & Rietveld, 1989; Nooteboom & Eefting, 1994), but these studies leave considerable room for improvement, as will be discussed below.

As a first approximation to an answer, we could inspect tempo differences between speakers. If speakers are observed to differ in their basic tempo (e.g., Goldman-Eisler, 1968; Den Os, 1985; Van Heuven, 2003), then these between-speaker differences must exceed the observer's JND. Hence, the range of tempi in a large speech corpus can help us to establish the JND for speech tempo. The variation in speaking rate was investigated by means of the Corpus of Spoken Dutch for this purpose.

2 Corpus analysis

The Corpus of Spoken Dutch (CGN, e.g. Oostdijk, 2000) was used to quantify the between-speaker variation in speaking rate. For this purpose, we concentrated on the sub-corpus containing interviews with high-school teachers of Dutch in the Netherlands and Belgium (Van Hout et al., 1999). Only the speakers from the Netherlands are discussed here. The relevant sub-corpus contains interviews with 80 speakers, each interview lasting about 15 minutes. Interviewed speakers ('interviewees') were stratified by dialect region (four regions within the Netherlands), sex, and age group (below 35 vs. over 45 years of age), with n=5 speakers in each cell. All 80 speakers are assumed to speak a variety of Standard Dutch as used in the Netherlands. All 80 interviews were conducted by the same interviewer (female, age 26), and similar topics were discussed across interviews. Hence, language variety, conversation partner, and conversation topic were eliminated as confounding factors.

For each interview, the orthographic transcript of the interviewee was extracted, and the speaking time of the interviewee was determined from the time marks in this transcript. Pauses were thus excluded from the interviewee's speaking time. On average, the interviewee spoke during about 3/4 of the total interview duration. Phonetic speaking rate is usually expressed in a syllables-per-second scale (e.g., Stetson, 1988) or average syllable duration (Goldman-Eisler, 1968). Because the necessary phonetic transcripts are not available for this part of the CGN, the words-per-minute (wpm) scale was used instead. The number of words spoken by the interviewee was counted in the orthographic transcripts, using standard word-processing software (TextPad 4.7.2). For each interview, the speaking rate was calculated from the speaking time and word count.

In these 80 interviews, the average speaking rates of the interviewees range between 151 and 281 words per minute, with an overall average of 220 wpm (s=25; the distribution is approximately normal, KS=0.068, p=.45). This range and variation between speakers is quite large, considering that the interviews were similar with respect to interviewer, topics discussed, and total duration of interview. An average speaker would require 4.55 minutes to produce 1000 words. The fastest speaker would require only 3.56 minutes (0.78 × average), and the slowest speaker 6.62 minutes (1.45 × average).

Further analysis of variance shows that the speaking rate varies significantly between male and female speakers [227 vs. 213 wpm, respectively; $F(1,76)=15.0$, $p<.001$], and between younger speakers (mean age 33.5 year; mean rate 211 wpm) and older speakers [mean age 51.6 years; mean rate 230 wpm; $F(1,76)=7.7$, $p=.007$]. No interaction was observed between the sex and age-group factors, $F(1,76)=1.2$, n.s.

Hence, speaking rates do vary considerably between speakers, as has been observed many times before (e.g. Goldman-Eisler, 1968; Den Os, 1985). The rate of the fastest speaker is almost double that of the slowest speaker, in this part of the CGN. It comes as no surprise, then, that these large between-speaker differences are highly noticeable. A rough estimate of JND is obtained from the normal distribution of speaking rates across speakers, as follows. Let us make the debatable assumption that perception mirrors production, and that listeners' JND corresponds to 1 standard deviation of the distribution of produced tempi. This would amount to an estimated JND of 25/220 or 11%. In reality, however, listeners seem to perform better in speech tempo discrimination (Nooteboom & Eefting, 1994), and hence the JND is smaller, by some unknown amount.

3 On the Just Noticeable Difference for tempo

In the preceding section, it was assumed that listeners discriminate between the slowest and fastest decatile of the population. This assumption was necessary because the true JND for speech tempo remains to be determined. Hence the main question of this study: what is the just noticeable difference (JND) for tempo in speech? How much do speaking rates have to deviate from a reference in order to become noticeable, and hence relevant in speech communication? In the following sections, two experiments are reported that were aimed at establishing the JND for tempo in speech.

Research on JNDs for tempo has concentrated on *music* perception. Ellis (1991) presented listeners with a 6-bar, 24-beat musical fragment at various base tempi. After a stable period (of random duration), the tempo of the fragment started to drift gradually (up or down, with +2% or –2% on each subsequent beat, to extremes of either +16% or –10%). Using the staircase adjustment method, JNDs of 5.1% to 13.9% change in tempo were found, depending on the direction of drift and on the base tempo[2].

Drake and Botte (1993, Experiment 3) presented listeners with two 5-tone sequences to compare (2IFC paradigm). Using the staircase adjustment method, they found JNDs for this type of stimuli to be 6% to 10% for nonmusical listeners, and 3% to 8% for musicians, with the lowest JND at a base rate of 100 beats per minute (inter-onset interval 0.6 s). In an ERP study having a similar design, Pfeuty, Ragot and Pouthas (2003) report that a 4% change in inter-onset interval in a 7-tone sequence yields a discriminability value *d'* of 1.52, which suggests that the JND is smaller than 4% in their experiment.

Levitin and Cook (1996) cite an unpublished study by Perron (1994) involving computer sequencers or drum machines, as used in popular music. Although such machines turned out to have an average tempo deviation of 3.5%, most listeners do not notice these deviations (and neither do professional drummers). This suggests that the JND for musical tempo is at least 3.5%.

These studies (and others not discussed here) indicate that the JND for musical tempo is approximately 6% to 8% of the base tempo. For *speech* tempo, however, only a few estimates of JND are available. Benguerel and D'Arcy (1986, Experiment 4) presented a few pre-selected listeners with reiterant speech stimuli [nanananananana], with exponentially decreasing or increasing duration of each subsequent syllable. Listeners judged sequences as "regular", even for decelerating sequences with increasing syllable duration. However, a conventional

[2] This could be considered as a 2IFC paradigm, in which the beginning and the ending of the musical fragment constitute the two intervals to be compared.

JND value cannot be derived easily from their results. In addition, it is not clear whether and how this generalizes to normal speech and to other listeners.

Eefting and Rietveld (1989) presented listeners with two versions of a speech utterance (2IFC paradigm); tempo was always unchanged in one version. The reported JND of 4.4% is remarkably low, and even lower than some values reported above for musical tempo. Eefting and Rietveld argue that this may be due to listeners' adaptation to the stimuli. Interestingly, a JND this low would render the 4% change in tempo reported by Eefting and Nooteboom (1993) perceptually relevant. Again, generalization to other speech stimuli is questionable.

Assessing JNDs for tempo is surely more difficult in speech than in music, where presenting several intervals between beats requires only a few seconds. In speech, each syllable has its intrinsic duration and temporal structure, depending on the articulatory gestures involved in producing that syllable. This obscures the temporal regularity to some extent. Consequently, listeners may well need more than a few (stressed) syllables to determine the speaking rate in a speech fragment.

The classical method to determine a JND is to present two stimuli shortly after each other, which are either identical or slightly different in one property, and to ask listeners whether these are the same or different (2IAX paradigm) (e.g. Nooteboom & Eefting, 1994), or which one of the two has more of the property under investigation (2IFC paradigm) (e.g. Den Os, 1985; Drake & Botte, 1993). Hence the assumption is that listeners have an active memory trace of *both* stimuli when they respond. This assumption may well be unwarranted if longer stimuli are used, such as musical melodies or speech fragments that typically last at least a few seconds. An experimental paradigm involving a single stimulus presentation is to be preferred in these cases.

Hence, it will be attempted below to establish the JND for speaking tempo, by means of drift detection. In this task, the tempo in a stimulus starts to drift (i.e., to accelerate or decelerate), and the listeners' task is to respond as fast as possible to a perceived change in tempo. The JND is then calculated from the amount of change in the stimulus, at the time the listener responds. This method has already been successful for investigating musical tempo (Ellis, 1991; Grondin, 2003).

4 Detection of tempo drift

4.1 Method

Stimulus materials consisted of 36 text passages that resembled short news items, consumer reports, personal anecdotes, etc. Four female speakers spoke nine passages each, at a normal rate. Their productions were recorded on DAT, and re-digitized (16 kHz, 16 bits). Each passage contained about 10 s of speech in two or three sentences.

Between recorded text passages, the onset of tempo drift was varied between 2 and 4 s after onset of speech, to avoid strategic effects. PSOLA manipulation was used to create both the accelerating and the decelerating version of each recording. Relative duration of the speech recording was compressed (by linear interpolation from 1.0× to an extreme of 0.8× original; tempo acceleration) and expanded (to 1.2×; tempo deceleration) over a 5-s time window[3]. The two versions were counterbalanced over two stimulus lists, each of which also contained four practice items. Within each stimulus list, the blocking order of accelerating and decelerating items was also counterbalanced.

[3] Hence the tempo drifts up or down by 4% per second, during 5 seconds.

Each list was presented to 17 listeners, who reported no hearing defects (age mean 26, median 23 years). Listeners were instructed to press a button (with the index finger of their preferred hand) as soon as they detected a change in tempo in the speech stimulus. Response latencies were measured from the onset of tempo drift. Data from 3 participants were lost for various reasons.

4.2 Results and discussion

The JND for speech tempo can be obtained from the response latencies, as follows. If a listener responds at 2 s after onset of drift, then he has detected an 8% change in duration or tempo. A listener's JND is defined here as the amount of tempo drift at which he has responded in 50% of the test items. Figures 1 and 2 below show the listeners' proportions of hits, as a function of time after onset of drift. This procedure yields average JNDs of -15.4% for accelerating tempo (relative duration 0.846; Figure 1) and +17.2% for decelerating tempo (relative duration 1.172; Figure 2).

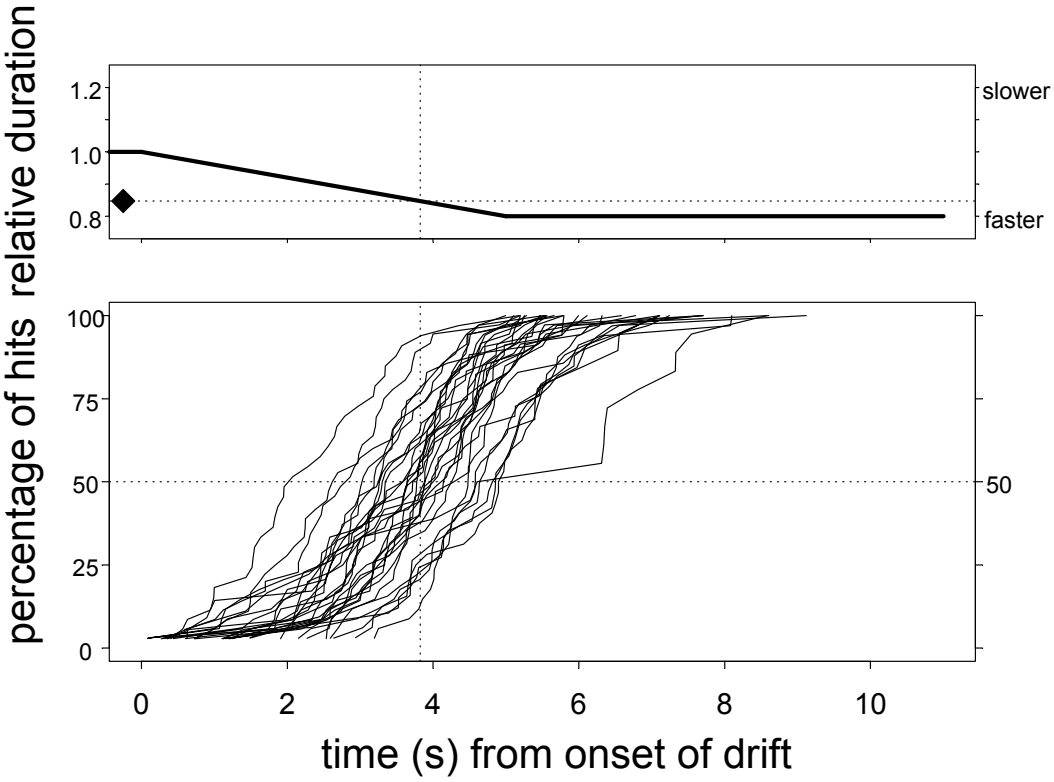


Figure 1. Individual listeners' proportions of hits (tempo drift detected) as a function of time (in seconds) after onset of tempo acceleration. The top panel shows the relative duration of the speech stimulus on the same time axis, with the average JND marked.

These results suggest that the average JND for speech tempo across listeners is about 15%, which is considerably larger than the values reported for musical tempo mentioned above. This larger JND value could mean two things. Perhaps the JND for speech tempo is indeed larger than that for music tempo; this could be due to the relatively large pertubatory effects

of articulatory gestures on 'underlying' tempo. As an alternative explanation, the method of drift detection used here could be less reliable for speech tempo than for music tempo.

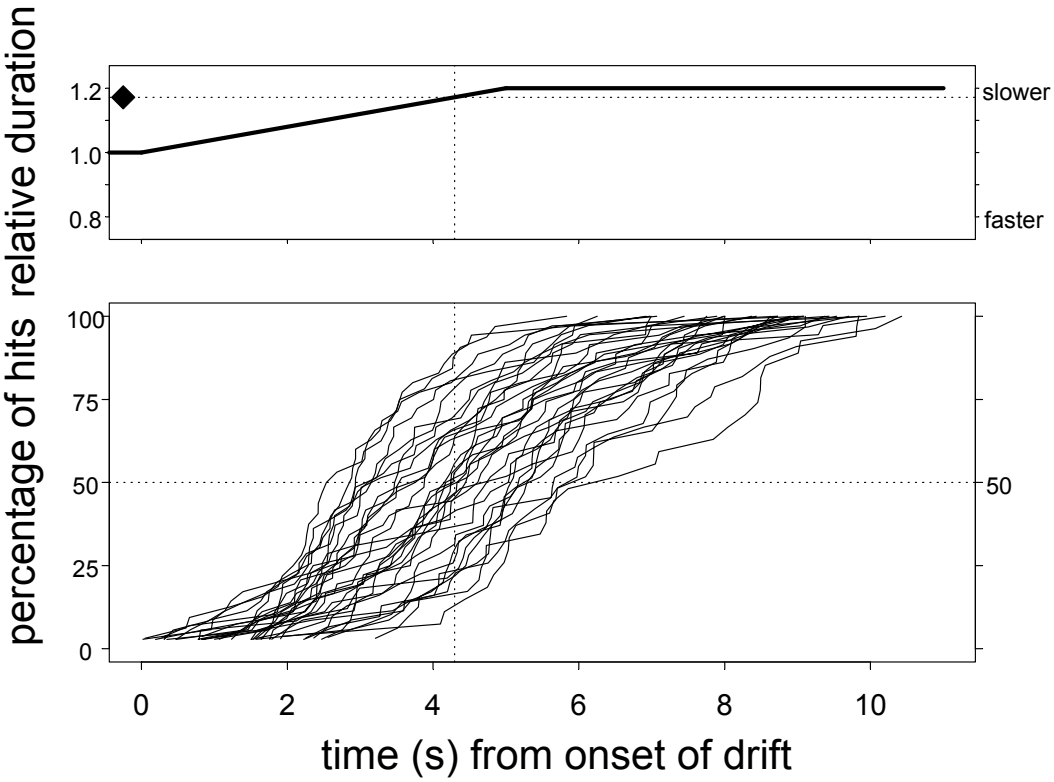


Figure 2. Individual listeners' proportions of hits (tempo drift detected) as a function of time (in seconds) after onset of tempo deceleration. The top panel shows the relative duration of the speech stimulus on the same time axis, with the average JND marked.

Several arguments suggest that this latter explanation may be the more plausible one. First, there is considerable variation among texts or items in their JNDs (for accelerations, range is 8% to 21%, for decelerations 8% to 23%). The tempo manipulation consisted in increasing compression or expansion, over a 5 s time window, irrespective of the contents of that window. This might have increased variability in the obtained JNDs, because speech pauses have not been taken into account. If the manipulation window contained a speech pause of 1 second, from t=1 (where acceleration is 4%) to t=2 (acceleration 8%), then the listener cannot detect intermediate values, because acceleration is not audible within a pause. The onset of the manipulation window was varied, but this variation plus the random presence of pauses within that window may well have increased the noise in listeners' response times, which increased the resulting JND values[4].

[4] If listeners did not respond before such pauses, then they could only respond *after* such a pause, and not at an intermediate point in time (during the pause). Hence, the effects described in the text could only have *increased* response times and their JNDs.

Consequently, it was attempted to establish the JND for speech tempo by a more conventional method, in which listeners have to compare two speech stimuli that vary only in the property under investigation, viz. speech tempo.

5 Pairwise comparison

5.1 Method

Stimulus materials consisted of 20 text passages, selected from the above experiment (5 passages from each speaker). Each was pruned to a fragment that does not contain major pauses (the resulting fragment usually corresponds to a major phrase). Means and standard deviations over the 20 fragments were as follows: duration 3.035 s (0.553), length 8 words (2), length 13.5 syllables (3.4), tempo 159.9 wpm (37.4), average syllable duration 239 ms (77).

These fragments were then accelerated to 0.80, 0.85, 0.87, 0.89, 0.91, 0.93, and 0.95 of the original duration, and decelerated to 1.05, 1.07, 1.09, 1.11, 1.13, 1.15 and 1.20 of the original duration, yielding 7×2 manipulated plus 1 unmanipulated version for each fragment. Temporal compression or expansion was uniform throughout the fragment.

Listeners were 24 students of a PABO college in Utrecht, who reported no hearing defects. They heard two versions of the same passage, with a 600 ms interval between the versions. One version was always the original or reference version, the other was one of the 14 manipulated versions or the original version. Each pair was presented in two orders, with the reference version as either the first or last member of the presentation pair. Listeners' task was to indicate whether the two versions were the same or different (Nooteboom & Eefting, 1994). This task was chosen because of its similarity with the detection of tempo drift. The order of the 640 pairs (20 passages × 8 versions × 2 orders × 2 directions) was randomized anew for each listener. Listeners were tested individually in a quiet room. They indicated their same-or-different response by pressing one of two keys, of which the "different" key was always under their dominant hand. Total time of each session was approx. 1.5 hours, including 3 short pauses at regular intervals.

Data from two listeners were discarded: one because of her high miss rate, and the other because of a mild 'cluttering' disorder in his speech. Responses on two manipulated versions that turned out to be defective, were also discarded.

5.2 Results and discussion

In a 2IAX design, the JND is sometimes defined as the difference in tempo or duration at which half of the responses are "different". The upper panel of Figure 3 below shows the listeners' percentages of "different" responses, as well as their average. The present experiment also contains a bias, however, which renders the above procedure inappropriate. If both members of a pair are the same, then no "different" responses are expected. Because all versions were compared with an unmanipulated reference version, this situation only occurred with both versions unmanipulated. Interestingly, listeners incorrectly judged these pairs as "different" in 15.1% of their responses; this false-alarm rate deviates significantly from zero (with standard error of 3.7% between listeners' average hit rates, $p<.001$). This possible bias may stem from the 2IAX (same~different) design, which refers to a criterion internal to the

listener (a criterion of equality or difference), which makes this type of experiment susceptible to bias[5].

Hence, the JND was defined here by means of *d'* values, because these are based on the percentages of hits as well as false alarms. The *d'* values are plotted in the lower panel of Figure 3. The JND is defined as the difference in tempo or duration at which *d'=1*. By this procedure, average JNDs are –9.0% for accelerating tempo, and +11.5% for decelerating tempo.

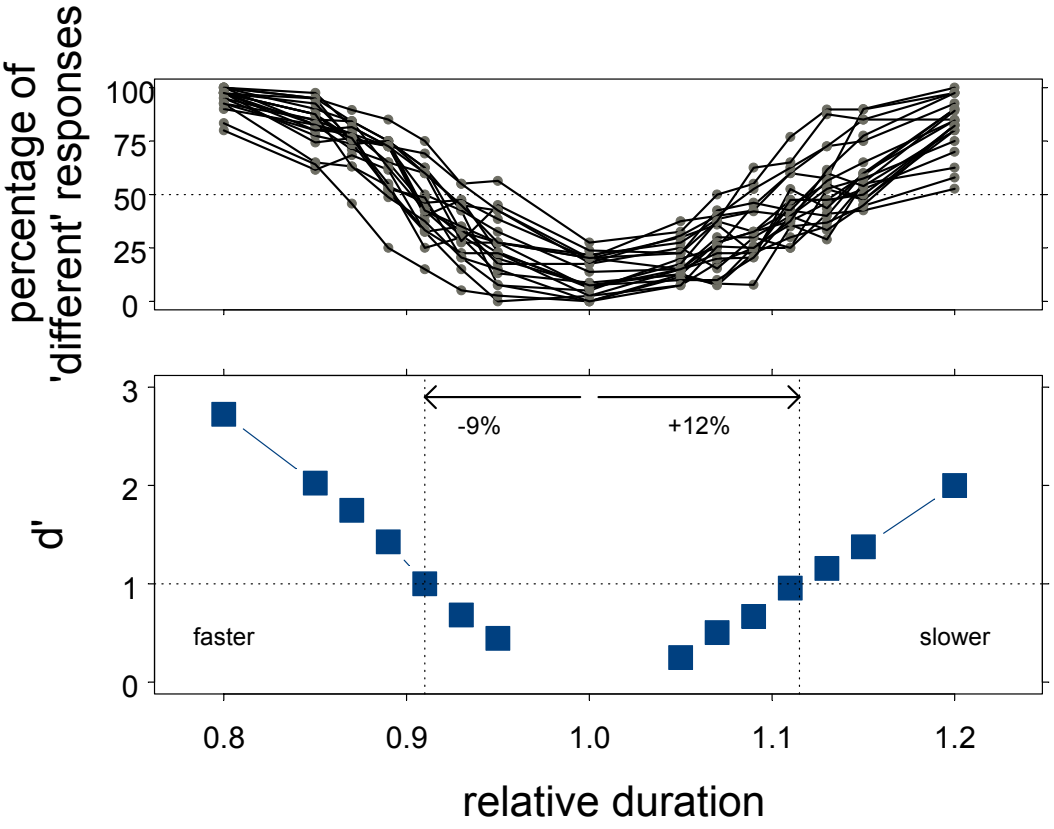


Figure 3. Individual listeners' percentages of "different" responses (*top*), and average values of *d'* (*bottom*), as a function of relative duration or tempo.

The lower JNDs obtained in this 2IAX experiment suggest that the method in the previous experiment, viz. detection of tempo drift, is indeed not valid, in that it over-estimates the JND for speech tempo. The present JNDs are more in line with those reported for musical tempo, especially if we allow for the larger variability in speech timing (as compared to musical timing) due to articulatory constraints. A provisional conclusion of this present experiment is that the JND for speech tempo is about 10%.

[5] The 2IFC task does not suffer from this disadvantage: the two versions of a pair are only compared with each other (first~second is faster), without reference to a subjective criterion. This latter method is therefore to be preferred. We hope to re-run the present 2IAX experiment as a 2IFC experiment.

6 Discussion

The observed JND of 10% change in speech tempo appears to match the between-speaker distribution of speech tempo in the Spoken Dutch Corpus. One third of the speakers differs by more than 10% from the overall average speaking rate (14 of them are slower than –10%, and another 14 are faster than +10%). Hence, between-speaker differences should be easily noticeable, as indeed they are.

How large do within-speaker variations in speaking rate have to be, in order to be perceptually relevant? The present study suggests that changes in speech tempo smaller than 10% are not noticeable. Hence, the 4% change in tempo reported by Eefting and Nooteboom (1993) was indeed not perceptually relevant — in fact, it would be hardly noticeable even in tone sequences or drum machines.

Because some within-speaker changes in tempo *are* noticed, and relevant (see below), these changes must be larger than 10%. A detailed analysis of within-speaker tempo variation in the CGN sample is still pending, but other studies suggest that speakers do indeed accelerate and decelerate by a considerable amount. For example, Nooteboom and Eefting (1994) report on a professional speaker, whose average syllable duration per phrase ranged between 118 ms and 289 ms (with mean 176, s.d. 35); this amounts to a change by –33% and +64%. Chafe (2002) reports on a spontaneous conversation in which one speaker accelerates by 33% to convey her high emotional involvement (her average syllable duration decreases from 150 to 100 ms).

Such tempo variation is indeed communicatively relevant, as exemplified by a recent study on the effects of speaking rate on responses to a spoken advertisement, with young adult listeners (Megehee, Dobie, & Grant, 2003). Relative to the original version at normal rate (100%), the time-compressed version (by –15%, to 85% of the original duration) yielded a more positive attitude to the speaker ("trustworthy, secure, favorable", etc.), and a higher number of affective responses to an open-ended question about the advertised product. The time-expanded version (by +15%, to 115%) yielded a more positive attitude to the message, and a higher number of "cognitive" responses to the open-ended question.

The considerable tempo changes discussed above do not pose great problems for the listener. Speech compressed to 65% of its original duration (by –35%) is still reported to be "perfectly intelligible" (Janse, 2004:160). And if speech is time-compressed even further, to 35% of its original duration (by –65%), which amounts to a highly unnatural speaking rate, even then intelligibility does not fall below 53% correct identifications for real words (Janse, Nooteboom, & Quené, 2003).

In conclusion, this study suggests that the just noticeable difference for tempo in speech is about 10%. Further research is necessary, however, to eliminate a possible downward bias in this reported JND. Tempo variations between speakers and within speakers typically exceed this provisional JND, which adds to the importance of these tempo variations in speech communication.

Acknowledgements

My sincere thanks are due to Grace Postma and Petra van de Ree, for conducting both experiments reported here, to Vincent van Heuven and Sieb Nooteboom for helpful comments, to Hans Van de Velde for assistance in the corpus analysis, and to Theo Veenker for technical assistance.

References

Benguerel, A.-P., & D'Arcy, J. (1986). Time-warping and the perception of rhythm in speech. *Journal of Phonetics, 14*(2), 231-246.

Chafe, W. (2002). Prosody and emotion in a sample of real speech. In P. H. Fries, M. Cummings, D. Lockwood & W. Spruiell (Eds.), *Relations and Functions Within and Around Language* (pp. 277-315). London: Continuum.

Den Os, E. A. (1985). Perception of speech rate of Dutch and Italian utterances. *Phonetica, 42*, 124-134.

Drake, C., & Botte, M. C. (1993). Tempo sensitivity in auditory sequences: Evidence for a multiple-look model. *Perception & Psychophysics, 54*(3), 277-286.

Eefting, W., & Nooteboom, S. G. (1993). Accentuation, information value and word duration: effects on speech production, naturalness and sentence processing. In V. J. Van Heuven & L. C. W. Pols (Eds.), *Analysis and synthesis of speech: Strategic research towards high-quality text-to-speech generation* (pp. 225-240). Berlin: Mouton de Gruyter. Speech Research; 11.

Eefting, W., & Rietveld, A. C. M. (1989). Just Noticeable Differences of articulation rate at sentence level. *Speech Communication, 8*, 355-361.

Ellis, M. C. (1991). Thresholds for detecting tempo change. *Psychology of Music, 19*(1), 164-169.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in Spontaneous Speech*. London: Academic Press.

Grondin, S. (2003). *Detection of tempo accelerations and decelerations with abrupt and gradual variations.* Paper presented at the Rhythm Perception and Production Workshop, Île de Tatihou, France.

Humes, J. C. (2002). *Speak Like Churchill, Stand Like Lincoln: 21 Powerful secrets of history's greatest speakers*. New York: Three Rivers Press.

Janse, E. (2004). Word perception in fast speech: Artificially time-compressed vs. naturally produced fast speech. *Speech Communication, 42*, 155-173.

Janse, E., Nooteboom, S. G., & Quené, H. (2003). Word-level intelligibility of time-compressed speech: Prosodic and segmental factors. *Speech Communication, 41*(2-3), 287-301.

Levitin, D. J., & Cook, P. R. (1996). Memory for musical tempo: Additional evidence that auditory memory is absolute. *Perception & Psychophysics, 58*(6), 927-935.

Lindblom, B. E. F. (1989). Explaining phonetic variation: A sketch of the H&H theory. In W. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 403-439). Dordrecht: Kluwer.

Megehee, C. M., Dobie, K., & Grant, J. (2003). Time versus pause manipulation in communications directed to the young adult population: Does it matter? *Journal of Advertising Research, 43*(3), 281-292.

Nooteboom, S. G., & Cohen, A. (1984). *Spreken en Verstaan: Een nieuwe inleiding tot de experimentele fonetiek* (2nd ed.). Assen: Van Gorcum.

Nooteboom, S. G., & Eefting, W. (1994). Evidence for the adaptive nature of speech on the phrase level and below. *Phonetica, 51*(1-3), 92-98.

Oostdijk, N. (2000). Het Corpus Gesproken Nederlands. *Nederlandse Taalkunde 5* (3), 280-284.

Perron, M. (1994). *Checking tempo stability of MIDI sequencers.* Paper presented at the 97th Convention of the Audio Engineering Society, San Francisco.

Pfeuty, M., Ragot, R., & Pouthas, V. (2003). Processes involved in tempo perception: A CNV analysis. *Psychophysiology, 40*(1), 69-76.

Stetson, R. H. (1988). *Motor Phonetics* (retrospective edition, edited by J.A.S. Kelso & K.G. Munhall). Boston: Little, Brown and Company.

Van Heuven, V. J. (2003). Vervlakt het Nederlands? Over intonatie. In J. Stroop (Ed.), *Waar gaat het Nederlands naartoe? Panorama van een taal* (pp. 215-223). Amsterdam: Bert Bakker.

Van Hout, R., De Schutter, G., De Crom, E., Huinck, W., Kloots, H., & Van de Velde, H. (1999). De uitspraak van het Standaard-Nederlands: variatie en varianten in Vlaanderen en Nederland. In E. Huls & B. Weltens (Eds.), *Artikelen van de Derde Sociolinguïstische Conferentie* (pp. 183-196). Delft: Eburon.

Zwaardemaker, H., & Eijkman, L. P. H. (1928). *Leerboek der Phonetiek: inzonderheid met betrekking tot het Standaard-Nederlandsch*. Haarlem: Erven F. Bohn.

Do H*L and L*H accents have similar target positions?

Toni Rietveld & Joop Kerkhoff

Radboud University Nijmegen

Abstract

Targets of pitch accents have generally been defined on the basis of production data. Obviously, most of the research on tonal targets has focussed on peaks associated with H*L accents as these are very well identifiable, in contrast to the lows associated with L*. In the perception experiments reported here it is made clear that, in Dutch, the conventional target positions of L* and H* (being the F_0-valley and the F_0-peak respectively) do not coincide. The valley associated with L* which gives rise to the perception of a sentence accent occurs much earlier in a syllable than the peak associated with H*. No significant effect was found for the position of the accent: nuclear or prenuclear. Confirmation was found for the existence of the phonological contrast between H* H% and L*H H% in Dutch.

1 Introduction

In the autosegmental tradition F_0-targets have been generally defined on the basis of production data; they are "considered to be identifiable points in the F_0-contour which are aligned with the segmental string in extremely consistent ways" (Ladd, 1996:67). Knowledge about the alignment of targets in production does not necessarily mean that we know which part of a pitch contour associated with a pitch accent is used by the listener to *locate* the accent. As a matter of fact, quite a number of 'turning points' are candidates. For an H*L accent obvious candidates are the start of the rise and the position of the F_0-peak. For L*H-accents the relevant turning points are less evident. The fall associated with a %L L*H-accent is quite small, and is sometimes hardly distinguishable from the low onset. Obviously, most of the research on tonal targets has focussed on peaks associated with H*L accents as these are very well identifiable, in contrast to the lows associated with L*. Arvatani, Ladd & Mennen (1998) presented one of the few investigations on the alignment of L*-accents (for Greek); the F_0-minimum was found to be aligned in a consistent way: 5 ms before the onset of the accented syllable. In many respects the pitch accent L*H takes a somewhat special position in intonation: it seems to be acquired in a rather late stage of speech development (Bolinger, 1989). To 'novices' in the study of intonation it looks often counterintuitive that focus can be signalled by a low target: in pilot studies we find again and again that naive subjects tend to assign a pitch accent to the syllable in which the final F_0-point of the H of L*H is located.

Doubts about the validity of the assumed equivalence of high and low target F_0-values for H*L and L*H might come from the difference in tonal prominence between reaching the 'peak' and the 'valley', respectively, of the two pitch accents at issue. We think, therefore, that the concept of 'target' should also be considered from a perceptual point of view; the comparison of the perceptual processing of the pitch accents H*L and L*H might shed more light on the correspondence between productive and perceptual targets. Furthermore, it is difficult to determine which tonal event in a nuclear L*H-accent gives rise to the perceived accent: the early valley associated with nuclear L*, or the high F_0-value associated with the following H, which is (nearly) in the same syllable (σ_1), see Figure 1.

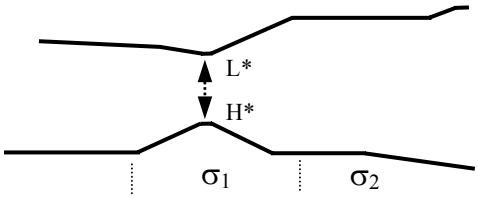


Figure 1. Possible targets of L*H and H*L.

House (1989) developed a model according to which tonal movements through areas of spectral change will be optimally categorized as levels, instead of movements: a falling movement as Low, and a rising movement as High. House assumes that at vowel onset the perceptual mechanism is maximally loaded with the task of resolving spectral information; thus its capacity to resolve F_0-movements is decreased. This would mean that a falling F_0-movement extending over the CV-boundary leads to the perception of a Low in the vowel, and a rise extending over the CV-boundary to the perception of a High. Thus, in a nuclear L*H-accent, a steep rise extending over the VC-boundary, might be perceived as a 'high' target on the vowel following the V associated with L*. As a consequence this vowel might be perceived as carrying the accent, because in the competition of high and low targets it is likely to dominate (Wales & Taylor, 1987).

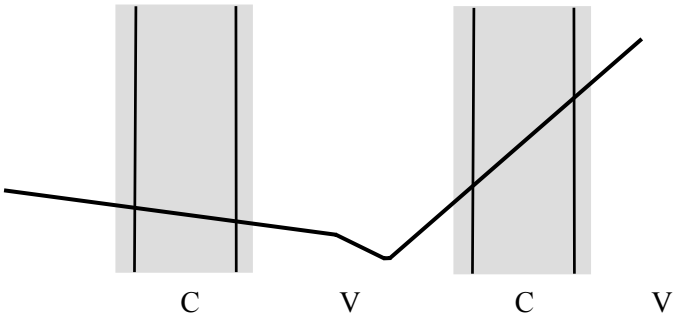


Figure 2. According to House (1989) tonal movements through areas of spectral change (CV and VC) will be categorized as levels, instead of movements: a falling movement as L, a rising movement as H.

In the experiments to be reported here we assessed to what extent the perceptual effects of H*L and L*H accents are similar when the initial boundary tone is %L; this initial boundary tone was chosen as it yields the smallest fall associated with L*. Perceptual effects are restricted to the perceived position of a pitch accent in the syllable as a function of specific tonal events, like reaching a target F_0-value (F_0-maximum for H*L and F_0-minimum for L*H). Furthermore we want to show that the perceptual effects of the pitch accent L*H differ as a function of the location of this accent in the sentence. We expect nuclear L*H-accents, with quite steep F_0-movements to the target of H, to yield different perceptual reactions to shifts in the position of the L*-target from non-final L*H-accents, in which the linking with the following accent leads to smoothly rising F_0-movements (cf. Gussenhoven & Rietveld, 2000). This specific position also creates the possibility to assess to what extent listeners use global information, based on the perception of a whole pitch contour, or use specific tonal events, like a pitch valley, associated with L*. The question whether subjects use global

information to locate pitch accents can be answered by deleting the valley associated with L* in %L L*H H%-sequences, and presenting the resulting pitch contour to listeners.

There is another, less general, because language-specific question which needs confirming evidence. Both in the descriptions of Dutch intonation of 't Hart, Collier & Cohen (1990) and in that of Gussenhoven & Rietveld (1991) the contrast between H* H% and L*H H% was not recognized; the latter equated Pierrehumbert's (1980) H* H% with an L*H H% contour pronounced with wide span in a higher pitch register. Gussenhoven & Rietveld (2000) concluded on the basis of scores on gradient paralinguistic attributes (like the degree of 'SURPRISE' conveyed by these contours) that this contrast does exist in Dutch. Establishing differences in the target positions of both contours would be regarded as final confirmation.

To summarize, we aim at finding answers to the following questions:

Q1: Do the perceptual target spaces of H*L and L*H have the same alignment?

Q2: Is the perceptual target of L*H the same in nuclear and prenuclear positions?

Q3: Do listeners use global F_0-information when specific local F_0-cues are missing?

Q4: Does the contrast between H* H% and L* H H% contours exist in Dutch?

2 Method

2.1 Speech materials

Two source utterances were synthesized with the Nijmegen/MBROLA diphone synthesis system (male voice; sampling frequency 16 kHz): *hij wil mo mo verlaten* ('he wants to leave mo mo') and *hij wil mo mo voor een tijdje verlaten* ('he wants to leave mo mo for some time') with the quasi nonsense syllables *mo1* (the first *mo* in the utterance) and *mo2* (the second *mo*), which together suggest a name,). No durational, intensity or spectral cues were assigned to either *mo1* or *mo2* to signal the location of the accent.

Six experimental contours were assigned to the source utterances, the first three to the shorter, the latter three to the longer utterance:

(1)	%L H*L L%	('pointed hat')
(2)	%L H* H%	(equivalent to a %L L*H H%-contour without valley associated with L*)
(3)	%L L*H H%	(valley of default 12 Hz).
(4)	%L H*LH !H*LH H%	(prenuclear accent on '*la*' of *verlaten* - 'to leave' -)
(5)	%L L*H !H*LH H%	(prenuclear accent on '*la*' of *verlaten*: linked contour)
(6)	%L L_0*H !H*LH H%	(as (5), but without a valley associated with L*, L_0 stands for a deleted L-target)

In Figure 3 the six contours are depicted.

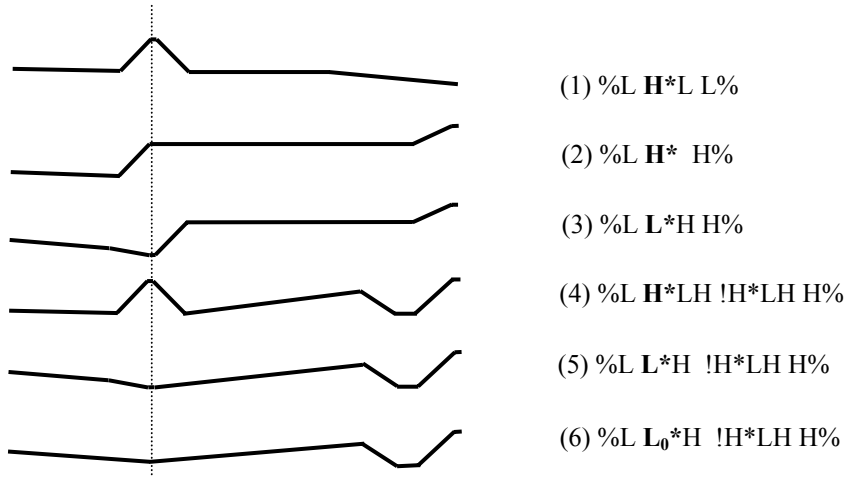


Figure 3. The alignments of the six contours used in the perception experiments. The dotted line marks the conventional targets of the pitch accent, at issue.

These contours were presented to subjects, with different (shifted) temporal locations of the targets; the step size was 20 ms, see Figure 4.

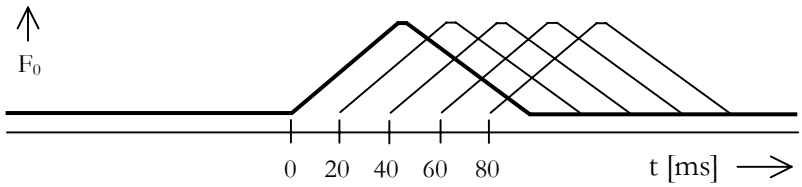


Figure 4. Shifts of the pitch configurations in steps of 20 ms: example for H*L .

The task of the subjects was to indicate whether the accent is on the first or on the second *mo*.

Below we give a detailed account of the information the responses to the six contours (see Figure 3) can give us:

(1) The first contour (with H*L) can be seen as an anchor contour, with a generally accepted target, the F_0-peak.

(2) The second contour (with H* H%) has nearly the same make-up as the L*H-contours, but is not realised with a valley. This contour enables us to find out two things:

a. whether the same F_0-movement, associated with reaching H*, will be interpreted differently as a function of the tonal material which follows; this would mean that listeners try to give an interpretation to the whole contour, and do not (only) use specific target values, and
b. to which extent this contour is processed differently from similar contours in which the valley is realized (contour 3). If not, than we have to assume that the valley itself is not a crucial part of this contour. The comparison with contour 3 (%L L*H H%) enables us to assess whether these contours constitute a phonological contrast.

(3) Contour 3 is a default realisation of the L*H-accent; this contour provides information on the question whether the hypothesis is correct that the highest and lowest F_0-value of H*L and L*H-accents, respectively, should be seen as the targets of these accents.

(4) Contour 4 is a realisation of the %L H*LH !H*LH H% accent; it is the two-accented stimulus to be used as 'standard' to contours (5) and (6) in which the prenuclear accent is L*H.

(5) Contour 5 is a default realisation of the L*H accent in prenuclear position (linked with H*LH on *la* of *verlaten*). Thus, a smooth pitch rise to the following accent can be realized. If shifting L*-targets yields accent assignments that differ from those obtained in the single L*H-accents of contour 3, we will have extra evidence that L*H is not a unit, which functions independently of tonal context, like – as we predict – in the case for H*L

(6) Contour 6 is a non-default realization of the L*H-accent, in that the valley associated with L* is deleted (L_0*). If subjects still exhibit the same perceptual patterns as with contour 5, we will have evidence that they use global, contour-based information to assign pitch accents, and not only specific tonal events, strongly associated with the 'accented'syllable.

2.1.1 Physical characteristics of the contours

The duration of the movement towards the target of H* is 100 ms (93 Hz to 147 Hz), the duration of the F_0-movement towards the valley is also 100 ms; the valleys in the %L L*H H% accents are 12 Hz below the preceding last F_0 value of the onset. The %L-onsets start at 104 Hz and end at 93 Hz. The duration of the H% movement is 120 ms, it covers 33 Hz with 180 Hz as endpoint. The F_0-maxima of H* and H in L*H are the same: 147 Hz.

The respective durations of [m] and [o:] are the same in both syllables *mo*. The positions of the conventional targets for L*H and H*L (start of the valley and position of the maximum F_0-value) are shifted in a variable number of steps (between 10 and 15) of 20 ms from different starting points.

In Table 1 we give the start and end of the segments in the two syllables *mo*, and positions of the turning points of the five experimental contours.

Table 1. Start and end of segments in the two syllables *mo*, and positions of turning points in ms (beginnings of conventional targets: maximum F_0 for H* and minimum F_0 for L*, see fig. 3) of the six experimental contours, starting from the 'left'.

	m	o	m	o
Start/end of segment	425-485	485-630	630-690	690-835
Contour				
(1) %L **H***L L%		560		740
(2) %L **H*** H%		600		780
(3) %L **L***H H%		500	680	
(4) %L **H***LH !H*LH H%		520		760
(5) %L **L***H !H*LH H%		520		760
(6) %L **L_0***H !H*LH H%		520		760

The total duration of the short sentences (with contours 1, 2 and 3) was 1510 ms, that of the long sentences (with contours 4, 5 and 6) was 2160 ms.

2.2 Task and procedure

Subjects had to tell whether they heard the name *MOmo* (*mo1* with accent ('stress') on the first syllable: 'a') or *moMO* (*mo2* with accent on the second syllable: 'b').

Two methods of presentation were used, (A) that of minimal changes, in which the stimuli of a specific contour are presented in a specific order, either with F_0-targets shifting to the right ('ascending') or to the left ('descending'), and (B) that of random presentation, in which the stimuli were presented in random order. The stimuli were presented to four groups of 10 subjects, to whom different methods of presentation were assigned (either method A or method B, with 20 subjects each) and different orders of stimuli. Each utterance is followed by an response interval of 3 seconds.

For *presentation method A*, each block of stimuli was preceded by two 'anchor stimuli': a default 'a' and a default 'b' contour with unambiguous accents on the first and the second *mo* syllable, respectively. The anchor stimuli always corresponded with the particular contour to be presented in that group. In the instruction for the listeners it was told that the number of stimuli per block could vary between 10 to 15; thus we tried to avoid expectations about the stimulus in which the change from *a* to *b* or vice-versa had to be expected. Accordingly the score form contained 15 answer positions.

For *presentation method B*, the whole experiment was preceded by default realisations of all contours to be presented, with 'clear' accents on syllables 'a' or 'b' respectively. Each presentation cost about 10 minutes. Stimuli were presented over earphones; subjects were free to determine the start of the presentation of each group of stimuli.

3 Results

Not all our data fit the psychometric curve; in the data obtained with method A the transition from 'a' to 'b' judgements took place in a very categorical way. That is why it was decided to carry out analysis of variance of the randomized block type on the data obtained with method A, followed by post-hoc comparisons. The six contours make up the within-subject factor 'treatment'. For each subject the time (in ms) is given at which the target of a particular contour gives rise to a 'b' judgement. Subsequent post-hoc comparisons show which contours have different or similar 'perceptual target values'. For method B the application of the conventional psychometric curve fitting procedure was not problematic for most contours, because the transitions from 'a' to 'b' judgements were much smoother, within subjects and pooled over subjects.

3.1 Results obtained with method A

In Table 2b we present the mean transition values, pooled over 20 subjects; for clarity's sake we reproduce part of Table 1 in Table 2a, with the starts and ends of the segments at issue.

Table 2a. Locations of starts and ends of the segments of *mo mo,* in ms.

m	o	m	o
425…485	485…630	630…690	690…835

Table 2b. Mean locations of transition from 'a' to 'b' responses, in ms, obtained with method A. Each value is based on 20 observations. Relevant targets (L* or H*) printed in bold.

Contour	Position of transition
(1) %L **H***L L%	685
(2) %L **H*** H%	701
(3) %L **L***H H%	628
(4) %L **H***LH !H*LH H%	704
(5) %L **L***H !H*LH H%	644
(6) %L **L_0***H !H*LH H%	655 (?)

The perceptual effects of shifting the target of **H*** in %L **H***L L% (contour 1) and %L **H*** H% (contour 2) are nearly the same. In %L **H***L L% the location of the target at 685 ms leads to the perception of an accent shift from *mo*1 to *mo*2, and in %L **H*** H% at 701 ms, a difference within the step size of 20 ms. This is the location of the onset of [o:] in *mo*2.

For **L*** in %L **L***H H% (contour 3) the position is clearly different: the perception of an accent shift from *mo*1 to *mo*2 takes place at 628 ms, much earlier than with H* (at 701 ms). This is apparently the position where H following L* reaches its maximum value (at 728 ms). This means that the perceptual anchors of H* and L* are not the same in final accents.

It is not surprising – at least from a perspective which focuses on the use listeners may make of physical cues – that the perceptual effect of shifting the target of **H*** in %L **H***LH !H*LH H% (contour 4) is about the same as in both %L **H***L L% (contour 1) and %L **H*** H% (contour 2): the perceptual shift took place around 700 ms. For **L*** in %L **L***H !H*LH H% (contour 5) the situation is not so clear. In the series in which the target was shifted from left to right, the target position of L* at which *mo*2 was perceived as accented was 670 ms, whereas in the series with right to left shifting, the location was 620 ms, a difference of about 2 stimulus shifts. They are not so different from the position of **L*** in %L **L***H H% (contour 3), while both are different from those observed with an H*-accent.

Very unclear are the perceptual effects of contour (6) %L **L_0***H !H*LH H% in which **L*** is not marked by a valley. Nine of the 20 subjects did not hear any transition, two others started a series with a perceptual localisation opposite to the direction of the shifts. The nine remaining subjects located the shift at 655 ms.

An ANOVA was carried out on the location of the transition for each contour, apart from contour (6): %L L_0*H !H*LH H%, for which a clear transition could not be established; the (fixed) factor 'contour' was significant: $F(4,56)=24.69$, $p<.01$. Post-hoc comparisons (Tukey's HSD procedure; $\alpha=.05$) yielded the following homogenous subsets (mean values of transition locations given in ms):

Table 3. Homogeneous subsets of contours (time position of accent transition, in ms).

Contour	subset 1	subset 2
(3) %L L*H H%	628	
(5) %L L*H !H*LH H%	644	
(1) %L H*L L%		685
(2) %L H* H%		701
(4) %L H*LH !H*LH H%		704

Table 3 reveals that there is a clear distinction between alignments associated with L* and H*-accents respectively. The existence of a contrast between %L L*H H% (3) and %L H* H% (2) is confirmed again.

3.2 Results obtained with method B

The results obtained with method B were processed in the conventional way: the percentages of judgements obtained by pooling over subjects were probit-transformed and the stimulus value which coincided with z=0 was regarded as the Point of Subjective Equality (PSE). None of the associated χ^2 values used to test the fit were significant at the .05 level, which reflects goodness-of-fit, except for the results of one contour: %L H*LH !H*LH H% (contour 4). For this contour $\chi^2(7)=30.86$, $p<.01$. The reason for this poor fit was the steep transition from 'a' to 'b' judgement. As estimate of the PSE we took the position at which the majority of the subjects had given a 'b' judgement.

For contour (6) %L L_0*H !H*LH H% (without a valley for L*) neither a conventional sigmoid was observed (its direction was even nearly reversed, and the percentages of 'b' responses varied around 60%), nor a clear transition in the perceived location of the accent. In fact, the situation was analogous to the one found with method A.

In Table 4 we present the mean transition values obtained for the six experimental contours, obtained with the two methods. Each mean value is based on 20 observations.

Table 4. Mean locations of transition from 'a' to 'b' responses in ms, obtained with methods A (series) and B (random order). Each value is based on 20 observations.

Contour	Transition (A)	Transition (B)
(1) %L H*L L%	685	671
(2) %L H* H%	701	704
(3) %L L*H H%	628	631
(4) %L H*LH !H*LH H%	704	700
(5) %L L*H !H*LH H%	644	640
(6) %L L_0*H !H*LH H%	655 (?)	737 (?)

Table 4 shows similar transition points for all but one contour (number 6); with both methods A and B we see nearly equal points of transition for accents with H* and with L* respectively.

In Figure 5 we summarise our findings by showing the alignments of the contours which mark the positions of Perceptual Subjective Equality; the alignment of contour %L L_0*H !H*LH H% (6) is not given, as we could not determine a clear PSE for it. (The slight differences of tonal height among the H*-contours only serve pictorial clarity).

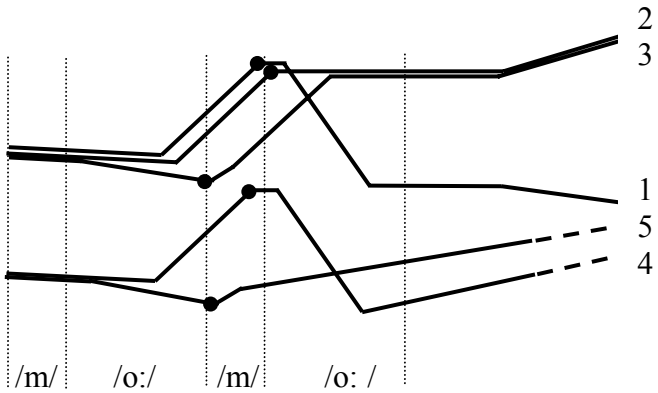


Figure 5. Positions of the conventional targets (for H* the beginning of the F_0-maximum and for L* the beginning of the valley, ● in the figure) that give rise to the perception of an accent on the second 'mo' in the series. The numbers refer to the contours given in Table 4.

The results summarised in Table 3 and Figure 5 make four things very clear. It is not correct to equate the conventional alignments of L* and H*-accents, viz. the location of the start of the valley and the location of the F_0 maximum. The small valley associated with L* is an important cue; we do not know, however, whether this cue plays a local role, or whether it triggers the use of the rest of the contour to locate the accent. No difference was found between the perceptual processing of nuclear and prenuclear pitch accents. The H* H% and L*H H% contours constitute a phonological contrast, as the timing of the associated target values is completely and significantly different.

4 Conclusion

The perceptual shift from an accent on the first syllable of the sequence *mo mo* to the second takes place earlier with L*H than with H*L accents, at least when the targets are expressed in the positions of the conventional targets of these accents. The difference, pooled over the variants of the two accent types amounts to about 50 ms, and confirms the phonological contrast between H* H% and L*H H% contours in Dutch, earlier established on the basis of paralinguistic scale judgements.

Whereas the presence of a valley at the onset of the sonorant consonant of the second syllable *mo* already gives rise to the perception of a sentence accent, the target of H* needs to be located at the onset of the vowel of that syllable, to be prominence lending. One of our hypotheses was that the H following the L* in nuclear accents might compete with the F_0-valley associated with L*, analogously to the way in which H-targets can compete with L*-targets in the degree of perceived prominence (cf. Gussenhoven & Rietveld, 2000). The introduction of L* in prenuclear position in contour (5) %L L*H !H*LH H% gave us the opportunity to test this hypothesis. Indeed, the target of prenuclear L* (with the following slowly rising H) gave rise to a perceived accent on the second syllable associated with a later target; the difference is 16 ms, but it is not statistically significant at the 5% level. Thus, until other evidence becomes available, we have to assume that the targets of L* in both prenuclear and nuclear positions behave similarly, but they should be accounted for by different explanations. For nuclear L*H, House's theory (1989) might throw light on the results. He assumes that at vowel onset the perceptual mechanism is maximally loaded due to the spectral instability at vowel onset; consequently its capacity to resolve F_0-movements is decreased. A rise extending over the onset of the vowel following the position of the low target might be 'categorized' as a high target in that vowel, and consequently lead to the

perception of a H*-accent on that vowel. If the steep rise leads to a relatively high F_0 value in the same vowel in which the F_0 valley occurs, it might even be perceived as delayed H*. Thus, an L*H accent might go through different perceived accents, L* and H*, depending on the location of both L*-targets and targets and movements associated with the following H.

However, this explanation does not account for the behaviour of L*H in prenuclear position. First of all, the presence/absence of the small valley in contours (5) %L L*H !H*LH H% and (6) %L L_0*H !H*LH H% appears to be crucial for the perception of a prenuclear L*H accent. At first sight, one might be inclined to think that the hypothesis of 'global perception' of contours must be rejected, as a local cue – the small valley of 12 Hz with a duration of only 15 ms, preceded by a downwards slope from 104 to 93 Hz over 100 ms – plays a role in the perception of the associated accent. We think however that it is not wise to dismiss the concept of 'global perception' altogether. In Dutch, %L (–) H*L is not a possible contour with a slowly rising pitch, starting somewhere at (–) in the contour before H*, unless the start of the slope is associated with either the realisation of an H*L or an L*H-accent. This constraint on possible contours might make listeners search for F_0-cues which mark the presence of an L*H-accent, as an H*L-accent is clearly lacking. Thus, 'global perception' might interact with the presence/absence of local cues, and give rise to the perception of accents.

References

Arvatani, A., Ladd, D.R., Mennen, I. (1998). Stability of tonal alignment: The case of Greek prenuclear accents. *Journal of Phonetics, 26*, 3-25.

Bolinger, D. (1989). *Intonation and its uses: melody in grammar and discourse*. London: Edward Arnold.

Gussenhoven, C.H.M. & Rietveld, A.C.M. (1991). An experimental evaluation of two nuclear tone taxonomies. *Linguistics, 29,* 423-449.

Gussenhoven, C. & Rietveld, T. (2000). The behaviour of H* and L* under variations in pitch range in Dutch rising contours. *Language and Speech, 43*, 183-203.

't Hart, J., Collier, R., Cohen, A. (1990). *A perceptual study of intonation.* Cambridge: Cambridge University Press.

House, D. (1989). Perceptual constraints and tonal features. *Working Papers, Department of Linguistics, Lund University, 35*: 113-120.

Ladd, D.R. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. PhD. thesis, MIT.

Wales, R. & Taylor, S. (1987). Intonation cues to questions and statements: how are they perceived? *Language and Speech*, *30*, 199-210.

The implicit prosody of Jabberwocky and the relative clause attachment riddle *

Frank Wijnen

Utrecht University

Abstract

This contribution reports on three sentence processing experiments involving structurally ambiguous *Jabberwocky* strings. The items are complex NPs, in which a relative clause can be attached to either of two nominal heads (NP1-preposition-NP2-RC). The phonological lengths of the heads, as well as that of the relative clause, are manipulated. The results show modest effects on relative clause attachment preferences of this manipulation, thus lending support to Fodor's (1998, 2002) *implicit-prosody hypothesis*.

1 Introduction

One of Sieb Nooteboom's relished criticisms of psycholinguistic research is that the distance between the primary data and the processes of interest is big, too big actually. He may be right – it often requires an inferential chain of considerable length and complexity to connect something pedestrian like reaction times to something lofty like lexical access or grammatical encoding. To my mind, however, psycholinguistics does not differ very much in this respect from other domains of investigation, including phonetics, definitely also including some of the work done by the distinguished scientist to whom this book is dedicated. Consequently, I consider it quite appropriate to present a study here in which the connection between data and the presumed underlying processes is exquisitely indirect.

The hypothesis under scrutiny is the implicit-prosody hypothesis, as put forth by Janet Fodor in several publications in the past few years. In Fodor (2002), the following summary is given:

> The *Implicit-Prosody Hypothesis* (IPH):
>
> In silent reading, a default prosodic contour is projected onto the stimulus, and it may influence syntactic ambiguity resolution. Other things being equal, the parser favors the syntactic analysis associated with the most natural (default) prosodic contour for the construction.

To avoid any misunderstandings, it is useful to point out that prosody (intonation, phrasing and rhythm) can affect sentence processing in the *auditory* modality (see Cutler, Dahan, & Van Donselaar, 1997 for an extensive review). In particular, various studies have demonstrated that accentuation and phrasing affect the on-line resolution of syntactic ambiguities, if at least the two parses can be associated with different prosodic realizations (e.g. Nagel, Shapiro, Tuller, & Nawy, 1996; Pynte & Prieur, 1996; Schafer, Carter, Clifton, & Frazier, 1996). A point of debate has been whether prosody has an immediate effect on

* Thanks are due to Carolien van den Hazelkamp, who ran Experiment III. I am also grateful to Janet Fodor, Jocelyn Ballantyne, Martin Corley, and this volume's editors, for helpful commentary on a first draft. The author was supported by a fellowship of the Royal Netherlands Academy of Arts and Sciences (KNAW).

ambiguity resolution, i.e., affecting the first-pass parse, or whether it is invoked only in cases where the perceiver is 'garden-pathed' (see e.g. Watt & Murray, 1996). Recent neurocognitive evidence suggests that prosodic cues are exploited immediately (Steinhauer, Alter, & Friederici, 1999).

The implicit-prosody hypothesis, being a well-behaved psycholinguistic hypothesis, makes reference to an invisible cognitive entity, viz. a phonological (prosodic) representation generated by the reader. Now, how should we go about collecting evidence for this construct? Fodor proposes the following research programme:

- isolate a factor F that can be manipulated and that measurably affects the prosodic structure of a sentence;
- show that this prosodic structure resulting from F has an effect on sentence parsing (e.g., ambiguity resolution);
- demonstrate that F does *not* affect parsing directly;
- include F in a reading task. Does the task reveal an effect of F on parsing (just as in an auditory task)? If so, we may infer that the prosody induced by F is in some sense 'real', even if not manifestly (physically) present.

The present study follows this recipe by exploring the effects of phonological weight (i.e., length in syllables) on the resolution of relative clause attachment ambiguities in a complex noun phrase with two possible relative clause (RC) heads, as illustrated in (1).

(1.) Iemand schoot op de bediende van de actrice die op het balkon zat.
'Someone shot the servant of the actress who was on the balcony.'

Clearly, the relative clause *die op het balkon zat* 'who was on the balcony' can be attached to either *de bediende* 'the servant' (NP1) or *de actrice* 'the actress' (NP2). The implicit-prosody hypothesis is of particular relevance to this ambiguity. Universalist theories of sentence processing (e.g. Frazier & Fodor, 1978; Kimball, 1973) predict a consistent preference for low (i.e., NP2) attachment of the relative clause, on the basis of a structurally defined strategy, *late closure*. Cuetos & Mitchell (1988) showed, however, that attachment preferences in the construction at hand vary across languages, and numerous studies since then have confirmed and extended this observation (see Ehrlich, Fernández, Fodor, Stenshoel, & Vinereanu, 1999; Fodor, 2002, for overviews).

Fodor has argued that the potential threat to universalist parsing models posed by these observations may be removed, and the cross-linguistic variation better understood at the same time, if we assume that relative-clause attachment in this context is guided by prosodic factors. In line with the implicit-prosody hypothesis, Fodor (1998) assumes that readers engage their phonological processor to 'chop up' the input into manageable chunks or packages. Prosodic packaging in silent reading is fully analogous to what happens in speaking. A phonological factor that leads to the insertion of a prosodic break in speaking will equally likely promote the insertion of a package boundary in reading. Furthermore, in processing written input, the syntactic processor is assumed to refrain from rearranging the packages delivered by the prosodic processor (but see Fodor, 2002).

One of the factors that affect prosodic packaging is phonological length (weight). Prosodic boundaries have a tendency to occur at roughly equal distances in the speech stream. A corollary of this tendency is that long, polysyllabic words tend to form their own

phonological phrases, whereas short words are more likely to be chunked together. Phonological length can affect the resolution of relative-clause attachment ambiguities of the type exemplified in (1) in various ways. First, the length of the relative clause itself matters: long (heavy) relative clauses tend to attach high, whereas short (light) ones tend to attach low, a phenomenon known as the 'anti-gravity effect'. The explanation Fodor advances is that a short RC is more easily packaged with the second of the two noun phrases, whereas a long RC promotes the insertion of a prosodic break right before it, which blocks its attachment to the second of the two NPs. The lengths of the potential head nouns (or NPs) are predicted to have an effect as well (Fodor, 1998). If, in the construction NP1-preposition-NP2-RC, NP1 is long and NP2 is short, there will be a tendency to insert a prosodic boundary after NP1. This boundary will render attachment of the relative clause to NP1 difficult, and consequently NP2-attachment will be the preferred option. If, however, NP1 is short and NP2 is long, a boundary is likely to be inserted after NP2, rendering the attachment of the relative clause to NP2 difficult, resulting in an NP1 attachment bias. To my knowledge, there is as yet no direct evidence in support of this prediction, either from auditory or visual language processing. Colonna & Pynte (2001) have reported data that seem to go against this prediction (from French), but their materials are not without problems.

The aim of the present study is to provide further evidence for the effects of phonological length on the resolution of relative-clause attachment ambiguities in the structure exemplified in (1). In an attempt to neutralize as much as possible the influence of lexical, semantic and pragmatic factors on attachment decisions, the three experiments reported here make use of "Jabberwocky prose", i.e., strings in which all content words have been replaced by non-existing (but phonologically legal) elements. Such strings pose no problem for the human parser. Recent ERP evidence shows that they engage the syntactic processing system just like normal sentences do (Hahne & Jescheniak, 2001). They cannot be interpreted, however, at least not in the normal sense of the word, as they prevent (normal) access to the lexicon and the semantic system.

The first experiment is a questionnaire study looking at the effect of relative phonological weights of the attachment sites (NP1 vs. NP2). Experiment II is a replication of Experiment I with relative clauses that are longer than those in Experiment I. These two off-line experiments involve globally ambiguous materials. This means that readers cannot be 'garden-pathed', i.e., pursue an analysis that is proven incorrect downstream in the input by lexical or structural cues. The main aim of Experiment III is to determine whether phonological length of the potential heads of a relative clause can have an effect in an on-line task (continuous acceptability judgment) as well.

2 Experiment I

2.1 Method

Participants. Twenty-two volunteers (10 female) took part, the majority of them being faculty and students at Utrecht University.

Materials. Thirty-two items were constructed according to the following template: NP1-preposition-NP2-Relative Clause. These 32 items were divided into four sets of eight, each of which corresponded to one of the four conditions that resulted from systematically varying the lengths of the two critical nouns, as exemplified in (2).

(2.) *long-short*: de kalambulo van de fup die verstritst was
long-long: de knilpatsiera van de astrublankor die verdrimd werd
short-long: de slos van de prefrastiaan die bedrept was
short-short: de vrink van de orcht die betrind werd

Sentence fragments (i.e., complex NPs) were used instead of complete sentences, in order to suppress as much as possible the influence of structural position and its (prosodic) correlates in terms of e.g. the presupposition~focus contrast. Both possible heads of the relative clause were always non-neuter nouns (as indicated by the determiner *de*), so that the relative pronoun (*die*) did not disambiguate the structure through grammatical gender. All relative clauses were passives (which could either be construed as adjectival or verbal), and consisted of three words (four syllables): the relative pronoun, a past participle (marked by an appropriate prefix) and one of the auxiliary verbs *werd* 'was' or *was* 'was'.

Procedure. The participants completed a questionnaire sent by electronic mail. The e-mail message contained the experimental items in a quasi-random order, and the participants were instructed to read the items one by one at a leisurely pace, and to immediately indicate whether they felt the relative clause modified the first or the second noun, by replying with either 1 or 2. They were explicitly requested not to read the stimuli aloud, and were instructed to utter a nonsense phrase after completing each item (the suggestion given was *abracadabra simsalabim*). Both of these recommendations were made in an attempt to counteract the emergence of response biases (or response priming).

2.2 Results and discussion

A total of 704 responses (22 participants × 32 items) were collected. Eight participants showed hardly any variation in their responses (they selected NP1 in merely one or two cases); their data were discarded, leaving 448 valid data points generated by 14 participants.

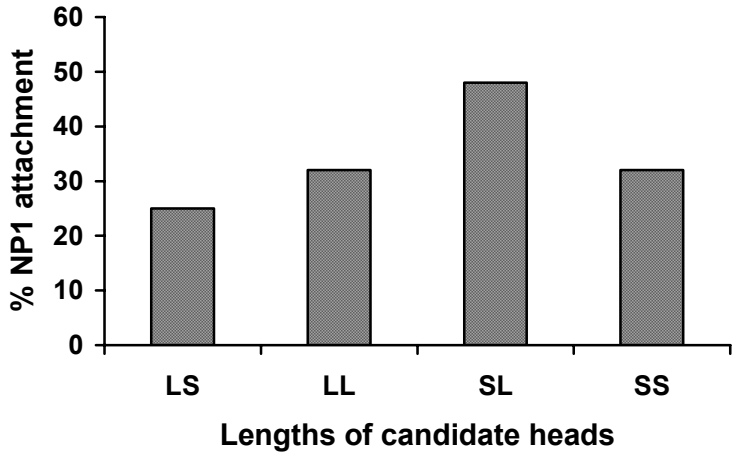


Figure 1. Experiment I: Percentage of NP1 attachment responses as a function of phonological length variation in the potential NP heads of the relative clause. LS: NP1 long, NP2 short; LL: both NPs long; SL: NP1 short, NP2 long; SS: both NPs short.

The first noun was indicated as the head of the relative clause 154 times (34%), reflecting a relatively strong low attachment preference, which is at odds with previous results on

relative-clause attachment in (real) Dutch (Brysbaert & Mitchell, 1996). The percentages of NP1 attachments per condition are displayed in Figure 1.

The distributions of NP1 and NP2 responses across conditions turn out to be significantly different, $\chi^2(3)=13.48$, $p=.004$. The conditions for which the implicit-phonology hypothesis provides clear predictions are LS (long NP1, short NP2) and SL (short NP1, long NP2). The percentage of NP1 responses in the latter (SL) is almost twice as high as in the former (LS): 48% vs. 25%. This difference is significant, $\chi^2(1)=6.08$, $p=.014$.

The first conclusion that can be drawn from these results is that phonological length of the potential nominal heads affects relative clause attachment in reading. It should be stressed that the manipulated variable (i.e., length in syllables) is, indeed, of a purely phonological nature. Moreover, in this particular instance, we may assume the effect is unconfounded by lexico-semantic or pragmatic factors. Therefore, the result can be plausibly ascribed to the phrasing contrast in implicit prosody that results from the length manipulations.

It is also important to reiterate that the present experiment involved globally ambiguous items only. Reanalysis on the basis of lexical or grammatical disambiguation can therefore not have occurred. Consequently, the length contrasts, i.e., their effects on implicit prosody, can not be associated with a 'last resort strategy', applied only in the case of parsing breakdown.

3 Experiment II

3.1 Method

This experiment was conducted in exactly the same way as Experiment I. Eighteen new volunteers (from the same population as in Experiment I; 9 female) participated. The sentence fragments used here were adapted from those in Experiment I by inserting a prepositional phrase before the past participle in the relative clause, consisting of a preposition, a definite determiner and a trisyllabic pseudo-noun, as illustrated in (3) below.

(3.) *long-short*: de kalambulo van de fup die in de strindosnees verstritst was
long-long: de knilpatsiera van de astrublankor die van de spatterplons verdrimd werd
short-long: de slos van de prefrastiaan die voor de elkodrator bedrept was
short-short: de vrink van de orcht die door een kaskadeur betrind werd

3.2 Results and discussion

Two of the participants displayed virtually no variation in attachment decisions; their data were discarded, leaving 512 valid responses. A chi-squared analysis indicates that the distributions of attachment responses across conditions are not reliably different, $\chi^2(3)=1.71$, $p=.64$. Across all conditions and participants, NP1 attachment was chosen in 209 cases (41%). This is a conspicuously higher percentage than in Experiment I; the difference between the response distributions in Experiments I and II is significant [$\chi^2(1)=4.22$, $p=.039$], in agreement with the 'anti-gravity effect'. This effect is explained as the result of a tendency to insert a prosodic break before a long RC, blocking its attachment to NP2. Apparently, this tendency outweighs the effects of the two potential heads' lengths. This issue, obviously, deserves further scrutiny, but it should be clear that, taken together, the results of experiments I and II agree with the implicit-prosody hypothesis.

4 Experiment III

Up to this point, we have considered Dutch native speakers' relative-clause attachment preferences in an unconstrained, off-line judgment situation. The next step in our exploration

is to determine whether the effect of phonological length found in Experiment I generalizes to a more rigorously controlled experimental task. The task used in the present experiment, continuous acceptability judgment, has a hybrid character. On the one hand, it requires participants to deliberately evaluate the acceptability of stimuli, which renders it similar to off-line judgment tasks. On the other hand, the stimuli are presented fragment by fragment, emulating normal reading in some respects (notably: sequential, incremental processing). The continuous acceptability judgment task has been shown to be sensitive to readers' on-line parsing preferences. Crucially, readers tend to judge a sentence fragment as unacceptable not only if it is incorrect in some objective way (e.g. when it is ungrammatical), but also when it contradicts an ambiguity resolution preference. In addition, fragments that cause a parsing problem but are nonetheless accepted, yield longer average reaction times than those of unproblematic fragments.

In contrast to Experiments I and II, the relative-clause attachment in the items used here is disambiguated, by means of grammatical number agreement of the auxiliary verb in the relative clause with one of the two potential head NPs. In all stimuli, the potential head NPs differ in length. The prediction is that LS items (long NP1, short NP2) yield a low (NP2) attachment preference, whereas SL items will promote high (NP1) attachment. Thus, forced NP1 attachment (by the auxiliary) will negatively affect the acceptance rate and average reading times in the LS items, as compared to the SL items. The mirror image of this pattern should be observed for forced NP2 attachment.

4.1 Method

Participants. Fifty-five paid volunteers, all of them students at Utrecht University, took part (18-40 years of age; 11 male). None of them had participated in experiments I or II.

Materials. Thirty-two items were adapted from those used in Experiment I. Each item contained two RC-attachment sites, one containing a long noun (four or five syllables) and the other a monosyllabic noun. Four variants of each item were made by varying the order of the long and the short noun and the attachment disambiguation (NP1 vs. NP2; see example 4). All long nouns had a plural affix (*-s*), whereas the short nouns were always singular. (Most monosyllabic nouns require the syllabic plural affix *-en*, which would have interfered with the length manipulation.) All relative clauses consisted of a relative pronoun, a past participle, and a plural or singular version (both monosyllabic) of the auxiliary *zijn* 'be'.

(4.)(slashes indicate fragment boundaries)

a.	de carambulo's / van / de fup / die / verstritst / is	*long, short, NP2*
b.	de carambulo's / van / de fup / die / verstritst / zijn	*long, short, NP1*
c.	de fup / van / de carambulo's / die / verstritst / is	*short, long, NP1*
d.	de fup / van / de carambulo's / die / verstritst / zijn	*short, long, NP2*

The four versions of the 32 experimental items were divided across four lists, in such a way that each list contained all four conditions in equal numbers, but only one version of each experimental item. The experimental items were interspersed with 64 filler items, all of which had the same structure as the experimental items, but varied with regard to the lengths of the critical nouns, association of grammatical number with long and short nouns, and the identity of the auxiliary verb in the relative clause. Participants were randomly assigned to lists, and items within lists were presented in random order, different for each participant.

Procedure. Participants were seated in front of a computer monitor in a dimly lit, sound-attenuating room. Each trial started with a blank screen, on which appeared a small gray window containing the text 'push a button to proceed'. Upon pressing a button on a button box, the screen was refreshed, and a black window was presented slightly above the vertical midpoint of the screen, with a fixed height of 2 cm, and a variable width, dependent on the length of the item to be presented. The items were presented fragment by fragment, in non-incremental moving-window mode. Participants were instructed to press the right-hand button on the box if they thought the fragment was an acceptable continuation of the preceding material, or the left-hand button if the fragment was an unacceptable continuation. A left-hand button press ended the trial. The intervals between any two subsequent button presses were computed and stored for later analysis. After a trial ended, the screen was erased, and the window with the text 'push a button to proceed' reappeared.

Participants were instructed to read at a pace that allowed them to judge the acceptability of the items. It was stressed they should rely on their first impression, and not unnecessarily delay their response. They were allowed to take breaks between trials. Completion of the experiment took about 25 minutes.

4.2 Results and discussion

Rejection rates. All items consisted of six fragments, of which only the last one (containing the disambiguating auxiliary) is critical. No effects were observed in fragments 1 to 5. In fragment 6, the rejection rates were 37.5% in the LS (long NP1, short NP2) condition, and 38.4% in the SL condition (short NP1, long NP2). This difference is not statistically reliable, by participants $F(1,54)=.22$; by items $F(1,31)=.09$. Across phonological length manipulations, the rejection rates for NP1 and NP2 disambiguation were 31.9% and 43.9%, respectively, which constitutes a significant difference, by participants $F(1,54)=4.09$, $p=.048$; by items $F(1,31)=25.13$, $p<.001$.

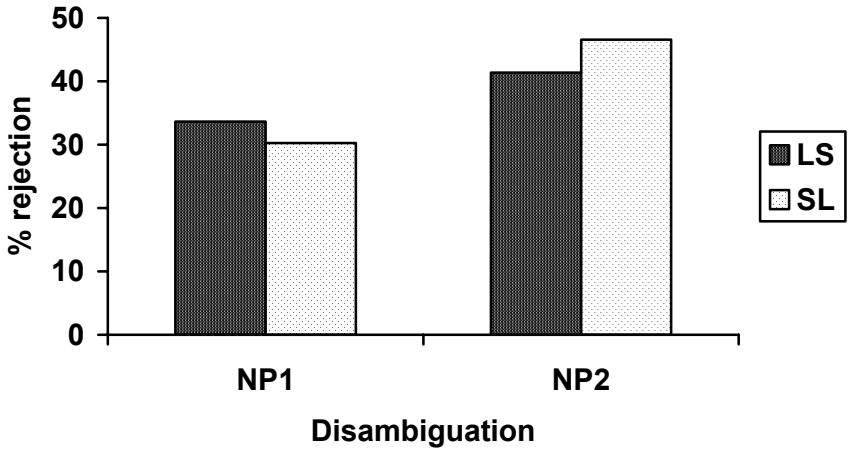


Figure 2. Experiment III: Mean percentage of rejection responses at the disambiguating fragment 6 (auxiliary), as a function of the lengths of NP1 and NP2 (LS: long NP1, short NP2; SL: short NP1, long NP2) and disambiguation direction (NP1 or NP2).

The critical question is whether RC-attachment disambiguation interacts with the manipulation of the lengths of the attachment sites. The relevant data are displayed in Figure 2, which shows that the difference in rejection scores between NP1 and NP2 disambiguations

is slightly smaller in the LS (long NP1, short NP2) condition than in the SL (short NP1, long NP2) condition. It would seem, then, that the overall NP1 attachment preference is modulated by the difference in length between the two attachment sites. Unfortunately, this effect is not statistically reliable, by participants $F(1,54)=2.55$, $p=.116$; by items $F(1,31)=2.2$, $p=.148$. When the results of the LS and SL conditions are analysed separately, we see a significant difference between NP1 and NP2 disambiguation in the SL condition [by participants $t(54)=2.59$, $p=.012$; by items $t(31)=4.13$, $p < .001$], whereas in the LS condition, the effect is marginal, by participants $t(54)=1.14$, $p=.258$, by items $t(31)=2.23$, $p=.034$. Admittedly, this is not exactly a strong result, but it suggests that it is slightly easier to accept an NP2-disambiguation when NP1 is long and NP2 short than when NP1 is shorter than NP2.

Reaction times. No effects were found in fragments 1 to 5. At the disambiguating auxiliary (fragment 6), the average reaction times are 1670 ms (s.e. 140) in the LS condition and 1566 ms (s.e. 102) in the SL condition. The difference is not significant, by participants $F(1,43)=1.92$; by items $F(1,31)=.231$. Across length manipulations, NP1 disambiguation yields an average reading time of 1708 ms (s.e. 144). The average reading time in the NP2 condition is 1554 ms (s.e. 123). This looks like a considerable difference, but it fails to reach significance, by participants $F(1,43)=1.12$, $p=.295$; by items $F(1,31)=3.43$, $p=.073$.

Figure 3 shows the reaction times for the two disambiguation conditions, as a function of the NP length manipulations. In the LS condition, NP1 disambiguation leads to a longer reading time than NP2 disambiguation: 1933 ms (s.e. 216) vs. 1460 ms (s.e. 145). In the SL condition the effect is reversed: 1484 ms (s.e. 101) for NP1 disambiguation, and 1649 ms (s.e. 138) for NP2 disambiguation. This interaction is statistically significant, by participants $F(1,43)=6.5$, $p=.014$; by items $F(1,31)=8.98$, $p=.005$.

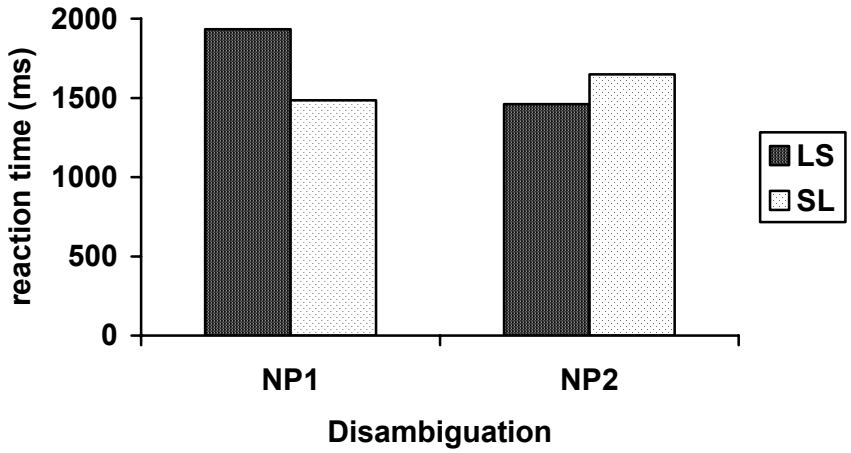


Figure 3. Experiment III: Mean reaction time of 'accept' responses at the disambiguating fragment 6 (auxiliary), as a function of the lengths of NP1 and NP2 (LS: long NP1, short NP2; SL: short NP1, long NP2) and disambiguation direction (NP1 or NP2).

Summarizing, the rejection rates indicate an overall preference for NP1 attachment, which agrees with previous findings on relative clause attachment in Dutch. The effect of the phonological length manipulation is weak: The apparent penalty on NP2 disambiguation is somewhat increased in cases where the two potential heads of the relative clause are short

and long, respectively. The direction of this effect is in agreement with the hypothesis that a long second NP will invite the insertion of a prosodic break, which blocks low attachment of the relative clause.

It is somewhat surprising that the disambiguation contrast did not produce a reliable difference in reaction time. There is a numerical trend, which, oddly enough, runs counter to the effect in the rejection rates. It would seem, however, that this trend is primarily due to the very long average reaction time in the case where NP1 disambiguation occurs in the LS condition. This long reaction time most likely also contributes considerably to the interaction of the phonological length manipulation and relative clause attachment disambiguation, which goes in the predicted direction: The LS condition is predicted to induce an NP2 attachment bias, and the data indicate that there is a strong penalty on forced NP1 attachment. The SL condition yields a mirror image effect, albeit less articulated.

5 General Discussion

The primary aim of this study was to test the implicit-prosody hypothesis, the idea that readers generate a phonological representation on the basis of the written input. This representation contains prosodic boundaries that impact on syntactic parsing, notably with respect to structural-ambiguity resolution. The present study used a somewhat notorious parsing problem as a window on this matter, viz. the resolution of relative-clause attachment ambiguity in a structure with two potential heads. In order to suppress as much as possible the effects of lexico-semantic and pragmatic factors, the experiments reported here made use of Jabberwocky prose. The crucial manipulation concerned the phonological lengths of the NPs that are accessible as head of the relative clause. The rationale behind this was that if an effect of this phonological manipulation was to be observed (in attachment preference), there would be no other way to account for the data than by reference to an implicit, self-generated phonological representation, matching the phonological structure that would be most naturally generated in the spoken realisation of the string under scrutiny.

The picture that emerges from the results is that in resolving relative-clause attachment ambiguity, in silent reading, the phonological lengths of the potential head nouns (Exp. I, III), as well as that of the relative clause itself (Exp. I vs. Exp. II), have an effect. The effects are not very strong; we see modulations of overall attachment preferences. The difference in outcome between experiments I and II agrees with the 'anti-gravity' effect reported earlier. The results of experiments I and III are compatible with the idea that the relative lengths of conjoined NPs affects prosodic phrasing, which in turn modulates RC-attachment bias. The results of this study contradict those reported by Colonna & Pynte (2001), which suggests that adding an adjective to an NP (as Colonna & Pynte did) impacts differently on processing than the purely phonological manipulation applied in this study.

Surprisingly, the overall attachment preference (NP1 vs. NP2) varies over tasks. Taking the two questionnaire studies together, we see quite a strong low-attachment preference (modulated by length of the relative clause), which is at odds with previous results on (real) Dutch. The continuous acceptability-judgment task, by contrast, yielded a bias toward NP1 attachment. I have no satisfactory explanation for this difference at present. Fodor (p.c.) suggests that processing Jabberwocky could be such a strain that subjects may resort to visual search in finding the attachment site for the relative clause. This would favor recency, i.e., NP2 attachment. Clearly, this strategy can only be applied when the stimuli are presented in full, not with a moving window presentation; hence the difference in bias between the off-line and on-line experiments. It would be interesting to see what will happen when the task of experiments I and II is combined with the presentation mode of experiment III.

The overall effects of the phonological manipulations are modest, and we may ask whether this is a result of the materials, the participants, or both. Conceivably, even in Jabberwocky prose, in which semantic and pragmatic influences on parsing decisions are believed to be virtually absent, other factors (beside syntax and prosody) can affect processing. It is also possible that generating implicit phonology is subject to much variation. Some readers may do it quite systematically, whereas others may never do it. Also, some stimuli may more often induce phonological encoding (in particular readers) than others. A replication of Experiment I using 'syntactic prose' items yielded a similar pattern of results, but much less articulated. One explanation for this difference is that syntactic prose, which contains existing content words (in nonsensical combinations) allows lexical access, and therefore does not engage the phonological processor as much as Jabberwocky — in which lexical access is ruled out. This is just to suggest that it might be worthwhile to further explore the conditions that stimulate or suppress phonological effects in visual language processing.

In concluding this contribution, I am considering whether Sieb Nooteboom might be inclined to believe my account. Probably not. After all, the inferential chain linking the data to the underlying processes is long and, possibly, seriously convoluted. Nonetheless, I have the firm conviction that the strife of the carmactrolophy that was plimpered in this lasteract has a snuzzling porrifritch.

References

Brysbaert, M., & Mitchell, D. G. (1996). Modifier attachment in sentence parsing: Evidence from Dutch. *Quarterly Journal of Experimental Psychology, 49(A)*, 644-695.

Colonna, S., & Pynte, J. (2001). *Relative clause attachment in French: The role of Fodor's "same-size-sister" constraint.* Paper presented at the Workshop on Prosody in Processing, Utrecht.

Cuetos, F., & Mitchell, D. G. (1988). Cross-linguistic differences in parsing: Restrictions on the use of the late closure strategy. *Cognition, 30*, 73-105.

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech, 40*, 141-201.

Ehrlich, K., Fernández, E., Fodor, J. D., Stenshoel, E., & Vinereanu, M. (1999). *Low attachment of relative clauses: New data from Swedish, Norwegian and Romanian.* Paper presented at the 12th Annual CUNY conference on Human Sentence Processing, New York.

Fodor, J. D. (1998). Learning to parse? *Journal of Psycholinguistic Research, 27*, 285-317.

Fodor, J. D. (2002). Prosodic disambiguation in silent reading. In M. Hirotani (Ed.), *Proceedings of NELS 32* (113-132). Amherst, MA: GLSA.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition, 6*, 291-326.

Hahne, A., & Jescheniak, J. D. (2001). What's left if the Jabberwock gets the semantics? An ERP investigation into semantic and syntactic processes during auditory sentence comprehension. *Cognitive Brain Research, 11*, 199-212.

Kimball, J. (1973). Seven principles of surface structure parsing in a natural language. *Cognition, 2*, 15-47.

Nagel, N. H., Shapiro, L. P., Tuller, B., & Nawy, R. (1996). Prosodic influences on the resolution of temporary ambiguity during on-line sentence processing. *Journal of Psycholinguistic Research, 25*, 319-343.

Pynte, J., & Prieur, B. (1996). Prosodic breaks and attachment decisions in sentence parsing. *Language and Cognitive Processes, 11*, 165-191.

Schafer, A., Carter, J., Clifton, C., & Frazier, L. (1996). Focus in relative clause construal. *Language and Cognitive Processes, 11*, 135-163.

Steinhauer, K., Alter, K., & Friederici, A. D. (1999). Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience, 2*, 191-196.

Watt, S. M., & Murray, W. S. (1996). Prosodic form and parsing commitments. *Journal of Psycholinguistic Research, 25*, 291-318.

Bibliography of Sieb G. Nooteboom

Hugo Quené

Utrecht University

Nooteboom, S. G. (in press). Listening to one-self: Monitoring speech production. In R. J. Hartsuiker, Y. Bastiaanse, A. Postma & F. N. K. Wijnen (Eds.), *Phonological encoding and monitoring in normal and pathological speech*. Hove: Psychology Press.

Nooteboom, S.G. (2004) *Waar komen de letters van het alfabet vandaan?* Afscheidscollege, 23 April. Utrecht: Universiteit Utrecht.

Janse, E., Nooteboom, S. G., & Quené, H. (2003). Word-level intelligibility of time-compressed speech: prosodic and segmental factors. *Speech Communication, 41*(2-3), 287-301.

Nooteboom, S. G. (2003a). Self-monitoring is the main cause of lexical bias in phonological speech errors. In R. Eklund (Ed.), *Proceedings of the Workshop on Disfluency in Spontaneous Speech (DiSS'03); Göteborg, 5-8 September 2003* (pp. 25-28). Gothenburg Papers in Theoretical Linguistics; 89.

---. (2003b). Lexical Bias in Phonological Speech Errors: Phoneme-to-Word Feedback or Output Editing? In M. J. Solé, D. Recasens & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences; Barcelona, 3-9 Augustus 2003* (pp. 2249-2252).

Nooteboom, S. G., Weerman, F., & Wijnen, F. N. K. (2002). Minimising or maximising storage? An introduction. In S. G. Nooteboom, F. Weerman & F. N. K. Wijnen (Eds.), *Storage and Computation in the Language Faculty* (pp. 1-19). Dordrecht: Kluwer.

Nooteboom, S. G., Weerman, F. P., & Wijnen, F. N. K. (Eds.) (2002). *Storage and Computation in the Language Faculty*. Dordrecht: Kluwer. Studies in theoretical psycholinguistics; 30.

van Rossum, M. A., de Krom, G., Nooteboom, S. G., & Quené, H. (2002). "Pitch" accent in alaryngeal speech. *Journal of Speech Language and Hearing Research, 45*(6), 1106-1118.

Nooteboom, S. G., & Vermeulen, K. (2000). Heads and tails of Dutch spoken words: An experiment on the relative contribution of word beginnings and endings to word recognition. In T. F. Shannon & J. P. Snapper (Eds.), *Proceedings of the Berkeley Conference on Dutch Linguistics 1997: The Dutch Language at the Milennium* (pp. 1-19). Lanham, MD: University Press of America. Publications of the American Association for Netherlandic Studies; 12. [Preview (1998) published as UiL OTS Working Paper; 98002-FON].

Nooteboom, S. G. (1999). Sloppiness in uttering stock phrases. In J.J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A.C. Bailey (Eds.), *Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, August 1-7, 1999* (Vol. 1, pp. 683-687).

---. (1998). Woorden in de wachtkamer. In E. Blom & A.-L. Jansen (Eds.), *Tijdperk Taal!: 10 jaar linguïstiek in Utrecht* (pp. 92-103). Den Haag: Holland Academic Graphics.

Nooteboom, S. G., & Van Dijk, M. (1998). Heads and tails in word perception: Evidence for 'early-to-late' processing in listening and reading. In *Proceedings of the International Congress of Spoken Language Processing (ICSLP), Sydney, 30 November - 4 December 1998* (Nr. 117).

Nooteboom, S. G. (1997). The prosody of speech: Melody and rhythm. In W. J. Hardcastle & J. Laver (Eds.), *The Handbook of Phonetic Sciences* (pp. 640-673). Oxford: Blackwell.

---. (1996a). Antonie Cohen: Obituary. *Phonetica, 53*(4), 230-232.

---. (1996b). Text and prosody: An overview. In J. P. H. Van Santen, R. W. Sproat, J. P. Olive & J. Hirschberg (Eds.), *Progress in Speech Synthesis* (pp. 431-434). New York: Springer.

Nooteboom, S. G., Cohen, A., & Eggen, J. H. (1996). Synthesizing personal characteristics of speech. In Z. Palkovà (Ed.), *Charisteria viro doctissimo Premysl Janota* (pp. 191-206). Prague: Charles University.

Nooteboom, S. G. (1995a). Limited lookahead in speech production. In F. Bell-Berti & L. R. Raphael (Eds.), *Producing speech: Contemporary issues: For Katherine Safford Harris* (pp. 3-18). New York: AIP

Press. AIP Series in Modern Acoustics and Signal Processing. [ISBN 1563962861. Preview (1994) in OTS Working Papers; OTS-WP-FON-94-003].

---. (1995b). How far do we look ahead while speaking? In K. Ellenius & P. Branderud (Eds.), *Proceedings of the XIIIth International Congress of Phonetic Sciences; Stockholm, August 1995* (Vol. 4, pp. 578-581).

---. (1995c). Hot topics in the field of speech perception. In G. Bloothooft, V. Hazan, D. Huber & J. Llisterri (Eds.), *European Studies in Phonetics and Speech Communication* (pp. 71-76). Utrecht: OTS Publications.

Nooteboom, S. G., & Cohen, A. (1994). *Spreken en Verstaan: Een nieuwe inleiding tot de experimentele fonetiek* (4th ed.). Assen: Van Gorcum.

Nooteboom, S. G., & Eefting, W. (1994). Evidence for the adaptive nature of speech on the phrase level and below. *Phonetica, 51*(1-3), 92-98.

Cohen, A., & Nooteboom, S. G. (1993). A five year research program "Analysis and Synthesis of Speech". In V. J. Van Heuven & L. C. W. Pols (Eds.), *Analysis and synthesis of speech: Strategic research towards high-quality text-to-speech generation* (pp. 1-10). Berlin: Mouton de Gruyter. Speech Research; 11.

Eefting, W. Z. F., & Nooteboom, S. G. (1993). Accentuation, information value and word duration : effects on speech production, naturalness and sentence processing. In V. J. Van Heuven & L. C. W. Pols (Eds.), *Analysis and synthesis of speech: Strategic research towards high-quality text-to-speech generation* (pp. 225-240). Berlin: Mouton de Gruyter. Speech Research; 11.

Eggen, B., & Nooteboom, S. G. (1993). Speech quality and speaker characteristics. In V. J. Van Heuven & L. C. W. Pols (Eds.), *Analysis and synthesis of speech: Strategic research towards high-quality text-to-speech generation* (pp. 279-288). Berlin: Mouton de Gruyter. Speech Research; 11.

Nooteboom, S. G. (1993). Computers, onderzoek en onderwijs in een letterenfaculteit: De fonetiek als onderdeel van de taalwetenschap. *Universiteit & Hogeschool, Tijdschrift voor wetenschappelijk onderwijs, 39*(6), 251-262.

Eggen, B. J. H., Nooteboom, S. G., & Houtsma, A. H. M. (1992). Contributions of voice-source and vocal-characteristics to speaker identity. *Journal of the Acoustical Society of America, 92*(4), 2301.

Nooteboom, S. G. (1992). Comment on the session "Contextual effects in vowel perception" of the ATR Workshop on Speech Perception and Production, Kyoto, December 15-16, 1990. In Y. Tohkura, E. Vatikiotis-Bateson & Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure* (pp. 83-85). Tokyo: Ohmsha.

Nooteboom, S. G., & Eefting, W. (1992). To what extent is speech production controlled by speech perception? Some questions and some experimental evidence. In Y. Tohkura, E. Vatikiotis-Bateson & Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure* (pp. 439-449). Tokyo: Ohmsha.

Eefting, W., & Nooteboom, S. G. (1991). The effect of accentedness and information value on word durations: a production and a perception study. In *Proceedings of the XIIth International Congress of Phonetic Sciences; Aix-en-Provence, August 19-24.* (Vol. 3, pp. 302-305).

Nooteboom, S. G. (1991a). Perceptual goals of speech production. In *Proceedings of the XIIth International Congress of Phonetic Sciences; Aix-en-Provence, August 19-24.* (Vol. 1, pp. 107-110).

---. (1991b). Words are produced in order to be perceived: the listener in the speaker's mind. In O. Engstrand & C. Kylander (Eds.), *Papers from the symposium "Current Phonetic Research Paradigms: Implications for Speech Motor Control", held in Stockholm, August 13-16, 1991.* (pp. 149-152). [Perilus XIV].

---. (1991c). Some observations on the temporal organization and rhythm of speech. In *Proceedings of the XIIth International Congress of Phonetic Sciences; Aix-en-Provence, August 19-24* (pp. 228-237).

Nooteboom, S. G., & Van Bezooijen, R. (1991). Five years of coordinated research on text-to-speech conversion for Dutch: An overview. In *Proceedings of the XIIth International Congress of Phonetic Sciences; Aix-en-Provence, August 19-24.* (Vol. 3, pp. 470-473).

Nooteboom, S. G., Scharpff, P., & Van Heuven, V. J. (1990). Effects of several pausing strategies on the recognizability of words in synthetic speech. In *Proceedings of the International Conference on Spoken Language Processing; Kobe* (pp. 385-387).

Nooteboom, S. G. (1988a). *Over het uitspreken van een rede*. Inaugural address. Utrecht: Utrecht University.

---. (1988b). Herkenning van gesproken woorden. In M. P. R. Van den Broecke (Ed.), *Ter Sprake: Spraak als betekenisvol geluid in 36 thematische hoofdstukken* (pp. 151-158). Dordrecht: Foris.

---. (1988c). Speech coding, synthesis and voice quality. In B. A. G. Elsendoorn & H. Bouma (Eds.), *Working Models of Human Perception* (pp. 173-181). London: Academic Press.

Nooteboom, S. G., & Cohen, A. (1988). *Spreken en Verstaan: Een nieuwe inleiding tot de experimentele fonetiek* (3rd ed.). Assen: Van Gorcum.

Nooteboom, S. G., & Van der Vlugt, M. J. (1988). A search for a word-beginning superiority effect. *Journal of the Acoustical Society of America, 84*(6), 2018-2032.

Nooteboom, S. G. (1987). Developments. *IPO Annual Progress Report, 22*, 13-14.

Nooteboom, S. G., & Kruyt, J. G. (1987). Accents, focus distribution, and the perceived distribution of given and new information. *Journal of the Acoustical Society of America, 82*(5), 1512-1524.

Terken, J. M. B., & Nooteboom, S. G. (1987). Opposite effects of accentuation and deaccentuation on verification latencies for given and new information. *Language and Cognitive Processes, 2*(3-4), 145-163.

Nooteboom, S. G. (1986). Developments. *IPO Annual Progress Report, 21*, 11-13.

---. (1985a). A functional view of prosodic timing in speech. In J. A. Michon & J. L. Jackson (Eds.), *Time, Mind and Behavior* (pp. 242-252). Berlin: Springer.

---. (1985b). Developments. *IPO Annual Progress Report, 20*, 11-13.

Nooteboom, S. G., & Van der Vlugt, M. J. (1985). Prefixes versus suffixes: a search for a word-beginning superiority effect in word recogntion from degraded speech. *IPO Annual Progress Report, 20*, 43-52.

Kruyt, J. G., & Nooteboom, S. G. (1984). Acceptability of accenting 'new' and 'given' in Dutch. In M. P. R. Van den Broecke & A. Cohen (Eds.), *Proceedings of the Tenth International Congress of Phonetic Sciences;Utrecht, 1-6 August* (pp. 549-553). Dordrecht: Foris. Netherlands Phonetic Archives; 2B.

Nooteboom, S. G. (1984). Developments. *IPO Annual Progress Report, 19*, 11-13.

Nooteboom, S. G., & Cohen, A. (1984). *Spreken en Verstaan: Een nieuwe inleiding tot de experimentele fonetiek* (2nd, revised ed.). Assen: Van Gorcum.

Nooteboom, S. G., & Doodeman, G. J. N. (1984). Speech quality and the gating paradigm. In M. P. R. Van den Broecke & A. Cohen (Eds.), *Proceedings of the Tenth International Congress of Phonetic Sciences;Utrecht, 1-6 August* (pp. 481-485). Dordrecht: Foris. Netherlands Phonetic Archives; 2B.

Nooteboom, S. G. (1983a). Is speech production controlled by speech perception? In M. P. R. Van den Broecke, V. J. Van Heuven & W. Zonneveld (Eds.), *Sound structures: Studies for Antonie Cohen* (pp. 183-194). Dordrecht: Foris.

---. (1983b). Mens en machine praten nog al eens langs elkaar heen. *TNO Project, maandblad voor toegepaste wetenschappen, 11*(12), 413-416.

---. (1983c). Developments. *IPO Annual Progress Report, 18*, 11-13.

---. (1983d). The temporal organisation of speech and the process of spoken-word recognition. *IPO Annual Progress Report, 18*, 32-36.

Bouma, H., & Nooteboom, S. G. (1982). *Ben L. Cardozo: Publicaties 1962-1982*. Eindhoven: Instituut voor Perceptie Onderzoek (IPO). (Selectie van publicaties aangeboden aan B.L. Cardozo t.g.v. zijn pensionering en afscheid van het IPO op 19 november 1982).

Bouma, H., Nooteboom, S. G., & Willems, L. F. (Eds.) (1982). *25 jaar IPO: Opstellen ter gelegenheid van het 25-jarig bestaan van het Instituut voor Percpetie Onderzoek 1957-1982*. Eindhoven: IPO.

Brokx, J. P. L., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics, 10*, 23-36.

Nooteboom, S. G. (1982a). *Fonetiek op het grensvlak tussen geluid en betekenis*. Inaugural address, Rijksuniversiteit Leiden. Leiden: Brill.

---. (1982b). Developments. *IPO Annual Progress Report, 17*, 39-40.

Nooteboom, S. G., & Doodeman, G. J. N. (1982). Speech quality and word recognition from fragments of spoken words. *IPO Annual Progress Report, 17*, 46-50.

Nooteboom, S. G., & Terken, J. M. B. (1982). What makes speakers omit pitch accent?: An experiment. *Phonetica, 39*(4-5), 317-336.

't Hart, J., Nooteboom, S. G., Vogten, L. L. M., & Willems, L. F. (1982). Manipulaties met spraakgeluid. *Philips Technisch Tijdschrift, 40*(4), 108-119.

Nooteboom, S. G. (1981a). Speech rate and segmental perception, or, The role of words in phoneme identification. In T. Myers, J. Laver & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 143-150). Amsterdam: North Holland. Advances in Psychology; 7.

---. (1981b). Lexical retrieval from fragments of spoken words : beginnings vs. endings. *Journal of Phonetics, 9*(4), 407-424.

---. (1981c). Developments. *IPO Annual Progress Report, 16*, 13-14.

Nooteboom, S. G., Kruyt, J. G., & Terken, J. M. B. (1981). What speakers and listeners do with pitch accents: Some explorations. In T. Fretheim (Ed.), *Nordic Prosody II: Papers from a symposium* (pp. 9-32). Trondheim: Tapir.

Nooteboom, S. G. (1980a). Speaking and unspeaking: detection and correction of phonological and lexical errors of speech. In V. A. Fromkin (Ed.), *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen, and Hand* (pp. 87-96). New York: Academic Press.

---. (1980b). Developments. *IPO Annual Progress Report, 15*, 11.

Nooteboom, S. G., & Doodeman, G. J. N. (1980). Production and perception of vowel length in spoken sentences. *Journal of the Acoustical Society of America, 67*(1), 276-287.

Nooteboom, S. G., & Truin, P. J. M. (1980). The recognition of fragments of spoken words by native and non-native listeners. *IPO Annual Progress Report, 15*, 42-47.

Brokx, J. P. L., Nooteboom, S. G., & Cohen, A. (1979). Pitch differences and the intelligibility of speech masked by speech. *IPO Annual Progress Report, 14*, 55-60.

Duifhuis, H., & Nooteboom, S. G. (1979). Developments. *IPO Annual Progress Report, 14*, 11.

Nooteboom, S. G. (1979a). The time course of speech perception. In W. J. Barry & K. Kohler (Eds.), *"Time" in the Production and Perception of Speech.* (pp. 113-151). Kiel: Phonetics Department of Kiel University. Arbeitsberichte; 12.

---. (1979b). Perceptual adjustment to speech rate: a case of backward perceptual normalization. In *Anniversaries in Phonetics: Studia Gratulatoria dedicated to Henrik Mol* (pp. 255-269). Amsterdam: Institute of Phonetic Sciences, University of Amsterdam.

---. (1979c). More attention for words in speech communication research. In B. Lindblom & S. Öhman (Eds.), *Frontiers of Speech Communication Research: Festschrift for Gunnar Fant* (pp. 203-211). London: Academic Press.

---. (1979d). Complex control of simple decisions in the perception of vowel length. In *Proceedings of the IXth International Congress of Phonetic Sciences; Copenhagen, 6-11 August* (Vol. II, pp. 298-304).

Duifhuis, H., & Nooteboom, S. G. (1978). Developments. *IPO Annual Progress Report, 13*, 11.

Nooteboom, S. G., Brokx, J. P. L., & De Rooij, J. J. (1978). Contributions of prosody to speech perception. In W. J. M. Levelt & G. B. Flores d'Arcais (Eds.), *Studies in the perception of language* (pp. 75-107). Chichester: Wiley. (Also appeared (1976) in IPO Annual Progress Report, 11, 34-54).

Muller, H. F., Nooteboom, S. G., & Willems, L. F. (1977). An experimental system for man-machine communication by means of speech. *IPO Annual Progress Report, 12*, 41-46.

Nooteboom, S. G., & Cohen, A. (1976). *Spreken en Verstaan: Een inleiding tot de experimentele fonetiek* (1st ed.). Assen: Van Gorcum.

Cohen, A., & Nooteboom, S. G. (Eds.) (1975). *Structure and Process in Speech Perception: Proceedings of the symposium on dynamic aspects of speech perception; Eindhoven, August 4-6*. Berlin: Springer. Communication and Cybernetics; 11. [See also: A. Cohen & S. G. Nooteboom (1975) A symposium on dynamic aspects of speech perception; IPO, August 1975. IPO Annual Progress Report, 10, 45-47].

Nooteboom, S. G. (1975a). Anticipation in speech production and its implications for perception. In A. Cohen & S. G. Nooteboom (Eds.), *Structure and Process in Speech Perception: Proceedings of the symposium on dynamic aspects of speech perception; Eindhoven, August 4-6* (pp. 124-145). Berlin: Springer.

---. (1975b). Contextual variation and the perception of phonetic vowel length. In *Proceedings of the second seminar on Speech Communication; Stockholm, August 1-3, 1974* (pp. 149-153). Stockholm: Almqvist & Wiksell.

---. (1975c). On the internal auditory representation of syllable nucleus durations. In G. Fant & M. A. A. Tatham (Eds.), *Auditory Analysis and Perception of Speech* (pp. 413-430). London: Academic Press.

---. (1975d). Chairman's Review, Session V. In G. Fant & M. A. A. Tatham (Eds.), *Auditory Analysis and Perception of Speech* (pp. 383-386). London: Academic Press.

Nooteboom, S. G., Brokx, J. P. L., Doodeman, G. J. N., De Jong, T. A., 't Hart, J., Van Katwijk, A. F. V., et al. (1975). Research on speech perception in the IPO 1975. *IPO Annual Progress Report, 10*, 25-26.

Nooteboom, S. G. (1974a). Book review: P. Ladefoged, Preliminaries to Linguistic Phonetics. *Lingua, 34*, 253-264.

---. (1974b). Some context effects in phonemic categorization of vowel duration. *IPO Annual Progress Report, 9*, 47-55.

Nooteboom, S. G., Eggermont, J. P. M., 't Hart, J., Van Katwijk, A. F. V., & Slis, I. H. (1974). Time in speech perception. *IPO Annual Progress Report, 9*, 39-46.

Nooteboom, S. G. (1973a). Experimentele bijdragen aan de fonologie. *Forum der Letteren, 15*, 73-99.

---. (1973b). Spraaksynthese in het fonetisch onderzoek. *Informatie, 15*, 136.

---. (1973c). The perceptual reality of some prosodic durations. *Journal of Phonetics, 1*, 25-45.

---. (1973d). The tongue slips into patterns. In V. A. Fromkin (Ed.), *Speech Errors as Linguistic Evidence* (pp. 144-156). The Hague: Mouton. [Also published (1969) in A. Sciarone, et al. (Eds.) *Nomen: Leyden Studies in Linguistics and Phonetics* (pp. 114-132). The Hague: Mouton.]

Nooteboom, S. G., Slis, I. H., & Willems, L. F. (1973). Speech synthesis by rule; why, what, and how? *IPO Annual Progress Report, 9*, 3-13.

Nooteboom, S. G. (1972a). *Production and perception of vowel duration: A study of durational properties of vowels in Dutch.* Unpublished PhD thesis, Rijksuniversiteit Utrecht. (Also published as: Philips Research Reports; 5).

---. (1972b). Temporal patterns in Dutch. In *Proceedings of the VIIth International Congress of Phonetic Sciences; Montreal, 22-28 August, 1971* (pp. 984-989). The Hague: Mouton.

---. (1972c). Some timing factors in the production and perception of vowels. *Occasional Papers, University of Essex, 13*, 49-76.

---. (1972d). De besturing van spraak: invariante motorcommando's of vaste doelposities? In *Taalwetenschap in Nederland 1971* (pp. 89-92). Amsterdam: Universiteit van Amsterdam.

---. (1972e). The interaction of some intra-syllable and extra-syllable factors acting on syllable nucleus durations. *IPO Annual Progress Report, 7*, 30-39.

---. (1972f). A brief survey of some investigations into the temporal organisation of speech. *IPO Annual Progress Report, 7*, 17-29.

Nooteboom, S. G., & Slis, I. H. (1972). The phonetic feature of vowel length in Dutch. *Language and Speech, 15*(4), 301-316.

Nooteboom, S. G. (1971a). Over de lengte van korte klinkers, lange klinkers en tweeklanken in het Nederlands. *Nieuwe Taalgids, 64*, 396-402. (Available: http://www.dbnl.org/tekst/noot004leng01/index.htm).

---. (1971b). Book review: Georg Heike, Sprachliche Kommunikation und linguistische Analyse. *Lingua, 27*, 282-287.

---. (1971c). Enkele opmerkingen over de relatie tussen generatieve fonologie en eperimentele fonetiek. *Studia Neerlandica, 6*, 169-178.

---. (1971d). Whistling rhythmic patterns of speech. *IPO Annual Progress Report, 6*, 37-40.

---. (1970a). Measurements on the movement of the lower jaw in speech. *IPO Annual Progress Report, 5*, 59-64.

---. (1970b). The target theory of speech production. *IPO Annual Progress Report, 5*, 51-55.

Nooteboom, S. G., & Slis, I. H. (1970). A note on the degree of opening and the duration of vowels in normal and 'pipe' speech. *IPO Annual Progress Report, 5*, 55-58.

Nooteboom, S. G. (1969a). Onderzoekingen over het normale lezen: linguïstische aspecten. *Nederlands Tijdschrift voor de Geneeskunde, 13*, 1628.

---. (1969b). Hardop lezen als vorm van continu taalgebruik. *Forum der Letteren, 10*, 95-107.

Nooteboom, S. G., & Eggermont, J. P. M. (1969). Syllable-timing or stress-timing in Dutch? *IPO Annual Progress Report, 4*, 41-43.

Nooteboom, S. G., & Slis, I. H. (1969). A note on rate of speech. *IPO Annual Progress Report, 4*, 58-60.

Slis, I. H., & Nooteboom, S. G. (1969). Articulatory timing in Dutch stressed words. *IPO Annual Progress Report, 4*, 52-57.

Bouma, H., & Nooteboom, S. G. (1968). On reading letters and random letter combinations from a relatively long distance. *IPO Annual Progress Report, 3*, 89-94.

Nooteboom, S. G. (1968). Perceptual confusions among Dutch vowels presented in noise. *IPO Annual Progress Report, 3*, 68-71.

Nooteboom, S. G., & Bouma, H. (1968). On reading nonsense syllables, whole words and coherent text from a relatively long distance. *IPO Annual Progress Report, 3*, 47-54.

Nooteboom, S. G. (1967). Some regularities in phonemic speech errors. *IPO Annual Progress Report, 2*, 65-70.

Nooteboom, S. G., & Zonneveld, F. W. (1967). On differences between prevocalic and postvocalic consonants. *IPO Annual Progress Report, 2*, 92-93.

Promoti and promovendi (Ph.D. graduates and students)

E. Janse, *Production and perception of fast speech* (2003, Utrecht).

P.A.M. Gerrits, *The categorisation of speech sounds by adults and children* (2001, Utrecht).

S.M.M. te Riele, *Early context effects in spoken-word perception* (1999, Utrecht).

A. Sanderman, *Prosodic phrasing: Production, perception, acceptability and comprehension* (1996, Eindhoven, together with R. Collier).

M.J. Sanders, *Intonation contour choice in English* (1996, Utrecht, together with R. Collier, Eindhoven).

H.H. Rump, *Prominence of pitch-accented syllables* (1996, Eindhoven, together with R. Collier).

W. A. van Donselaar, *Effects of accentuation and given/new information on word processing* (1995, Utrecht).

E. Blaauw, *On the perceptual classification of spontaneous and read speech* (1995, Utrecht).

P.J.A.M. Scharpff, *Het effect van spreekpauzes op de herkenning van woorden in voorgelezen zinnen*, (1994, Leiden, together with J.G. Kooij).

L. Menert, *Experiments on voice assimilation in Dutch* (1994, Utrecht, together with W. Zonneveld).

G. de Krom, *Acoustic correlates of breathiness and roughness* (1994, Utrecht).

J. Caspers, *Pitch movements under time pressure* (1994, Leiden, together with J.G. Kooij).

J.H. Eggen, *On the quality of synthetic speech* (1992, Eindhoven, together with H. Bouma).

I. Petric, *Here is the news. Predicting listening performance for news texts* (1992, Utrecht).

A.J. van Hessen, *Discrimination of familiar and unfamiliar speech sounds* (1992, Utrecht).

W.Z.F. Eefting, *Timing in talking* (1991, Utrecht).

L.M.H. Adriaens, *Ein Modell deutscher Intonation* (1991, Eindhoven, together with H. Bouma).

R.J.H. Deliège, *The "Tiepstem", an experimental Dutch keyboard-to-speech system for the speech impaired* (1989, Eindhoven, together with H. Bouma).

M.L. van Dijk-Kappers, *Temporal decomposition of speech and its relation to phonetic information* (1989, Eindhoven, together with H. Bouma).

H.C. van Leeuwen, *Toolip: a development system for linguistic rules* (1989, Eindhoven, together with H. Bouma).

S.J. Langeweg, *The stress system of Dutch* (1988, Leiden, together with J.G. Kooij).

M.C.H. Dupuis, *Perceptual effects of phonetic and phonological accommodation* (1988, Leiden).

M.J. van der Vlugt, *Spraakgeluid en woordherkenning* (1987, Eindhoven).

J.L.G. Baart, *Focus, syntax and accent placement* (1987, Leiden).

J.M.B. Terken, *Use and function of accentuation* (1985, Leiden).

J.G. Kruyt, *Accents from speakers to listeners* (1985, Leiden).

M.T.M.Scheffers, *Sifting vowels* (1983, Groningen, together with H. Duifhuis).

W.F.L. Heeren, *The acquisition of new phoneme contrasts by children and adults*.

A.W. Hoekstra, *Intonation and interpretation.*

M.A. van Rossum, *Prosody in alaryngeal speech*.

Universiteit Utrecht

Aankondiging & uitnodiging

Feestelijk programma ter gelegenheid van het afscheid van
Sieb G. Nooteboom als hoogleraar Fonetiek aan de Universiteit Utrecht,
op vrijdag 23 april 2004, te Utrecht (*wijzigingen voorbehouden*)

ochtend: Drift 21, collegezaal 0.32

9:30	ontvangst met koffie	
10:00	opening	
	voordrachten:	
	dr. Johanneke Caspers	*A matter of secondary importance: intonation of backchannels*
	dr. Jacques Terken	*Speech in the interface*
	dr. Esther Janse	*Auditory word perception in healthy listeners and in aphasic patients*
	dr. Arjan van Hessen	*Pandora's box*
11:30	pauze, koffie	
12:00	voordracht door prof. J.J. Ohala (University of California, Berkeley, U.S.A.) *Phonetics then, and then, and now*	
13:00	einde ochtendprogramma	

middag: Academiegebouw, Domplein, Aula

16:00	afscheidscollege van prof. dr. S.G. Nooteboom *Waar komen de letters van het alfabet vandaan?*
17:15	receptie (Senaatszaal)

Utrecht Institute of Linguistics OTS

Tabula Gratulatoria

Petra van Alphen
S. Avrutin
Joan Baart
Jocelyn Ballantyne
Florien J. van Beinum
Hans Bertens
Renée van Bezooijen
Gerrit Bloothooft
Paul Boersma
Geert Booij
Herman Bouma
Tina Cambier-Langeveld
Johanneke Caspers
Aoju Chen
René Collier
Peter Coopmans
Norbert Corver
Onno Crasborn
Anne Cutler
Jo Daan
Gary Dell
Arthur Dirksen
Mirjam Ernestus
Martin Everaert
Simo Goddijn
Carlos Gussenhoven
Willemijn Heeren
Caroline Henton
Vincen van Heuven
Arjan van Hessen
Dik Hermes
Heleen Hoekstra
Esther Janse
Joop Kerkhoff
Guus de Krom
Truus Kruyt
Cecile Kuijpers
Paul Kuyper
Bob Ladd
Pim Levelt
Björn Lindblom
Jean-Pierre Martens
James McQueen
Ludmila Menert
Holger Mitterer
Pieter Muysken
Anneke Neijt
Cecilia Odé
John Ohala
Peter Pabon
Jos J.A. Pacilly
Louis Pols
Mieke Postma - van Wijck
Hugo Quené
Pieter van Reenen
Eric Reuland
Toni Rietveld
Maya van Rossum
Margreet Sanders
Niels O. Schiller
Henk Schultink
Iman Slis
Anneke Slis
Hans Smits
Helmer Strik
Janienke Sturm
Jacques Terken
Hans Van de Velde
Ger Vellekoop
Jos F.M. Vermeulen
Jill Warker
Lei Willems
Ton van der Wouden
Frank Wijnen
Ellen van Zanten
Meertens Instituut