Oxford University Working Papers in Linguistics, Philology & Phonetics

Volume 8
2003

Edited by
Esther Grabe
David G.S. Wright

Contents

Editorial note

The Oxford University Working Papers in Linguistics, Philology and Phonetics is an annual publication presenting some of the research carried out in the field of Linguistics, Philology and Phonetics at the University of Oxford, sometimes in collaboration with researchers from other universities. The current volume contains seven papers from the Phonetics Laboratory and four papers from the Computational Linguistics Group.

Individual authors may be contacted at the addresses listed on page v. For general enquiries about the Oxford University Working Papers, please contact enquiries@ling-phil.ox.ac.uk.

Esther Grabe David G.S. Wright

Contributors' addresses

Photini Coutsougera
Phonetics Laboratory
41 Wellington Square
Oxford University
Oxford
OX1 2JF
photini.coutsougera@phon.ox.ac.uk

Esther Grabe
Phonetics Laboratory
41 Wellington Square
Oxford University
Oxford
OX1 2JF
esther.grabe@phon.ox.ac.uk

John Hajek
French and Italian Studies
Arts Centre Building
The University of Melbourne
Melbourne
3010 Australia
j.hajek@unimelb.edu.au

Maciej Karpinski
Institute of Linguistics
Adam Mickiewicz University
ul. Miedzychodzka 5
Poznan
60-371 Poland
maciejk@amu.edu.pl

Elinor Keane
Christ Church
Oxford
OX1 1DP
elinor.keane@phon.ox.ac.uk

Keith Owen
Worcester College
Oxford
OX1 2HB
keith.owen@worcester.oxford.ac.uk

Hannele Nicholson
Theoretical and Applied Linguistics
The University of Edinburgh
Adam Ferguson Building
40 George Square
Edinburgh
EH8 9LL
hannele@ling.ed.ac.uk

Brechtje Post
Phonetics Laboratory
41 Wellington Square
Oxford University
Oxford
OX1 2JF
brechtje.post@phon.ox.ac.uk

Andreas Hilmo Teig
The University of Edinburgh
New College
Mound Place
Edinburgh
EH1 2LX
teig@nvg.org

Burton Rosner
Phonetics Laboratory
41 Wellington Square
Oxford University
Oxford
OX1 2JF
burton.rosner@phon.ox.ac.uk

Ian Watson
Phonetics Laboratory
41 Wellington Square
Oxford University
Oxford
OX1 2JF
ian.watson@phon.ox.ac.uk

Simon Dobnik
Centre for Linguistics and Philology
Walton Street
Oxford
OX1 2HG
simon.dobnik@clg.ox.ac.uk

Dimitar Kazakov
Department of Computer Science
University of York
Heslington
York
YO10 5DD
kazakov@cs.york.ac.uk

Maria Liakata
Centre for Linguistics and Philology
Walton Street
Oxford
OX1 2HG
maria.liakata@clg.ox.ac.uk

Stephen Pulman
Centre for Linguistics and Philology
Walton Street
Oxford
OX1 2HG
stephen.pulman@clg.ox.ac.uk

Jana Sukkarieh
Centre for Linguistics and Philology
Walton Street
Oxford
OX1 2HG
jana.sukkarieh@clg.ox.ac.uk

David Wright
Centre for Linguistics and Philology
Walton Street
Oxford
OX1 2HG
david.wright@clg.ox.ac.uk

The Cypriot Greek Syllable

PHOTINI COUTSOUGERA

1. Introduction

The present paper studies the syllable of Cypriot Greek (henceforth CG), a variety of Greek spoken on the island of Cyprus.

It will be argued that the CG syllable prefers to have an onset rather than be onsetless while, on the other hand, it strongly prefers to be codaless rather than have a coda. The word-internal CG coda position can be filled only by a sonorant segment.

More specifically, it will be argued that, contra Malikouti-Drachman (1999b, 2000), who has proposed that the syllabification of CG is performed by virtue of the Sonority Sequencing Generalisation (henceforth SSG), sonority is not crucial for a correct account of the CG phonology.

The model of syllable that will be employed is an onset-rhyme model, which assumes hierarchical internal syllabic structure as set up by Fudge (1969, 1987).

2. The phonotaxis of Cypriot Greek

In CG the inventory of bi-consonantal clusters of obstruents is severely limited, unlike the inventory of sonorant sequences or sonorant obstruent ones, which is considerably richer.

Newton (1970: 30) gives us the following figures regarding permissible obstruent bi-consonantal clusters in CG:

13 fricative+stop clusters, 6 stop+fricative ones, 2 stop+stop ones[1] and 1 fricative+fricative cluster (1 voiceless and 1 voiced fricative+fricative cluster). See Newton (1970: 31-32) for a plethora of examples. Furthermore, 'mixed' clusters consisting of fricative+stop sequences are the most frequently occurring obstruent clusters in the dialect.

[1] Newton (1970) includes [pt^h] in stop+stop clusters, which is highly marginal.

2.1 Obstruent clusters in Cypriot Greek

The following clusters are ***surface clusters***, which may be derived from two (or more) distinct underlying representations. They can be found both ***word-initially*** and ***word-internally*** (with the exception of the rather rare [xc] and [sc] which do not appear word-initially). Homorganic obstruent clusters are disallowed (*θt, *xk etc.). See also appendix for obstruent clusters in Standard Greek (hereafter SG) and Attic Greek.

(1) Fricative + stop

*fp	*θp	*xp
ft	*θt	xt
fk fc	*θk θc	*xk xc

sp st sk sc [2]

CG does not possess underlying voiced stops. The latter appear exclusively after a nasal.

(2) Stop + fricative:

ps	ts	ks
pʃ		kʃ

([tʃ] is the allophone of /k/ before a front vowel)

There has been disagreement in the past as to whether /ps/, /ks/, /ts/ are sequences of two phonemes or contour segments in Greek[3]. It appears, though,

[2] Also ʃp, ʃt, ʃk if we were to follow Newton 1970 - the clusters in question are highly marginal.

[3] The status of Modern Greek /ps/, /ks/, /ts/ has been widely discussed in the literature. There are two opposing views: (1) Newton (1961), Koutsoudas (1962), Setatos (1969), among others, hold the position that they are sequences of phonemes while (2) Householder (1964) and Malikouti-Drachman (1970), among others, claim that they are single units (see Pagoni-Tetlow 1996: 71-72 for more details). Pagoni-Tetlow (1995, 1996), along the lines of a Government Phonology analysis, has argued that SG /ps/, /ks/ are sequences of phonemes while, on the other hand, /ts/ is a contour segment. Her analysis has been supported by phonetic evidence in Arvaniti (1987), which has shown that the latter is significantly shorter in duration than the former two.

that in CG they clearly behave like a cluster phonotactically (see also Malikouti-Drachman 2001): in the event of a yield of an illegal three-consonant cluster after the negator 'ðen' has been placed before the verb, an epenthetic [i] – called 'euphonic' in traditional grammar – separates the cluster. Thus, e.g. /ðen xtízo/, 'I don't build' becomes [en ixtízo][4]. There are plenty of examples of this kind. In all the examples that I have come across (in everyday speech or in folk literature (in, among others, Liasides 1997, Michaelides 1998, Christodoulou 1987 (ed.)), /ps/, /ks/, /ts/ are all treated identically, i.e. as sequences of two segments: e.g. [en itsilló], 'I don't push' from /ðen tsilló/; [en ipsállo], 'I don't chant' from /ðen psállo/; [en ikséro], 'I don't know' from /ðen kséro/.

Likewise, in article+noun concatenation in accusative case the illegal three-consonant clusters which arise are always simplified: /ton psépti/; /ton kséno/; /ton tsáro/ → [to psépti]; [to kséno]; [to tsáro] ('the liar', 'the stranger', 'the tzar'). When the final /n/ of the masculine article /ton/ in accusative case precedes a single word-initial consonant there are no deletions; /n/ either causes a following *stop* to become voiced (e.g. /to**n** **k**ósmon/ → [to**ŋ** **g**ózmon] 'the world' *acc* vs. [o **k**ózmos] *nom*) or itself gets totally assimilated to a following *fricative* or *sonorant* (e.g. /to**n** **f**óvon/ → [to**f** **f**óon] 'the fear' *acc*; /to**n** **l**óɣon/ → [to**l** **l**óon] 'the word' *acc*). For example, it causes the single unit [tʃ] from /k/ to be voiced in [to**n** **dʒ**erón] from /to**n** **k**erón/ *acc* 'the time' (**vs.** [o **tʃ**erós] *nom*), because the yielded sequence [ntʃ] is not offensive as it consists of *two* and not three elements.

Stop+/s/ are the only legal stop+fricative clusters. All the remaining [-cont] [+cont] clusters, i.e. clusters of *obstruent + fricative other than /s/* (i.e. *pf, *tf, *kf, *pθ, *tθ, *kθ, *px, *tx, *kx, *pv, *tv, *kv, *pð, *tð, *kð, *pɣ, *tɣ, *kɣ) are illegal. The preferred pattern in CG obstruent clusters is therefore [+cont] [-cont], i.e. a dissimilated in terms of continuancy pattern.

All the elements of the clusters in (1) and (2) must share the same value for the feature of voicing, or, rather, they must share voicelessness. Given that voiced stops do not have phonemic status in CG and also that in obstruent clusters a cluster-second member is almost always a stop, voiced obstruents clusters are impossible. [zv] (fricative + fricative) in (3) below is the only voiced obstruent cluster attested in CG precisely because its second member is not a stop.

Furthermore, obstruent clusters *sharing manner of articulation* (the same value for the feature of continuancy) are illegal (i.e. *pt, *tp, *kp, *pk, *tk, *kt, *fθ, *θf, *xf, *fx, *θx, *xθ, *vð, *ðv, *ɣv, *vɣ, *ðɣ, *ɣð).

[4] The boxes are used for highlighting or isolating certain segments for reasons of clarity.

There are three clusters, however, which do not fit this pattern and, instead, do share the same value for the feature of continuancy:

(3)

sf	*sθ	*sx	(fricative + fricative)
zv	*zð	(zɣ)	(fricative + fricative)
pc			(stop + stop)

Also:

ftʃ	*θtʃ	*xtʃ	(fricative + affricate)

Out of all obstruents, the labials /f/, /v/ and the sibilant /s/ are the only non-stops attested in a cluster-second position of a two-consonant obstruent cluster. [zɣ] clusters are extremely rare.

As can be seen, there is the following asymmetry: [sf] clusters are legal while *[sθ], *[sx] are illegal (SG [sθ], [sx] will be taken to be underlyingly /st/, /sk/ in CG as there is total absence of alternation between them, and surface unchanged).

Malikouti-Drachman treats [st], [sk] as outputs of a dissimilation rule (/sθ/, /sx/ → [st], [sk] respectively). We will support the position that in CG dissimilation applies only in verb stem+suffix morpheme boundaries where we get /ɣráf+so/ → [ɣráp+so], 'write' *dependent form, perfective aspect*; /víx+so/ → [vík+so], 'cough' *dependent form, perfective aspect* etc. Note that the outcome is [-cont] [+cont] here. Therefore p/f alternate in verb stem+suffix morpheme boundaries while /sf/ and /pc/ surface unchanged defying the dissimilation constraint by sharing the same value for continuancy. Thus we get [sféra], 'sphere' and not *[spéra] (while in SG both e.g. [asθenís] ~ [astenís]; [sxára] ~ [skára] appear) or [kupcá] and not *[kufcá] as opposed to [mmáθca] and not *[mmátca] from /mmátia/.

The presence of [zv] clusters (from /sv/) makes the system more symmetrical (since [sf] clusters *are* legal). These are the only ***voiced* obstruent clusters** that can be found in CG precisely because neither element in them is a stop.

Another important observation is that the sibilants /s/, /z/ have always behaved in an aberrant way and never in line with the other fricatives. That raises a question as to whether they form a class with them.

2.2 Sonorant + obstruent clusters

Nasal + oral stop

All nasal+oral stop clusters listed below can be either *word-initial* or *word-internal,* with the exception of [ɲɟ], which is *word-internal* only:

(4)

mb
nd
ɲɟ
ndʒ
ŋg

Liquid + obstruent

[rt] and [rk] clusters can be either ***word-initial*** or ***word-internal,*** while [rp] ones are ***word-internal*** only. [rc] and [rtʃ] clusters are ***word-internal*** only, too:

(5)

rp
rt
rk
rc
rtʃ
rf
rv
rs

[rf], [rv], [rs] can be underlying or non underlying (from /lf/, /lv/, /ls/). The non underlying **[rf], [rv], [rs]** clusters have become extremely unusual in young people's speech. It appears, though, that the newer forms with [lf], [lv], [ls] do not constitute part of the phonological system of CG that we are discussing here. They are SG forms which are becoming more and more dominant in present-day CG. Young people up to the age of 20+ now use a fair amount of SG forms in their speech, sometimes in parallel with the CG equivalent forms. As far as my observations can tell, SG [+cont][+cont] clusters are more tolerated than

[-cont][-cont] ones. For example, [arçî], [sxolío] seem to be used by young Cypriots along with [arcí], [skolío] but never SG [aftoktonó] instead of CG [aftoxtonó].

2.3 Obstruent + sonorant clusters

(6) Stop + sonorant

pr	tr	kr
pl		kl
	(tm)	(km)
pn		kn

(The clusters in parentheses are extremely rare, if at all existent in spoken language)

(7) Fricative + sonorant

fr	θr	xr[5]
fl	(θl)?	xl
fn	(θn)?	xn
vr	ðr	ɣr
vl		ɣl
(vn)		ɣn

zm (from underlying /sm/)

[5] In the past one could find [vr], [ðr], [ɣr] in place of underlying /fr/, /θr/, /xr/. This phenomenon, though, is becoming old and is no longer attested at least in young people's speech. One can find hundreds of such examples in older Cypriot poetry. It appears that the system was even more uniform throughout: e.g. we had [r+C] from both /r+C/ and /l+C/ while [C+l] was the output from both /C+l/ and /C+r/ throughout in older CG. Other processes that used to be productive in obstruent+sonorant clusters in older CG are fricative devoicing before /n/ (/ɣn/ → [xn] etc.) and voicing of intervocalic voiceless fricatives (e.g. [kavenés] for [kafenés] 'coffee shop').

2.4 Sonorant clusters

(8) Liquid+nasal

The clusters below are found only ***word-internally***:

rm (lm)

rn

(9) Liquid+liquid

The cluster below is found only ***word-internally***:

rl

(10) Nasal+nasal

The cluster [mɲ] is not underlying and is found only ***word-internally*** (/miV/ → [mɲV]) while [mn] can be found both ***word-initially*** and ***word-internally***:

mn

mɲ

2.5 Geminates

Geminate stops, fricatives, affricates and sonorants are attested in CG (see Malikouti 1998, Arvaniti 1999, 2001, Arvaniti & Tserdanelis 2000, Tserdanelis & Arvaniti 2001).

Geminate stops

Geminate stops (/p:^{h}/, /t:^{h}/, /k:^{h}/) are underlying and can be either ***word-initial*** or ***word-internal*** (SG lacks geminates). They behave like clusters when preceded by word-final /n/; hence Newton's (1970) position that the CG geminates are double consonants. More specifically, when an offensive three-element cluster is yielded it is not tolerated and causes the nasal to be deleted (e.g. /tin tt^{h}inú/ → [ti tt^{h}inú] 'the-Christina' *dim*). On the other hand, they behave like single units in morphophonemic alternations (see Malikouti 1998). Malikouti (1998), Arvaniti (1999, 2001), Arvaniti & Tserdanelis (2000) and Tserdanelis & Arvaniti (2001) take them to be monosegmental; hence the different notation in the literature. All

geminate stops are *aspirated* [6]. Length in consonants is *contrastive* as minimal pairs suggest (e.g. [péfti], 'Thursday' vs. [p:^{h}éfti], 'fall' *3rd sg present* / [míla], 'apples' vs. [mîl:a], 'fat' *noun* etc.).

Geminate fricatives

Geminate fricatives are derived by assimilation. When a nasal is followed by a fricative the former gets totally assimilated to the latter. With the exception of [s:], which is also found word-initially (e.g. /s:ázo/, 'slaughter') and can be underlying or not, all geminate fricatives listed below are not underlying, are found only ***word-internally*** and are heteromorphemic: [θ:]', [f:], [ɣ:], [s:], [ʃ:].

Geminate affricates

tʃ:^{h}

Geminate sonorants

The following are geminate sonorants, which can be found both *word-initially* and *word-internally:* /n:/, /m:/, /r:/, /l:/. They can be underlying or not. [r:] is an exception: it is not underlying and can be found only *word-internally.*

2.6 Three-consonant clusters

Three-consonant clusters are restricted, with one exception, to C+stop+sonorant in CG, which can be found either ***word-initially*** or ***word-internally***:

[6] See Arvaniti & Tserdanelis (1999) for more on the acoustic characteristics of geminates in CG. They have found that the closure duration of geminates is considerably longer than that of singletons but that they do not affect the duration of the preceding vowel. As far as *geminate stops* are concerned, they are heavily aspirated and lengthening affects both closure duration and VOT. See also Arvaniti (1999).

[7] Forms with [t:^{h}] from underlying /nθ/ are also attested in some CG varieties.

(11)

str	spr	skr
spl	*stl	skl

mbr (from /mpr/)
ndr (from /ntr/)
mbl
ŋgl

Clusters like [mbl] as in [mblázo], 'approach' and [ŋgl] as in [ŋgléo], 'pick' are not particularly frequent.

The clusters [rfc] and [xtr] are found only ***word-internally*** and are not as common as the ones in (11). [rfc] is not underlying and since it does not conform to the pattern C+stop+sonorant it is safely counted as heterosyllabic: [r.fc]. It is attested in words derived from underlying /rfiV/ (e.g. /aðérfia/ etc.), which used to surface as [a(ð)érca] and now are [aðérfca] etc. It is allowed to surface because both [rf] and [rc], at both edges of the cluster, are legal in CG.

The latter cluster is attested, at least to my knowledge, only in derivatives of the stem /oxtr-/ ([oxtrós], 'enemy'; [oxtrévume] 'be hostile' etc.). It is allowed to surface for the same reasons as [rfc]. Clusters like, for instance, *[rθc] or *[rxc] are not allowed because *[rθ]/*[rx] are offensive combinations.

3. The syllabification of Cypriot Greek - Previous accounts

Malikouti-Drachman (1999b, 2000) has proposed that the CG syllable features non-branching codas and branching onsets as long as the latter are of rising sonority. Branching onsets can consist of up to two segments. A tri-consonantal sequence is syllabified heterosyllabically. Two consonants forming a sequence of falling sonority span the syllable boundary and therefore occupy the coda and onset position of two adjacent syllables. Codas cannot be less sonorant or more complex than following onsets (also supported in Drachman & Malikouti-Drachman 1996) and therefore a 'maximal' consonant (= voiceless stop) is an ideal onset of the CG syllable when the coda of the preceding syllable is filled. She proposes in effect that medial clusters of *rising sonority* are *tautosyllabic* while those of *falling sonority* are *heterosyllabic* as dictated by SSG.

Therefore, according to her, *word-internal coda* position in CG can be occupied either by a *sonorant* or an *obstruent*. Obstruent clusters of falling or equal sonority are syllabified heterosyllabically:

(12)

e.g. obstruent codas:
[ás.tron]
[mmáθ.ca]
[kup.cá] etc.

sonorant codas:
[kar.cá]
[kum.bí] etc.

Her syllabification is based on the concept that, in CG, an onset must be ***'stronger'*** than a coda to its left. By 'strong' she means a 'maximal' consonant and therefore the 'weaker' the segment the more sonorant it is. In OT terms, the syllabification of CG is regulated by two constraints: **strong onset** and **weak coda**, ranked in the constraint hierarchy of CG as follows: Strong Onset > Weak Coda. The above two constraints work in conjunction with SSG (expressed in the form of a sonority constraint (SON SEQ) in her OT analysis).

According to Malikouti-Drachman (2000), the 'strong onset' and 'weak coda' constraints hold not only in CG but Greek in general. They are employed to resolve three issues: firstly, the obligatory process of dissimilation in oral clusters is formally expressed through these two constraints. More specifically, e.g. in CG /ɣð/, /vð/ etc. dissimilate in terms of continuancy (as well as becoming voiceless) and are syllabified *heterosyllabically* as [x.t], [f.t] respectively, because of the pressure exercised on them by the above two constraints. Essentially, the 'strong onset' and 'weak coda' constraints do the job of an **OCP constraint**, which was initially employed by Drachman & Malikouti-Drachman (1996) in an earlier OT version of her analysis as a way of accounting for formalising continuancy dissimilation but has been abandoned in the second author 's recent studies.

Secondly, the 'strong onset' and 'weak coda' constraints practically determine the *direction* of continuancy dissimilation in obstruent clusters, i.e. [+cont][-cont] and not the other way round. Sequences of [-cont][+cont], which could have been

yielded by dissimilation, are avoided. Thus one-way as well as two-way continuancy dissimilations are accounted for:

(13)

/plek+tó/ 'knitted item' → [plextó] not *[plekθó] (one-way dissimilation)
/plék+θike/ 'be knitted' *3rd p sg past* → [pléxtike] (two-way dissimilation)
(/eplék+θi/ 'be knitted' 3rd p sg past → [epléxti] is the equivalent CG form)

In the latter example, the underlying /kθ/ is a [-cont] [+cont] cluster and therefore its two segments are already dissimilated in terms of continuancy. Here the constraints of 'strong onset' and 'weak coda' dictate that /kθ/ change into [xt] ([+cont] [-cont]). Therefore it is the coda which is 'weakened' (e.g. /plektó/ → [plex.tó]) while the onset gets strengthened (e.g. /plékθike/→ [pléx.tike]). Here the [-cont] [+cont] outputs as in verb stem+suffix morpheme boundaries (e.g. /ɣráf+so/ → [ɣráp+so]) is ignored.

Thirdly, the *voicing* of the obstruent cluster is also dictated by the same constraints. Namely, in an obstruent cluster whose members do not share voicing, it is the onset (the strong element) that imposes its value to the preceding coda and not vice versa: (e.g. /aniɣ+tó/ 'open' → [anix.tó]: here /t/ imposes voicelessness on /ɣ/ which becomes [x]). It is therefore the onset segment that imposes its features on the coda one to its left (Malikouti-Drachman 2000: 405).

Malikouti-Drachman (2000) has attempted to account for the syllabification of all varieties of Modern Greek by means of SSG and, in a way, extend Steriade's (1982) analysis of Attic Greek to Modern Greek. Steriade (1982) syllabifies a sequence of two word-internal obstruents heterosyllabically in Attic Greek, by taking the first obstruent to be a coda and the second one to be the onset of the following syllable:

(14)

as.tron
^{h}es.pe.ra
e.lek^{h}.t^{h}i
ok.to
ark.tos
ar.t^{h}ron

Steriade (1982) syllabifies VCCV strings as VC.CV when CC do not obey the SSG (i.e. are of falling sonority) and as V.CCV when CC obeys it (i.e. are of rising sonority). She arrives at this syllabification led by the fact that the first V in VCCV counts as *light* when it is followed by a cluster of rising sonority whereas, when it is followed by a cluster of falling sonority, V syllabifies with the C to its right (VC.CV), which makes the syllable heavy. The sonority scale proposed by Steriade (1982) for Attic is: **stops, fricative, nasals, liquids**. A minimum distance of two positions has to be observed by the members of a legal onset or coda.

It appears that Malikouti-Drachman (2000) espouses this analysis, with some modifications, for the syllabification of CG. She claims, however, that Steriade 's (1982) analysis does not give an explanation as to why a *dental* segment is disallowed from an Attic coda position while velars and labials are not. In particular, she evokes a phonotactic constraint which did not allow a *dental* segment in the coda position of the Attic syllable:

(15)
/tit.ko/ → [tik.to] 'bear' *verb*
/kekomiδ.ka/ → [kekomika] 'look after', *pluperfect*

She then goes on to account for it by proposing a gradation in the strength of Attic Greek obstruents, which is dependent upon their **place of articulation**. She proposes that labial segments were weaker than velar ones, which in turn were weaker than dental ones: **dental > velar > labial**.

Malikouti-Drachman (2000) argues that Morelli's (1998) approach fails to account for the metrical conventions of Attic Greek described above (with regard to light vs. heavy syllables).

Likewise, with regard to Modern Greek (including CG), Malikouti-Drachman (2000) maintains that the syllabification of obstruent clusters proposed by herself for CG is not only a consequence of the hierarchy of *strong onset* > *weak coda* (in OT terms) but also a consequence of a place of articulation hierarchy that holds in Modern Greek: **dental / velar > labial.**

The above hierarchy classifies labials as weaker than dentals/velars. As can be seen, the notions of ***'strong'*** and ***'weak'*** have been applied to *place of articulation*, too. Under that assumption, the *dental and velar stops*, which are the strongest stops, are possible cluster-second members because they *can occupy the onset position* of the CG syllable as they satisfy the 'strong onset' constraint

([f.t], [x.t], [f.k] etc.), while labials are not possible cluster-second members (in any variety of Modern Greek) because they *cannot occupy the onset position* (*[f.p], *[θ.p], *[x.p]) as they are weak and do not therefore satisfy this constraint.

According to Malikouti-Drachman, the fact that [sf] clusters[8] are attested in CG while [sθ]/[sx] are absent (SG [sθ]/[sx] appear in CG as [st]/[sk] respectively) is a consequence of the 'place of articulation hierarchy'. Because /f/ is incapable of satisfying the 'strong onset' constraint it stays unchanged: "The fact that the labial articulation does not get strengthened is due to the impossibility of satisfying the demand for a STRONG ONSET, which dictates that the optimal syllable has to begin with a low sonority consonant followed by a vowel. *(my translation)*" Malikouti-Drachman (2000: 404). By that it is essentially meant that, since /f/ in a /sf/ cluster is – being a labial – a weak segment and as such unable to satisfy the demand for strong onset, it will not change into [p].

The 'place of articulation hierarchy' is used by Malikouti-Drachman (2000) to also account for the discrepancy:

/xápp^{h}ia/ → [xáp.ca] 'pills' ([xáp.p^{h}in] *sg*) NOT *[xáf.ca]

vs. /putt^{h}ia/ → [puθ.cá] 'female genitalia' *colloq.* ([put.t^{h}in] *sg*) [9]

These outputs arise because the yielded three-element clusters [pp^{h}c], [tt^{h}c] are illegal.

Here she supports the stand that /p/ stays [p] in the coda of the penult in [xáp.ca] (which does not therefore become *[xáf.ca]) because by being labial and therefore weak, it is essentially very close to /f/. On the other hand, /t/ being dental and therefore strong, changes into [θ] ([puθ.cá]), so that it will satisfy the demand for a weak coda.

4. Discussion of Malikouti-Drachman's (1999b, 2000) syllabification of Cypriot Greek

In this section, following Morelli (1998), it will be proposed here that obstruent clusters are syllabified tautosyllabically in CG. Morelli (1998) argues that sonority is not what determines the syllabification of obstruent clusters: " …such principle is not relevant to obstruent clusters since it fails to account for both the markedness

[8] No reference to the problematic [rf] clusters, which do not dissimilate either, is made in Malikouti-Drachman (2000).

[9] Let us note here that examples of that kind are very scarce in CG.

relations and the implicational universals observed for onset obstruent clusters" (Morelli 1998: 4). See also Ohala (1992) for a critique of the concept of sonority hierarchy in phonology.

In particular, after looking at a number of 25 natural languages she proposes that fricative+stop clusters are to be considered *unmarked* in the languages of the world, while stop+fricative ones (which conform to the sonority principle) are to be considered *marked.* Thus there are languages which allow only fricative+stop clusters in the onset position of a syllable but not the opposite (stop+fricative), while, on the other hand, if stop+fricative clusters are tolerated in a language then fricative+stop are obligatorily tolerated in that language as well.

Obstruent clusters of the type fricative+stop, i.e. of falling sonority, are found in *great abundance* both word-initially and word-internally in CG. In fact, as stated earlier, they are by far the most common and most frequently-attested obstruent clusters to be found in this dialect. High frequency, to begin with, should be a strong indication that the clusters in question are *tautosyllabic* and that they therefore occupy the onset position of the CG syllable. Occam's razor, the principal of ontological economy, which suggests that we should keep things simple unless there is sufficient reason not to do so, would also point towards the same direction.

An important indication that points towards the tautosyllabicity of the clusters in question is that native speakers of CG would intuitively syllabify fricative+stop clusters as *onsets* of the same syllable and, as I have noticed, quite confidently.

That is corroborated by a traditional Cypriot song, which is linguistically telling. In this song, words are broken down into their syllables. More specifically, the tri-syllabic string *va ra va, ve re ve, vi ri vi, vo ro vo, vu ru vu* is inserted (almost like a kind of an 'infix', which is the term that we will informally use here to refer to it) into words, breaking them down into their syllables, or is placed after the final syllable of a word. The ***rhyme*** of the 'infixed' syllables has to agree with the ***rhyme*** of the (word-final or word-initial/medial) syllable preceding them. For instance, the verse [eʃé nan ástron tʃe mitsín] is broken down as follows:

(16)
eʃé ***ve re ve*** nan á ***va ra va*** stron

tʃe ***ve re ve*** mitsín...[10]

There was a little star...

The song does not go:
* ás ***va ra vas*** tron

As we can see, the crucial form [ástron] ('star') is syllabified as [á.stron] and not as *[ás.tron].

When, however, the 'infix' is placed after a word-final syllable ending in **/s/** (e.g. **tis** = 'of the' *fem gen*; **tus** = 'the' *masc acc pl*) it has to rhyme with it perfectly. This indicates that it is the ***rhymes*** of the infix and its preceding syllable which are involved and not just the vowels (nuclei):

(17)
tis ***vi ri vis*** pellís...
... the rash (young girl)...

... mes tus ***vu ru vus*** eftá ***va ra va*** planí ***vi ri vi*** tes...
... in the seven planets...

The difference between (16) and (17) is that the [s] in [á.stron] is in the *onset* of the ultima while the [s] in [tis], [tus] is in the *coda*. If the former were in the coda it would have to be *[ás varavas tron] instead.

This device of the tri-syllabic string insertion counts as a word game[11], which can reveal strong implications about the syllable structure of a language. In fact, it

---

[10] To be more accurate, the example (16) above should be [eʃé *ve re ve* nan á *va ra va* stro **ndʒe** *ve re ve* mitsín]. The concatenation of [ástron tʃe] gives [ástro **ndʒ**e].

[11] At this point I would like to emphasise how crucial this traditional song proved to be for the present study. The following problem arose when my Cypriot informants were confronted with my SG accent. They would switch to speaking SG to me, which is an *automatic* process for them impossible to be overridden. As a result, eliciting the information I needed from them proved difficult. I then attempted to use English when addressing them but this resulted in a rather awkward situation and was also abandoned. I finally resorted to another method which consisted in changing the words of the song in question and conveniently using the infix in crucial places. For instance, questions of the type: "Would you say [....má varava θca...] or [...máθ varava ca...]?" were asked. Given the fact that this is perhaps the most famous

falls under one of the types of word games described in Fudge (1987: 373) and in particular type (ii): "Typically word games seem to be of three kinds [...] (i) where the segments of words or roots are completely reversed [...] (ii) where whole syllables are moved [...] or inserted in such a way as to leave existing syllables undisturbed [...] (iii) where parts of syllables are moved or material is deleted and/or inserted so as to break up existing syllables." Incidentally, this CG language game also points towards the postulation of a rhyme constituent within the syllable, which is what Fudge (1987) argues for.

With regard to native speakers' intuitions on syllabification issues, Joseph & Philippaki-Warburton (1987: 241) mention that if native speakers of Greek were asked to follow their intuitions in syllabifying lexical items, they would syllabify even clusters non-attested in word-initial position as onsets of the same syllable (e.g. [va.θmós], 'grade' rather than [vaθ.mós]); a position that I would surely support.

Kaisse (1988)[12], too, is led to support the position that Greek follows the principle of **maximisation of onsets**: "If fricatives are more sonorous than stops, creation of onsets with fricatives preceding stops cannot be motivated by an improvement of the sonority profile of the syllable. Thus whatever the syllabification of the sequence ft in [ɣráftike], the consonants are unambiguously within a single onset in [fteró]. Moreover, the creation of stop+s clusters [...] goes quite in the wrong direction for the sonority differential hypothesis." Kaisse (1988: 15).

Another compelling piece of evidence favouring our position comes from the fact that CG *deletes unstressed word-initial vowels*, clearly showing a preference even for words starting with **1)** bi- consonantal clusters of falling sonority; **2)** tri-consonantal clusters; **3)** geminates:

(18)

[rtónno]	instead of	[ortónno]	'rise, achieve' (SG [orθóno])
[ndʒízo]	instead of	[andʒízo]	'touch'
[stráfti]	instead of	[astráfti]	'it's lightening'

traditional Cypriot song, this method proved to be successful as the intuitions of my informants were exploited in a rather natural and unforced way.

[12] Unpublished paper kindly provided to me by the author (for which I would like to thank her) and also cited in Kaisse (1992).

[m:áθca] instead of [am:áθca] 'eyes'
[ʃ:íl:os] from underlying /skíl:os/ 'dog'

Deletion of unstressed word-initial vowels is pervasive in the CG phonology. In fact, Menardos (1894) points out that the deletion of unstressed word-initial vowels is a rule in CG (Cypriot poetry is rife with examples of this kind) and provides a short list of exceptions to this rule, which includes a handful of words starting mainly with [e]. Another exception is the past tense augment (e or more rarely i), which never drops out, whether stressed or unstressed, as it marks a grammatical category. In OT terms, we would say that CG seems to rank 'ONSET' quite high up in its constraint hierarchy. It should be also noted that the phenomenon of unstressed word-initial vowel deletion remains totally healthy in present-day CG.

But the mere fact that underlying /(C)CiV/ strings, which do not violate any sonority constraints, surface as [CCV], whereby the CC cluster is always of falling sonority (word-initially or internally), should be indicative of the tendencies in CG phonology. If, after all, non-violation of SSG is of paramount importance then why do /(C)CiV/ strings not surface unchanged? It appears, therefore, that the presence of the onsetless V in the string is more offensive than SSG violation. In the [CCV] surface form, this V will cease to be onsetless.

Moreover, it is important to point out that in CG there are no consonant clusters which are attested *exclusively* word-initially. This indicates that the first element of a word-initial cluster should not be counted as an extra, word-peripheral consonant, i.e. a consonant that appears *only* at the periphery of a word (left periphery in our case). Such 'extra consonantal options' have been handled in the theory as appendices, degenerate syllables (in GP) etc. (see van der Hulst & Ritter 1999: 13-14). As van der Hulst & Ritter (1999: 13) explain: "An important discovery has been that the deviations from the simple schema [CV schema] can be limited in some languages to word edges only. It is well-known, for example, that extra consonants can occur on the left or right periphery of words, leading to initial or final clusters which we do not encounter word-internally as syllable-initial or syllable-final clusters, respectively." However, in CG deviations from a basic CV schema are not restricted to the periphery of the word. On the contrary, all word-initial clusters without exception whatsoever also appear word-internally while, on the other hand, there are word-internal clusters which do not appear word-initially. In other words, possible word-initial clusters are a ***subset*** of word-internal ones.

Evidence for the CG syllable can be also drawn from the fact that CG manifests a very strong tendency for word-final *open syllables*. Word-finally, there is virtually one *canonical* segment, i.e. /s/, which can occupy the coda position of a syllable. Even a word-final sonorant (/n/) drops out - with one exception - if no other lexical item follows it or gets totally/partially assimilated to a word-initial onset following it. Certainly, we would therefore anticipate that also the CG word-internal syllable would not deviate from the word-final one to such an extent that it would have to feature an obstruent coda.

It is surely odd to claim that a language which deletes word-initial vowels showing a preference to forms with the onset position filled (by three consonant clusters, two-consonant clusters of falling sonority or even geminate consonants) and shows a strong preference to open (codaless) word-final syllables can be regarded as favouring codas to onsets.

The 'place of articulation hierarchy' that Malikouti-Drachman (2000) proposes for CG, is to be rejected for the following reasons. To begin with, CG deletes *all three voiced fricatives* (/v/, /ð/, /ɣ/) intervocalically. If /ð/, /ɣ/ were stronger than /v/ they would most likely resist deletion between vowels. In Danish, for example, among the voiced stops, /b/ remains unchanged intervocalically, /d/ is weakened to /ð/ and /g/ is deleted altogether. On that basis it has been proposed that the place of articulation hierarchy for Danish is labial>alveolar>velar (see Katamba 1989: 106). Nevertheless, the evidence from CG with regard to the intervocalic behaviour of voiced fricatives does not point towards a place hierarchy of a similar kind.

Furthermore, Malikouti-Drachman treats [st], [sk] as outputs of a dissimilation rule which derives [st], [sk] from underlying /sθ/, /sx/. But this is clearly not the case: /st/, /sk/ are underlying clusters in CG which surface unchanged. But even if we were to treat them as outputs of dissimilation, Malikouti-Drachman's (2000: 404) argument that "labial articulation does not get strengthened [i.e. /sf/ does not become [sp]] because of the impossibility of satisfying the demand for a STRONG ONSET" *(my translation)* is a statement which sounds rather like special pleading. On the contrary, it is /sθ/ clusters which would be expected to stay unchanged, since the dental /θ/ would satisfy the constraint for strong onset, rather that /sf/ not being expected to change into [sp]. After all, p (in [sp]) is stronger than f (in [sf]) because p is a *stop* while f is a *fricative*. If anything, it looks as if it is f that behaves like an 'honorary stop' since it is *exceptionally* allowed as a cluster-second member in obstruent as well as in sonorant+obstruent clusters!

But apart from [sf] and [rf] clusters, which end in a labial, labials ARE possible cluster-second members in [sp] (from /sp/), [rp], [rv], which are quite common.

Let us now address the issue of the plural forms of [xápp^{h}in] / [xápca] from /xápp^{h}ia/ vs. [putt^{h}in] /[puθcá] from /putt^{h}ia/. These two are quite straightforward examples where degemination (see also Newton 1970: 56) has applied and has given rise to the intermediate forms /xápp^{h}ja/ → xápja ; /putt^{h}ja/ → putjá, which in turn yield [xápca] / [puθcá] respectively, just like any relevant example in our data. The p in [xápca] stays unchanged because [pc]/[θc] clusters are the *normal outputs* derived from intermediate pja/tja. *[tc] would be illegal as an output in the first place as it violates the phonotactic constraint of continuancy dissimilation.

What is odd is that this argument treats the above two outputs as distinct cases from [kupcá] / [mmáθca], which clearly they are not.

Moreover, in examples like /aniɣ+tó/ → [anixtó], claiming that /t/ i.e. the strong element, imposes its features (voicelessness) on /ɣ/ turning it into [x] does not explain the direction of any assimilation process. /ɣ/ would be incapable of imposing its voicedness on the segment to its right (/aniɣ+tó/ *→ *[aniɣdó]) simply because /t/ does not have a voiced counterpart and cannot become voiced in the first place. It goes without saying that obstruent clusters involving stops have to be *voiceless*. In absence of voiced stops at a phonemic level and since the members of an obstruent cluster have to agree in voicing, there is no other option than voicelessness. That is therefore a *default case* and 'place of articulation hierarchy' does not need to explain anything here.

Finally, evidence contra Malikouti-Drachman's syllabification for CG is drawn from the fact that new forms which have not undergone cluster reduction are emerging in present-day CG phonology. This applies only to former opaque examples of the type /...**rf**iV/, i.e. forms featuring postconsonantal, prevocalic labial **f**, now surface as [...rfcV] along with the older form [...rca]: e.g. [aðérfca], [omorfcá] etc. in parallel with the opaque forms [a(ð)érca], [omorcá] etc. These new forms could not possibly be syllabified with a complex coda *[aðérf.ca], *[omorf.cá] etc. Nevertheless, Malikouti-Drachman's syllabification proposal suggests exactly that. Since tautosyllabic [fc] clusters clearly violate SSG we are left with no other option but to syllabify the cluster as [rf.c].

5. The Syllabification of Cypriot Greek revisited

5.1 Word-final coda position

CG manifests a tendency for open syllables, which is in accordance with most varieties of Greek, including SG (see Malikouti-Drachman 1984, Kappa 1997). Word-finally, only three segments are allowed to occupy the coda position of a syllable, i.e. /s/, /n/, /r/ (and perhaps /l/) as shown in the examples below:

(19)

e.g.	[télos]	'end'
	[ðóron]	'gift'
	[ipér]	'hyper', 'for', 'pro'

As a word-final coda, /l/ is the most unusual out of all and is found in lexical items like [nepál], 'Nepal' etc. Thus, **/s/** appears to be the only *canonical* segment in a word-final coda position as **/r/** is restricted only to a handful of words and the fate of **/n/** as a word-final coda element is totally dependent upon the following word-initial segment across word boundaries. Thus, **/n/** surfaces word-finally only **1)** before a word-initial vowel with which it resyllabifies as in (20) and **2)** before a word-initial voiceless oral stop onset whose *place* of articulation it gets assimilated to as in (21). In the latter case, **/n/** undergoes *partial assimilation*, losing its features for place, although remaining [+nasal] (we can see in (23), that /n/ *can* lose [+nasal]). Thus **/n/** gets totally or partially assimilated to a following obstruent at both a lexical and a postlexical level.

(20)

/anθizusi**n** úla ta ðéntra/	[aθθizusi**n** úla ta ðéndra]	'all the trees are blooming'
/miðé**n** apotelésmata/	[miðé**n** apotelézmata]	'zero results'

(21)

/anθizusi**n** ta ðéntra/	[aθθizusi**n** da ðéndra]	'the trees are blooming'
/miðé**n** kópos/	[miðé**ŋ** gópos]	'zero pain'
/epîɣasi**n** péra/	[epîasi**m** béra]	'they went over there'

In all other cases, /**n**/ gets *totally assimilated* to a following consonant segment in a word-initial onset position as in (23) or *drops* when nothing follows it as in (22)[13]:

(22)

/ɪpó to miðé**n**/	[ipó to miðé]	'below zero'
/ðén kámno típota pléo**n**/	[éŋgamno típota pléo]	'I don't do anything any more'
/ta ðéntra anθízusi**n**/	[ta ðéndra aθθízusi]	'the trees are blooming'

An exception to that are *nouns*: **1)** neuter nouns not dropping /n/ in nominative and accusative singular e.g. /en kalón peði**n**/ → [eŋ galóm beði**n**], 'he's a good/nice guy' and **2)** masculine/feminine ones not dropping /n/ in accusative singular: e.g. /kalón áθropo**n**/ → [kalón áθropo**n**], 'good/nice person'; /stin meɣálin póli**n**/ → [stim me(ɣ)álim bóli**n**], 'in the big city'.

It is very likely that /n/ does not drop in the above cases because it marks grammatical categories and in particular it distinguishes masculine nouns from neuter ones in nominative and nominative from accusative case in feminine nouns.

The following examples feature a total assimilation of /n/ to a following obstruent:

(23)

/ðén kámno pléo**n** **x**orón/	[éŋgamno pléo**x** **x**orón]	'I don't do dance any more'
/éla prí**n** **f**iɣo/	[éla prí**f** **f**io]	'come before I go'

In (20) /n/ resyllabifies from the coda position to the unfilled onset position of the following onsetless syllable: e.g.

[mi.ðé.**na.**po.te.lé.zma.ta], [e.**na.**ɣa.pó] (from /ðe**n** **a**ɣapó/, 'I don't love') etc.

---

[13] Let us note that in SG instead of dropping the word-final n in verbs (in present tense, 3rd person plural), an epenthetic **e** usually appears to the right of a word-final n (e.g. [anθízun] or [anθízune] 'bloom' *3rd pl present*) changing the syllable from a closed one into an open one. Thus, the two varieties, CG and SG, employ two different means of arriving at the same end.

This conclusion is in accordance with the breaking down of the syllables in our phonologically informative song. Thus the verse [tʃe síkosen ajéran] is broken down as follows:

(24)
tʃe sí ***vi ri vi*** kose ***ve re ve*** najé***ve re ve*** ran
(not *tʃe sí ***vi ri vi*** kosen ***ve re ven*** ajéran)

...and the wind blew...

Thankfully, the song features this 'critical insertion' whereby *vereve* is wedged between the nucleus and the word-final n in [síkosen], which clearly shows that the word-final **n** of [síkose**n**] *resyllabifies* with the word-initial vowel to its right: [sí.ko.se. **na**.jé.ran].

5.2 Syllabification of word-internal consonant sequences

Syllabification of word-internal consonant sequences[14] is less straightforward and needs more careful consideration. Word-internal consonant sequences of the type **sonorant+obstruent** will be discussed first. But for us to fully understand the phonology of present-day CG, we should take into account the fact that it is going through a *transitional period* at the moment with regard to syllabification issues. If we look closely at older forms of CG we will see that some of its phonotactic constraints are being relaxed. For instance, word-initial [rk] clusters are now extremely rare even in old people's speech. On the other hand, [rt] clusters – another example of a sequence with two extreme sonority poles – is still healthy in present-day CG. Secondly, new tri-consonantal clusters of the type [**rfc**] are now emerging:

(25)

UF	**Newer SF**	**Older SF**	**Gloss**
/aðérfia/	[aðérfca]	[aðérca]	'siblings'
/omorfiá/	[omorfcá]	[omorcá]	'beauty' etc.

[14] With regard to SG, Joseph & Philippaki-Warburton (1987: 241) give [ðé.rma], 'skin' [á.lma], 'jump' etc. (i.e. not even sonorants are given as codas).

Although not all possible **r+obstruent** combinations are legal word-initially ([rp], [rc], [rf], [rv] are illegal word-initially), the presence of [rk], [rt] clusters in the word-initial onset position, would still render **/r/** in word-internal sequences a good candidate for the onset position rather than the coda one. Also, why would CG prefer [rtónno] to [ortónno]? Obviously, because it does not like onsetless syllables. Thus it is possible that, for instance, [ʃérka] was syllabified as [ʃé.rka] rather than [ʃér.ka]. On the other hand, the ultima in the alternative syllabification [ʃér.ka] does fulfil the onset necessity. It looks as if **/r/** is not *firmly* established as a coda or onset element in r+obstruent clusters word-internally. However, newer CG surface forms as in (25) seem to suggest that **/r/** is being pushed to the coda position: in [aðér.fca], for example, we have no option but syllabify it as coda. The unstable syllabic position of **/r/** is reflected in Cypriots' intuitions. When asked how, say, [peθcá] is broken down they would reply [pe.θcá] without hesitation, while for, say, [karcá] they would pause and think. The majority of them, though, would still incline towards [ka.rcá].

Let us now look at **nasal+oral stop** word-internal sequences. One would perhaps expect /n/ to abandon the CG coda position and resyllabify with a vowel of an onsetless syllable to its right, as we saw in (24) above. But what happens when a syllable-final /n/ is found before a syllable-initial stop onset to its right? Once again our song features another 'crucial insertion'. Let us look at it more closely. The verse [mes tiŋ garcán] from /mes tin karcán/ is broken down as follows (If we take [tin karcán] to be a prosodic word with one stress on the final syllable we should note that the behaviour of the prosodic word of, say, [tis karcás] is asymmetrical in that the final consonant in [tis] does not resyllabify as onset in, say, [tis.pellís] (*[ti.spellís]). It seems that there is a difference in *status* of coda /s/ and coda /n/ rather than of place of word boundary):

(26)
mes ti ***vi ri vi*** ŋgarcán
(not *[mes tin ***vi ri vin*** karcán])[15]

...in the heart...

[15] Remember that tin virivi would not be an option because the rhyming has to be perfect: we can have either **tin virivin** or ti virivi.

What is interesting is that again the n in [tin] resyllabifies with the consonant to its right: [mes ti.**ŋga**rcán]. Knowing that **nasal+oral stop** clusters are legal word-initially (unlike SG where the nasal drops out before a voiced stop word-initially), this syllabification should not come as a surprise.

We should not forget that nasals as cluster-first elements are exclusively found before oral stops (they totally assimilate to fricatives). Another argument which should allay our suspicions concerning the syllabification of nasals in **nasal+oral stop** clusters is the following: there are some CG words starting with N+oral stop clusters which have resulted from **N#voiceless stop** concatenation: e.g. [mbróeman], 'breakfast' instead of [próeman] (from concatenation of e.g. [enan # próeman], 'a breakfast'); [mbroú], 'a while ago' instead of [proú], [mbótis], 'drinker' instead of [pótis]; [ŋgléo], 'pick' instead of [kléo] from /kléɣo/ (SG /eklέɣo/, 'elect') etc. (see also Hadjiioannou 2000). That suggests that N# has been resyllabified to the following word-initial onset and is in accordance with (26).

Allowing a word-internal sonorant in sonorant+obstruent sequences to be syllabified as onset would suggest that an obstruent in the same position would make a much better candidate for the onset position. It has been shown that word-internal obstruent+obstruent clusters of falling or rising sonority occupy the onset position of the CG syllable. Finally, it goes without saying that obstruent+sonorant clusters are tautosyllabic, too.

5.3 Syllabification of CG geminates

Arvaniti (2000) and Arvaniti & Rose (2003) have argued that the CG geminates should be analysed as monosegmental (first proposed by Malikouti-Drachman 1987) and non-moraic. CG geminates are not a sequence of two separate consonants because they behave like single units in morphophonemic alternations.

It has also been supported by Arvaniti (2000) and Arvaniti & Rose (2003) that the CG geminates are non-moraic (against Hayes's 1989 proposal that geminates are *inherently* moraic) because **1)** stress is not affected by the presence of geminates and **2)** there is no minimal word requirement which dictates that geminates should be contained in a minimal CG word (see Davis 1999 for analogous examples in Trukese). As a result, Davis's (1999) model, which represents non-moraic geminates as *two* root nodes, does not accommodate the CG geminates.

Broselow et al's (1997) proposal for mora sharing geminates (non-moraic geminates *share* the mora of the preceding vowel) is also turned down by Arvaniti (2000) and Arvaniti & Rose (2003) for CG geminates because the latter do not shorten the vowel of the preceding syllable. Arvaniti & Tserdanelis (2000) have found that the duration of the vowel preceding a geminate is not systematic in CG.

Finally, Arvaniti (2000) and Arvaniti & Rose (2003) propose that CG geminates should therefore be *doubly linked to two timing positions,* agreeing with Hume et al (1997) that a representation of geminates should involve both a skeletal and a moraic tier.

Assuming that CG accent is analysed as a bitonal L+H (as in SG), Tserdanelis & Arvaniti (2001) looked at the *position* of the L(= low) tone in a stressed CG syllable (they checked whether L would appear at the onset of a stressed syllable or in the middle of a geminate's duration). It turned out that the L tone appeared in the middle of a geminate's duration. That evidence has led Arvaniti (2000) to maintain that CG geminates are *ambisyllabic*.

On the other hand, the native speakers' intuitions would suggest that CG geminates are tautosyllabic. Also our traditional Cypriot song features the following critical verse:

(27)

[pos é**nn**a páo péra]

pos e ***ve re ve*** **nn**a pa ***va ra va*** o pera

...that I will go out there...

Here the breaking down of the words would suggest that the geminate occupies the onset position.

5.4 Conclusion

The CG syllable prefers to have an onset rather than be onsetless while, on the other hand, it strongly prefers to be codaless rather than have a coda. The word-internal CG coda position can be filled only by a sonorant segment.

Appendix

Obstruent clusters in Standard Greek and Attic Greek

In SG, the inventory of obstruent clusters is larger mainly due to the fact that it permits both Katharevusa and Demotic phonological patterns. The following bi-consonantal obstruent clusters are attested in SG:

(5) Stop + stop

pt	*tp	*kp
*pk	*tk	kt

(6) Fricative + fricative

fθ	*θf	*xf
fx	*θx	xθ
vð	*ðv	*ɣv
vɣ	*ðɣ	ɣð

'Mixed' clusters of stops + fricatives are abundant in SG; they all have to share the same value for the feature of voicing:

(7) Fricative + stop

*fp	*θp	*xp
ft	*θt	xt
*fk	*θk	*xk
sp	st	sk

(8) Stop + fricative

ps	ts	ks

If we now look at Attic Greek, we will see that its consonant inventory did not include fricatives except for /s/, /z/. The Modern Greek phonemes /f/, /θ/, /x/ used to be the aspirated stops /p^h/, /t^h/, /k^h/ respectively, which through the centuries developed into voiceless fricatives. In Attic Greek there was a three-way contrast in stops: unaspirated voiceless stops, aspirated voiceless stops and voiced stops

(see also Steriade 1982). Two-consonant clusters of obstruents consisted exclusively of two segments that shared *aspiration* and *voicing* (or, in terms of feature geometry, clusters of stops had to share all their *laryngeal* features). They also shared *continuancy*. Therefore legal clusters were those which consisted of *two voiced stops* or *two unaspirated voiceless stops* or *two aspirated voiceless stops*:

(9)

pt	*tp	*kp
*pk	*tk	kt

(10)

pʰtʰ	*tʰpʰ	*kʰpʰ
*pʰkʰ	*tʰkʰ	kʰtʰ

(11)

bd	*db	*gb
*bg	*dg	gd

Stop clusters of voiced and voiceless stops were not permitted:

(12)

*pb	*tb	*kb
*pd	*td	*kd
*pg	*tg	*kg
*bp	*dp	*gp
*bt	*dt	*gt
*bk	*dk	*gk

Neither were stop clusters of aspirated and unaspirated voiceless stops permitted:

(13)

*pʰt	*tʰp	*kʰp
*pʰk	*tʰk	*kʰt

*pt^{h}	*tp^{h}	*kp^{h}
*pk^{h}	*tk^{h}	*kt^{h}

The exception to that is clusters consisting of the **fricative s+stop**, in which case the members of the cluster do not share all their laryngeal features: [sp], [st], [sk] and [sp^{h}], [st^{h}], [sk^{h}]. Also word-internal [zd] clusters occurred as a result of concatenation of prefix-final –s + stem-initial d.

As we saw earlier, in CG, this scenario has been reversed: obstruent clusters sharing continuancy are *illegal*. In a reversal of fortunes, for an obstruent cluster to be legal, its members have to be dissimilated in terms of continuancy; namely the cluster has to consist of a fricative+stop. Interestingly enough, the phonotactic pressure for obstruent clusters to share the same value of the feature of *voicing* still holds in all varieties of Greek.

Acknowledgements

I would like to thank John Coleman and Amalia Arvaniti for their useful comments regarding gemination. The usual disclaimers apply.

References

Arvaniti, A. 1999. Effects of speaking rate on the timing of single and geminate sonorants. In *Proceedings of the XIVth International Congress of Phonetic Sciences*: 599-602.

Arvaniti, A. 2000. Cypriot Greek and the phonetics and phonology of geminates. *Proceedings of International Conference for Modern Greek Dialects and Linguistic Theory*. Patras: 12-14 October 2000.

Arvaniti, A. 2001. Comparing the phonetics of single and geminate consonants in Cypriot and Standard Greek. *Proceedings of the Fourth International Conferenceon Greek Linguistics*. Thessaloniki: University Studio Press: 37-44.

Arvaniti, A. & S. Rose 2003. Two sources of evidence against moraic timing of geminates. Handout of speech presented in LSA Meeting, Atlanta, GA, January 2003.

Arvaniti, A. & G. Tserdanelis 2000. On the phonetics of geminates: evidence from Cypriot Greek. *Proceedings of International Conference on Spoken Language Processing ICSLP2000*, Beijing 23-27 October 2000.

Broselow, E., S. Chen & M. Huffman 1997b. Syllable weight: convergence of phonology and phonetics. *Phonology* 14. 47-82.

Davis, S. 1999. On the representation of initial geminates. *Phonology* 16. 93-104.

Drachman, G. & Malikouti-Drachman, A. 1996. Dissimilation in Cypriot Greek: competing analyses. *Studies in Greek Linguistics* 17: 57-71.

Fudge, E. C. 1969b. Syllables. *JL* 5. 253-286.

Fudge, E. C. 1977. Phonotactics and the syllable. In *Linguistic Studies offered to J. Greenberg on the Occasion of his Sixtieth Birthday*. Vol. 2: 381-398.

Fudge, E. C. 1987. Branching structure within the syllable. *JL* 23: 359-377.

Hayes, B. 1989. Compensatory lengthening in moraic phonology. *LI* 20: 353-306.

Hume, E., J. Muller & A. van Engelenhoven 1997. Non-moraic geminates in Leti. *Phonology* 14: 371-402.

van der Hurst, H. & Ritter, N. 1999. Theories of the syllable. In van der Hurst & Ritter eds., *The Syllable*: 13-52.

Joseph B. & Philippaki-Warburton, I. 1987. *Modern Greek.* Beckenham: Croom Helm Ltd.

Kaisse, E. 1988. Modern Greek continuant dissimilation and the OCP. Ms. Seattle: University of Washington.

Katamba, F. 1989. *An Introduction to Phonology*. London and New York: Longman. Liasides, P. 1997. *The Complete Works. Άπαντα.* Lefkosia: Cyprus Research Centre.

Malikouti-Drachman, A. 1984. Syllables in Modern Greek. In Dressler et al eds., *Phonologica 1984*: 181-186.

Malikouti-Drachman, A. 1987. The representation of double consonants in Modern Greek. *Studies in Greek Linguistics* 8: 275-291.

Malikouti-Drachman, A. 1999a. Observations on the dialectal retreat of Cypriot Greek. Παρατηρήσεις σε διαλεκτικές υποχωρήσεις της Κυπριακής. *Studies in Greek Linguistics* 20: 292-302.

Malikouti-Drachman, A. 1999b. Opaque interactions in Cypriot Greek. *Proceedings of the 4th International Conference on Greek Linguistics*. Lefkosia: University of Cyprus: 54-61.

Malikouti-Drachman, A. 2000. Syllable constraints and dialectal variety. Συλλαβικοί περιορισμοί και διαλεκτική ποικιλία. *Studies in Greek Linguistics* 21: 402-413.

Michaelides, V. 1998. *Poems. Ποιήματα*. Lefkosia: K. Epiphaniou Publications.

Morelli, F. 1998. Markedness relations and implicational universals in the typology of onset obstruent clusters. *NELS Proceedings* 28, Vol. 2. *[ROA-251-0398, http://roa.rutgers.edu/]*.

Newton, B. 1970. *Cypriot Greek: Its Phonology and Inflections.* The Hague: Mouton.

Newton, B. 1972. *The Generative Interpretation of Dialect*. Cambridge: CUP.

Ojala, J. J. 1992. Alternatives to the sonority hierarchy for explaining segmental sequential constraints. *Papers from the parasession on the syllable. Chicago Linguistic Society* 319-338.

Pagoni-Tetlow, S. 1995. The syllabic structure of Modern Greek *ts*: a Government Phonology approach. *Studies in Greek Linguistics* 15: 198-202.

Pagoni-Tetlow, S. 1996. The syllabic structure of Modern Greek *ps/ks*: a Government Phonology approach. *Studies in Greek Linguistics* 16: 71-91.

Steriade, D. 1982. Greek prosodies and the nature of syllabification. Unpublished doctoral dissertation. Cambridge Mass.: MIT.

Tserdanelis, G. & A. Arvaniti 2001. The acoustic characteristics of geminate consonants in Cypriot Greek. *Proceedings of the Fourth International Conference on Greek Linguistics.* Thessaloniki University Studio Press: 29-36.

Universal and Language-specific Aspects of Intonation in English and Polish

ESTHER GRABE AND MACIEJ KARPINSKI

1 Introduction

The intonation of Dutch, German and English is characterised by a trade-off between the number of morphosyntactic question markers in the text and high pitch. Utterances that contain fewer lexical or syntactic question markers are more likely to be produced with high pitch. This trade-off may be universal but evidence from other languages is required. In this paper, we provide data from Polish, a Slavic language. We compared nuclear accent production in English and Polish declaratives and in three types of questions. We predicted a cross-linguistic difference: English and Polish speakers would produce different distributions of nuclear accent shapes in the four types of utterances. The trade-off between high pitch and the number of question markers in the text was predicted to operate in both languages.

1.1 Background

The search for linguistic universals has attracted interest at least since the work of Greenberg (e.g. 1963). A summary of work on phonetic universals is given by Maddieson (1999). Prosodic universals were first discussed in detail by Bolinger (1978). Bolinger commented extensively on the link between high pitch and certain types of utterances. Since then, this link has become the most popular prosodic universal. The connection between high pitch and utterance type goes back to work by Hadding-Koch and M. Studdert-Kennedy (1964) and by Liberman (1967). Hadding-Koch and M. Studdert-Kennedy stated that high or rising tones co-occurred with interrogative messages. Liberman (1967) connected the observation to the control of breathing. He hypothesised that all linguistically meaningful uses of intonation can be explained in terms of the distinction between

marked and unmarked breath groups. Marked breath-groups are characterised by a final rise; unmarked breath-groups by a final fall. Cruttenden (1986) reached a similar conclusion. He suggested that all sentences could be divided into two groups, referred to as closed and open. Closed sentences are assertive and non-continuative, and tend to have a falling melody. Open sentences are non-assertive and continuative (e.g. questions) and tend to have falling contours.

A more recent and more wide-ranging approach to prosodic universals was put forward by Gussenhoven (2002). Gussenhoven suggested that the intonation of any language involves universal and language-specific components. Following earlier work by Ohala (1983), he argued that the phonetic implementation of fundamental frequency is affected by three universal biological codes: The effort code, the production code and the frequency code. The effort code involves the amount of energy expended. More effort leads to a wider excursion of pitch movements. Wider pitch movements can signal an increase in forcefulness, agitation or surprise. The production code is a consequence of the commonly observed correlation between the duration of utterances and the duration of breath groups. This code is responsible for the production of high pitch at the beginnings of utterances and low pitch at the ends and is widely used to signal topic shifts. The frequency code, finally, involves pitch height. Smaller larynxes produce higher pitch and since smaller creatures are often less powerful than larger creatures, high pitch can be used to signal submissiveness or a willingness to cooperate. Speech communities vary in the extent to which they employ the three codes and in the choices they make when the codes conflict.

The present study is restricted to the frequency code. An interpretation of the frequency code that differs from that put forward by Ohala (1983) and Gussenhoven (2002) has been offered by Chafe (1994) and Wennerstrom (2001). These authors suggest that the link between high pitch and questions stems from prelinguistic responses to important observations in the environment. Questions are usually produced as demands for interaction and they are intended to attract the listener's attention. Therefore, the speaker may want to make them especially prominent in the stream of speech. This interpretation takes us back to Bolinger (1978, 1986). Bolinger assumed that there are at least some aspects of intonation that can be accounted for in evolutionary terms. High or rising pitch can signal interest, arousal and consequently, incompleteness. Low or falling pitch can be used to express the absence of interest and consequently, finality.

Gussenhoven's (2002) language-specific component of intonation involves intonational phonology and morphology. In phonology and morphology, intonational meaning can be arbitrary. Form-function relationships may mimic the paralinguistic form-function relations in the universal component, but linguistic change can also produce arbitrary form-meaning relations. For instance, recent corpus-based research shows that in Belfast English, declaratives are produced with final rise-plateau pattern just as frequently as questions (Grabe, 2002). Cruttenden (1994) points out that in American or Australian English, the production of similar rise-plateaux represents a particular communicative choice. This is not the case in Belfast English or in a number of other dialects spoken in the North of Britain. In those dialects, statements with rising intonation are the default.

2 Earlier Work on High Pitch and Utterance Type

Haan and van Heuven (1999) and Haan (2002), Brinckmann and Benzmüller (1999) and Grabe (2002) showed that the frequency code operates in Dutch, German and English. Haan and van Heuven investigated the distribution of nuclear accent patterns in declaratives, wh-questions, yes/no questions and declarative questions (questions without morphosyntactic or lexical question markers). In their study, the production of intonation-phrase-final high pitch was affected significantly by the number of syntactic and/or lexical markers of interrogativity an utterance contained. High pitch was produced more frequently when there were fewer markers of interrogativity. Brinckmann and Benzmüller (1999) found comparable results for German. German speakers were more likely to produce high or rising f_0 at intonation phrase boundaries delimiting yes/no questions and declarative questions than at IP boundaries delimiting statements and wh-questions. Grabe (2002) showed that the pattern first uncovered in Dutch could also be observed in English. Moreover, the observation held for speakers from several dialects of English although the intonational phonological structures of these dialects differed, in some cases drastically.

3 Method

3.1 Stimuli

We constructed a set of Polish stimuli, comparable to those used in Grabe's study of English. Grabe's stimuli were taken from the Intonational Variation in English Corpus (IViE, http://www.phon.ox.ac.uk/~esther/ivyweb/, a publicly available corpus of prosodically annotated data from seven urban dialects of English spoken in the British Isles. The corpus contains data in five speaking styles; the data used for the purposes of the present study were taken from the read sentences. These included four utterance types: declaratives (e.g. *You remembered the lilies*), yes/no questions (e.g. *May I lean on the railings?*), wh-questions (e.g. *Where's the manual?*) and declarative questions (e.g. *You remembered the lilies?*). Data from six speakers were prosodically annotated.

The set of Polish stimuli consisted of statements, general questions with 'czy' (comparable to yes/no questions in English), wh-questions and general questions without 'czy' (comparable to questions without morphosyntactic question markers in English). Each Polish sentence was built to follow, as closely as possible, the syllabic frame of the corresponding IViE sentence, and it belonged to the same sentence category as its IViE counterpart. The number of syllables in the corresponding English and Polish utterances was equal. Efforts were made to keep at least the final syllables phonetically similar to those in the English sentences, because these were likely to be the domain of the nuclear melody. Note, however, the following differences between the English and the Polish phonetic and phonological systems:

The IViE sentences were deliberately composed entirely of voiced segments to facilitate auditory and acoustic analyses. Fully voiced Polish stimuli could not be designed. Polish speakers devoice the final consonant of the utterance, especially when it is followed by a pause. Even if the last phoneme of an isolated sentence is vocalic, it is often produced with creaky voice, especially when the melody declines. For instance, some of the English sentences ended in <-ing> (e.g. the place name *Ealing*). In Polish, the sequence <-ing> is untypical. Therefore, we used words of foreign origin, e.g. *doping* or *trening*. These words are relatively well grounded in the Polish lexical system. Orthographically, <-ing> endings were identical to their English prototypes, but in the Polish stimuli, speakers devoiced the final segment.

3.2 Participants

The English stimuli were produced by six speakers from each of the seven dialects. The speakers were 15 – 16 years of age. Each speaker read eight different declaratives and three examples each of wh-, yes/no-, and declarative questions. The utterances were read in random order and presented without context. Individual stimuli, rather than one single list, were presented to the subjects.

Six Polish speakers were recorded. They were 25-30 years of age. The group consisted of three female and three male subjects. They were asked to read the sentences in the same, quasi-random order in which the IViE sentences had been read. The speech signal was recorded in an anechoic chamber, using a CD recorder. The recording procedure was comparable to the one used for English speakers. Then the recordings were digitised and labelled intonationally.

3.3 Prosodic Annotation

We combined two approaches to intonation transcription: the IViE system, developed for the transcription of the IViE corpus (Grabe, Post and Nolan, 2001) and the approach to transcription taken in the *PoInt* Database project (Karpinski, 2002). This project provided the first prosodically annotated corpus of Polish speech data. In essence, our transcription was phonetic-descriptive rather than phonological. In the English and in the Polish utterances, we annotated the pitch movement equivalent to the nuclear melody. In English, the nuclear syllable is the last accented syllable in the intonation phrase, regardless of the position of that syllable in the utterance. In Polish, the unmarked position of the nucleus is the penultimate syllable of the phrase, or the last syllable if the last word is monosyllabic. This prominence may move towards the beginning of the phrase in emotional speech or in 'marked' utterances, when the speaker intends to change the default focus of a sentence. While such situations did not occur in the analysed data from Polish, questions concerning the location of the nuclear melody emerged. In some of sentences, a very strong prominence occurred early in the phrase. This prominence was especially strong in wh-questions. Although this early prominence may point to an early nucleus placement followed by a further, postnuclear accent, or the presence of two focus accents in one utterance (Ladd, 1996), in the present study, we restricted our analyses to the IP-final accent.

The shapes of the nuclear melodies observed were transcribed with the labels H, L and M. Labels were assigned from left to right. The pitch level preceding the nuclear syllable was taken into account when labels were assigned. A bitonal accent was transcribed as HL if the nuclear syllable was high and followed by a low, and if the low was equivalent to a low turning point immediately preceding the nuclear syllable. The transcription was LH if the nuclear syllable was low-high and if the high was equivalent to a high turning point immediately preceding the nuclear syllable. The transcription was HM if the low following the nuclear syllable was not as low as the low turning point preceding the nuclear syllable. The transcription was LM if the high following the low nuclear syllable was not as high as an earlier high. ML was used if the nuclear syllable was lower than the preceding high in the utterance and followed by a low, etc. If the accent was tritonal, one label was given to the nuclear syllable, one to the turning point and one to the IP-final pitch level. Again, the pitch level of prenuclear turning points in the utterance was taken into account.

4 Results

Our findings provided further evidence for the postulated combination of universal and language-specific components in intonation. We will begin with a summary of our language-specific findings.

4.1 Nuclear accent types in English and Polish

Our Polish speakers produced a smaller range of nuclear accent types than our English speakers (but note that we had data from seven English dialects). We found six nuclear accent types in Polish: HL, ML, LL, LH, LM, MH. In English, we found 13 types: HL, ML, LL, LH, LM, MH, HH, HM, LHL, LHM, HLH, MLH, MHL. We observed five types in Belfast, nine in Bradford, eight in Dublin and six in Leeds, Cambridge, London, and Newcastle. All nuclear accent types found in Polish were also observed in English. There was complete overlap between the 'standard' Polish data and the English data from Bradford, but the Bradford speakers produced three additional types. We found bitonal types HH and HM in English, but not in Polish. However, as we know from other studies, HH and HM accents *do* occur in Polish spontaneous speech (Francuzik, Karpinski and

Klesta, 2002). Finally, we observed tritonal accents in English, but not in Polish. Again, tritonal accents in Polish occur, but rarely and mostly in emotional speech. As yet, it is not clear whether tritonal accents are part of the Polish intonation phonological system or whether they are stylistic variants of bitonal accents (Karpinski, 2002).

4.2 Distribution of nuclear accent types

Table 1 shows the distribution of the nuclear accent types in Polish. Table 2 illustrates, in percent, the distribution of nuclear accent types in the Cambridge English data in the IViE corpus. This is the dialect of English in the corpus which is closest to an English 'standard' and similar to the one described in textbooks on English intonation. In Tables 1 and 2, we have highlighted the patterns that we observed most frequently for a particular sentence type. The distribution of nuclear accent shapes in the other dialects of English is given in (Grabe, 2002).

Tables 1 and 2 illustrate a number of differences in the distribution of nuclear accent shapes in standard varieties of English and Polish. The clearest difference involves the absence of tritonal accents in Polish. We also found differences in the distribution of bitonal accents, but these were not particularly striking. Polish declaratives predominantly ended in ML, the nuclear accent that seems to dominate in Polish declaratives more generally (Karpinski, 2002; Francuzik et al., 2002). Cambridge English declaratives predominantly ended in HL, but ML was also observed. Rising declaratives were observed in both languages. In English, these involved a fall-rise; in Polish, they involved a rise. In southern standard varieties of English, falling-rising declaratives are common (O'Connor and Arnold, 1973). Preliminary analyses based on a larger corpus of Polish data show that declaratives with rising nuclear accent are not unusual in Polish either (Karpinski, 2002).

In wh-questions questions, speakers from both languages used falling or rising contours. In Polish, two female and two male speakers were consistent in their choices: two of them always produced LH; the other two produced ML. This finding may reflect individual differences in speaking style or individual differences in the interpretation of the utterances. Steffen-Batogowa (1996) suggests that the variations of the melody in Polish statements may be related to the emotional load of the message. For instance, the high final tone is said to be perceived by listeners as a signal of irritation, disbelief, etc. Our speakers may

have attached certain emotional values to the sentences and these may be responsible for their choices. In recent studies of Karpinski (2002), two factors were found to coincide with the occurrence of rises in Polish statements: (a) a speaker's uncertainty about what s/he is saying or is about to say; (b) a speaker's intention or wish to continue speaking.

Polish yes/no questions ended predominantly in LH. In a recent study, Durand, Durand-Deska, Gubrynowicz and Marek (2002) state that the main characteristics of Polish 'czy' questions is a double climbing binary foot at the beginning and at the end of the sentence. He coded the intonation of the final binary foot in the analysed utterances as BT (Bottom-Top) or BH (Bottom-High), which is consistent with our results. English speakers used falling or rising contours in wh- questions.

Polish declarative questions ended in LH or MH. Cambridge English declarative questions predominantly ended in LH but in about 10% of utterances, HL was produced.

More generally, we found that our English speakers produced the widest range of contour types in declarative questions. In five of the seven dialects, we observed five or more nuclear accent types. In Polish, we found the opposite. In declarative questions, only two contour types were observed (LH and MH). Polish speakers produced the widest range of contour types in declaratives. This finding may reflect cross-linguistic differences in grammatical structure. English word order is somewhat stricter than Polish word order (Asher, 1994). Therefore, Polish speakers and listeners may rely more heavily on the use of high pitch in questions.

	DEC	WH	Y/N	DECQ
ML	79.2	50.0	5.6	
HL	10.4			
LL	4.2	5.6		
LH	2.1	38.9	66.7	94.4
MH	2.1	5.6	27.8	5.6
LM	2.1			

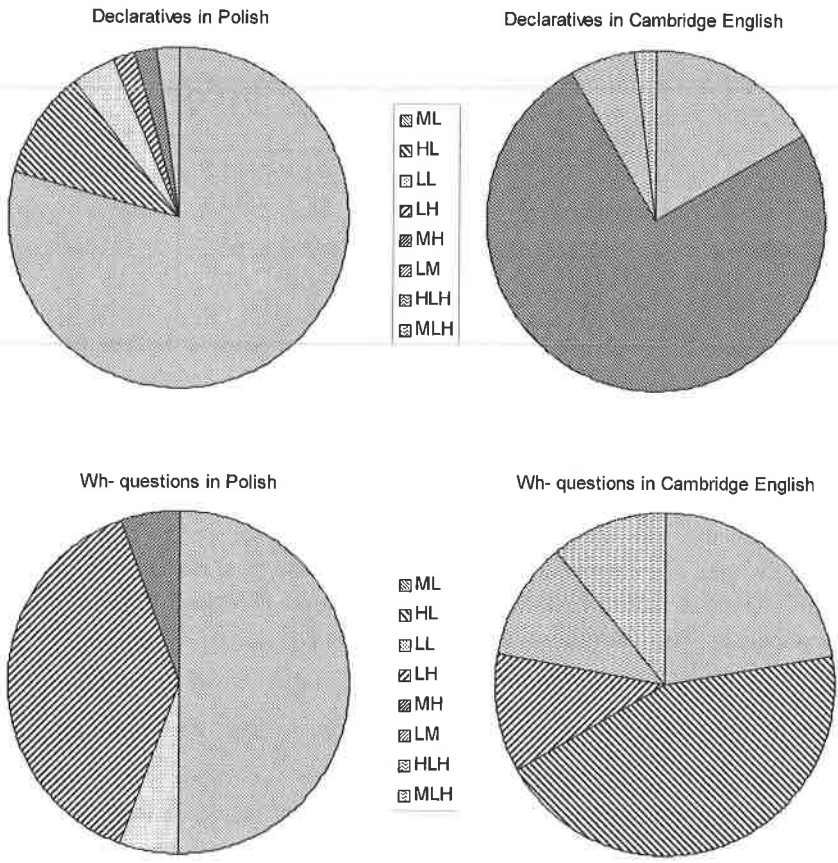
Table 1. Distribution of nuclear accent types in standard Polish, in percent. The nuclear melody is shown in the first column. DEC = declarative, WH = wh-question, Y/N = yes/no question, DECQ = declarative question).

	DEC	WH	Y/N	DECQ
ML	16.7	22.2		
HL	75.0	44.4	64.7	11.8
LH		11.1	23.5	58.8
MH				23.5
HLH	6.3	11.1	11.8	5.9
MLH	2.1	11.1		

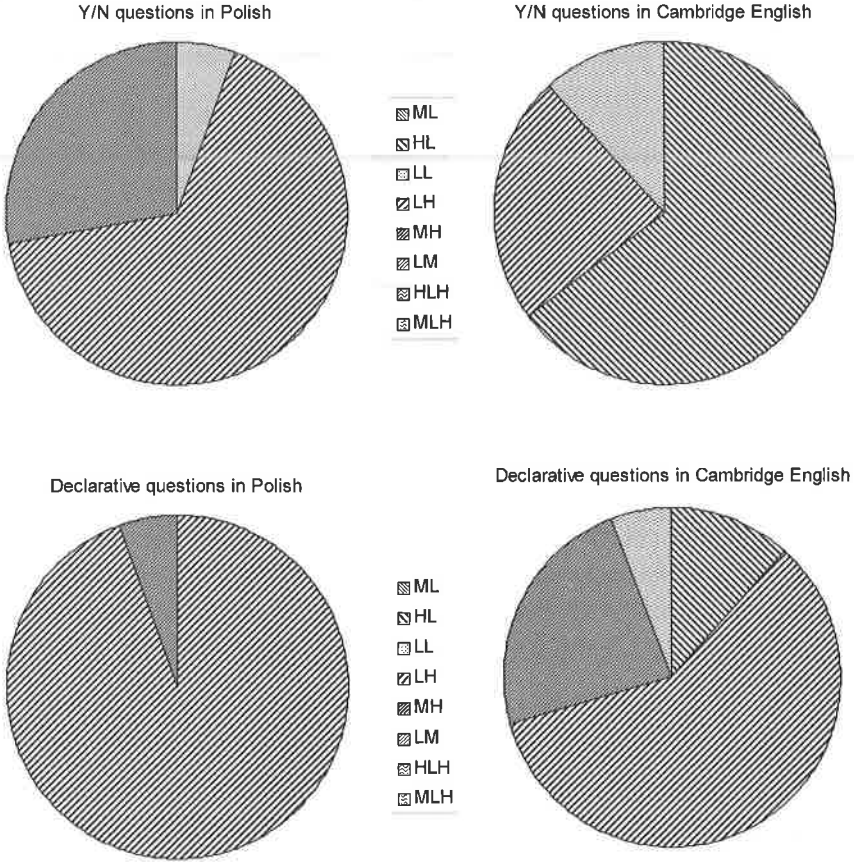
Table 2. The distribution of the nuclear accent types in the data from Cambridge English.

4.3 Cross-language generalisations

The data also provided evidence of a cross-language generalisation. Just as Dutch, German and English speakers, Polish speakers produced high pitch at the end of utterances more frequently when the utterances contained fewer markers of interrogativity. This finding is illustrated in Figure 1.



Figures 1(a)–(d). The distribution of nuclear accent types in standard Polish and Cambridge English (shown for each sentence type studied).


Y/N questions in Polish
Y/N questions in Cambridge English
ML
HL
LL
LH
MH
LM
HLH
MLH
Declarative questions in Polish
Declarative questions in Cambridge English
ML
HL
LL
LH
MH
LM
HLH
MLH

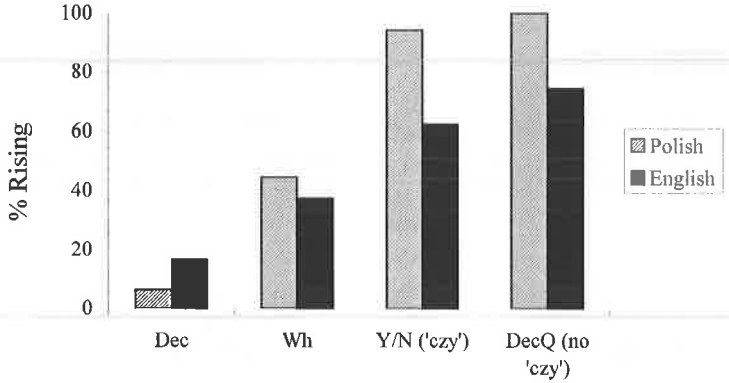


Figure 2. Incidence of rising nuclear accent types in English and Polish.

Figure 2 shows the incidence of a final rise in f_0 (LH or MH), plotted against the four utterance types. The figure illustrates a trade-off between the presence of lexical and/or syntactic markers of interrogativity in the text and the production of high pitch. English and Polish speakers were more likely to produce rising nuclear accents when the text contained fewer lexical or morphosyntactic question-markers.

5 Conclusion

We compared nuclear accent production in four types of utterances in English and Polish. Our findings support descriptions of intonation that combine universal with language-specific properties. We found evidence of cross-linguistic differences. In the four utterances types investigated, English and Polish speakers produced (a) different nuclear accent types and (b) different distributions of these types. We also found evidence of a cross-language generalisation. English, a Germanic language and Polish, a Slavic language, share a prosodic property. In both languages, the distribution of intonation patterns is affected by lexical and syntactic structure of the text.

References

Asher, R. E. (1994, Ed.).*The Encyclopaedia of Language and Linguistics*. Oxford-Tokyo, Pergamon Press, vol. IX.

Bolinger, D. 1978. Intonation across languages. In J. Greenberg (Ed.) *Universals of Human language*, Vol. II: Phonology. Palo Alto, CA: Stanford University Press: 471 – 524.

Bolinger, D. 1986. *Intonation and Its Parts*. Palo Alto, CA: Stanford University Press.

Brinckmann, C. and Benzmüller, R. 1999. The relationship between utterance type and F0 contour in German. *Proceedings of EUROSPEECH* 1999, vol. 1: 21-24.

Chafe, W. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing.* Chicago, University of Chicago Press.

Cruttenden, A. 1986. *Intonation*. Cambridge: Cambridge University Press.

Cruttenden, A. 1994. Rises in English. In J. Windsor Lewis (Ed.) *Studies in General and English Phonetics: Essays in Honour of J. D. O'Connor.* London: Routledge.

Durand, P., Durand-Deska, A., Gubrynowicz, R. and Marek, B. 2002. Polish: Prosodic Aspects of "Czy" Questions. *Proceedings of the Speech Prosody 2002* Conference: 255-258.

Francuzik, K., Karpinski and M., Klesta, J. 2002. A Preliminary Study of the Polish Intonational Phrase, Nuclear Melody and Pauses in Polish Semi-Spontaneous Narration. *Proceedings of the Speech Prosody 2002 Conference:* 303-306.

Grabe, E. 2002. Variation adds to prosodic typology. In B.Bel and I. Marlin (Eds.), *Proceedings of the Speech Prosody 2002 Conference*: 127-132.

Grabe, E., Post, B. and Nolan, F. 2001. Modelling intonational Variation in English. The IViE system. In Puppel, S. and Demenko, G. (eds.), *Prosody 2000.* Adam Mickiewicz University, Poznan, Poland, 2001.

Greenberg, Joseph H. (ed.) (1963). *Universals of Language*. Cambridge, Mass.: MIT Press.

Gussenhoven, C. 2002. Intonation and Interpretation: Phonetics and Phonology. *Proceedings of the Speech Prosody 2002 Conference*: 47 – 58.

Haan, J. 2002. *Speaking of Questions*. Utrecht, LOT.

Haan, J. and V.J. van Heuven. 1999. Male vs. female pitch range in Dutch questions. *Proceedings of the 14 International Congress of Phonetic Sciences.* San Francisco: 1581-1584

Hadding-Koch, K. and Studdert-Kennedy, M. 1964. An Experimental Study of Some Intonational Contours. *Phonetics*, 11, pp. 175 – 185.

Karpinski, M. 2002. Polish Intonational Database PoInt: Project Report. Available (in Polish) from http://main.amu.edu.pl/~maciejk/idbhome.html.

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

Maddieson, I. 1999. In Search of Universals. In *Proceedings of ICPhS*, San Francisco 1999: 2521 – 2528.

O'Connor, J .D. and Arnold, G.F. 1973. *The Intonation of colloquial English.* London, Longman.

Ohala, J. 1983. Cross-language Use of Pitch: An Ethological view. *Phonetica* 40: 1 – 18.

Steffen-Batogowa, M. 1996. *Struktura przebiegu melodii polskiego jezyka ogólnego.* Poznan: Instytut Lingwistyki UAM.

Wennerstrom, A. 2001. *The Music of Everyday Speech.* Oxford, Oxford University Press.

Inductive learning of lexical semantics with typed unification grammars

DIMITAR KAZAKOV AND SIMON DOBNIK*

Abstract

In the last decade machine learning techniques based on logic such as Inductive Logic Programming (ILP) have started being used in learning grammars from corpora. While the first approaches were based on the translation of grammar into first-order predicate logic, an attempt has been made recently to adapt the ILP learning schema to the feature constraint logic of typed-unification grammars. In this framework, the learner applies in turn generalisation and specialisation to the typed feature structures representing the grammar in order to improve its coverage of the training corpus provided. In this paper we demonstrate how the lexical resource WordNet could be incorporated into an existing grammar learning tool and show how lexical semantic constraints could be learned on the basis of the ontology information.

1 Introduction

In the last decade machine learning techniques based on logic such as Inductive Logic Programming (ILP) (Muggleton and De Raedt, 1994) have started being used in learning grammars from corpora (Cussens and Džeroski, 2000). Ciortuz (2002) implements and documents a system known as *ilp*LIGHT which combines a logic-based learning approach with *typed feature structure unification grammars* to modify the initial grammar provided and adjust its coverage to the training corpus.

Typed Feature Structure Grammars (TFSG) present both grammar rules and lexical items as symbols or attribute-value matrices (which are also called feature structures), and thus lexical semantic information can be represented in an identical

* The second author's work has been funded by the EPSRC ROPA grant *Machine Learning of Natural Language in a Computational Logic Framework.*

way, alongside the syntactic constraints. Furthermore, Feature Structures (FSs) are organised into hierarchies of *types* (represented as directed acyclic graphs) where each type inherits the feature specifications of those types higher in the hierarchy. This is a very concise and convenient way of representing both lexical entries and grammar rules.

The present work complements and extends the work of Ciortuz (2002) by outlining a technique of incorporating lexical semantic information within the existing typed feature structure grammar and describes how to use his learning tool *ilp*LIGHT to induce lexical semantic generalisations for verbal predicates. The approach demonstrates that TFSGs are well suited for incorporating lexical information available in online ontologies such as WordNet, since there the semantic information is organised in hierarchical structures that could be successfully incorporated within a FS hierarchy and referenced by the grammar rules. The approach also discusses the possibilities of reducing the ambiguity of grammars as additional lexical semantic constraints will make the parsing process (more) deterministic. While the present stage of research only offers a descriptive account, the future work will concentrate on its practical implementation within the *ilp*LIGHT system.

The article proceeds as follows: we start by giving a brief outline of the typed unification grammar formalism (Section 2) together with the details of the *ilp*LIGHT system (Section 3). We provide a short description of WordNet, a large-scale resource of lexical semantic information (Section 4), and outline our proposal for interfacing the two (Section 6).

2 Typed feature structure grammars

The TFSG is a logic framework first studied by Carpenter (1992) and Smolka (1992), which is widely used in computational linguistics. Several formalisms have been developed: Lexical Functional Grammar (LFG) (Kaplan and Bresnan, 1982), Categorial Unification Grammar (Uszkoreit, 1986), and Head-Driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994; Sag and Wasow, 1999). Finally, the Linguistic Knowledge Building (LKB) system described by Copestake (2002) is an open source TFSG development environment.

The linguistic objects (or *signs*) of TFSGs are represented as typed attribute value matrices (AVMs) or feature structures (FSs). For example, the simple

grammar included with the *ilp*LIGHT system includes the following information for the lexical item *girl* specified as a matrix of *attributes* or *features* and their values.[1]

```
girl:phrase_or_word
  [ PHON    <!"girl"!>,
    CAT     noun,
    SUBCAT <det> ]
```

One of the most important properties of the formalism is that it allows context-free grammar rules to be represented in an identical way: as *constraints* inside FSs in the form of indices (or *tags*) which indicate structure sharing. For example, *ilp*LIGHT has been used with a FS which encodes the following HPSG rules: the Head Feature Principle, the Subcategorisation Principle, and the Saturation Principle. The Head Feature Principle specifies the category of the phrase be token-identical (thus not simply share the same feature specifications) with the category of the head. The Subcategorisation Principles specifies the first member of the SUBCAT list to be identical with the category of the complement phrase. Finally, the Saturation Principle specifies that the value of COMP . SUBCAT is nil.

```
satisfy_HPSG_principles
  [ PHON    diff_list,
    CAT     #1:categ,
    SUBCAT #2:categ_list,
    HEAD    #4:phrase_or_word
     [CAT     #1
      SUBCAT #3|#2 ],
    COMP    #5:phrase_or_word
     [CAT     #3,
      SUBCAT nil ],
    ARGS    <#4, #5> ]
```

The FSs are *typed* or *sorted* (the types of the above feature structures are *phrase_or_word* and *satisfy_HPSG_principles* respectively) by the grammar which means that the grammar contains specifications or *appropriateness conditions* about which features are appropriate for each typed feature structure or *type/sort* and what are the possible values of these features.

Types can be atomic or non-atomic. A non-atomic type is one that has more specific instances. These are represented in a hierarchical relation to a more general non-atomic type, as a directed acyclic graph. The grammar used by

[1] The following examples are taken from Ciortuz (2001a).

*ilp*LIGHT is based on the sort hierarchy in Figure 1 (the origins of which can be traced back to Shieber (1986)), and is a simplified version of the HPSG sort hierarchy found in Sag and Wasow's book (1999), p.386.

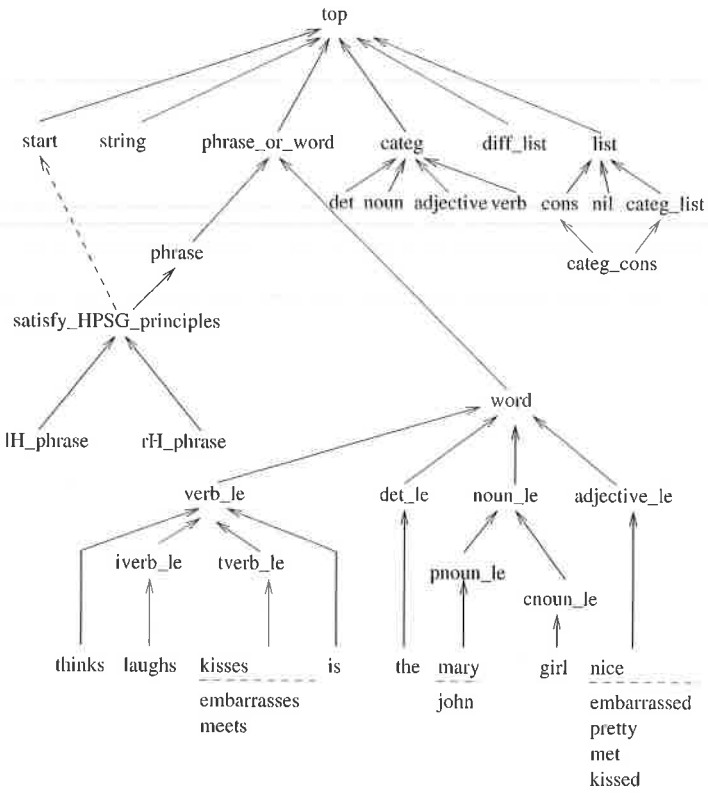


Fig. 1. A sample type/sort hierarchy

A more specific type contains more information in terms of the number and the value of features and thus constraints that apply for more than one type need only to be stated once for a parent type and are then inherited by its descendants. Stating this differently: the types higher in the type hierarchy *subsume* those below. This property enables the grammar to be constructed in a very concise way (see Figure 3). For example, the sample type hierarchy specifies that the type

lH_phrase is more specific than *satisfy_HPSG_principles*. It adds the following constraint:[2]

```
lH_phrase
[ PHON       #1!#3,
  HEAD.PHON  #1!#2,
  COMP.PHON  #2!#3 ]
```

Thus the knowledge specified in the type *lH_phrase* is as follows:

```
lH_phrase
[ PHON       #6!#8,
  CAT        #1:categ,
  SUBCAT     #2:categ_list,
  HEAD       #4:phrase_or_word
               [ PHON   #6!#7,
                 CAT    #1,
                 SUBCAT #3|#2 ],
  COMP       #5:phrase_or_word
               [ PHON   #7!#8,
                 CAT    #3,
                 SUBCAT nil ],
  ARGS       <#4, #5> ]
```

Additional types present in the grammar allow further expansion of the *lH_phrase* type until all the relevant features specified in the type hierarchy are present and their values are maximal or atomic types. Such feature structures are said to be *totally well typed* and *sort resolved* and as such are representations of linguistic signs. On the other hand, types can be partial and thus relating to a series of linguistic signs.

Finally, the grammar must allow two constraints (typed FSs) to be combined or *unified*. For example, the COMP feature in the *lH_phrase* must combine with the type *phrase_or_word* while satisfying certain constraints regarding the values of features CAT and SUBCAT. A unification of two feature structures F and F' is the greatest lower bound of F and F' in the collection of feature structures ordered by subsumption (Copestake, 2000). Thus, the result of the unification is the most general typed feature structure that is compatible with both input F and F' which is identical to the logical conjunction $F \wedge F'$. If no greatest lower bound of F and F'

[2] The ! notation is used to represent difference lists. The constraint specifies that the head of the phrase is its leftmost element. *rH_phrase* specifies the opposite and is required to represent subjects as complements of right-headed phrases.

exists, i.e. the information specified in both feature structures is incompatible, the unification fails.

The *ilp*LIGHT system uses a somewhat different logic than typed feature structure theory (Carpenter, 1992; Copestake, 2000; Copestake, 2002). Instead, it is based on Order-Sorted Feature (OSF) logic common in Constraint Logic Programming (CLP) and elaborated by Aït-Kaci *et al.* (1994). These authors define the notion of OSF-theory unification on *order consistent* OSF-theories as more general than well-typed feature structure unification. Ciortuz (2001a; 2001b) complements the OSF notion of *order consistency* with *type consistency* and requires both conditions to be satisfied by the OSF theories used in his tool. The reader is referred to Ciortuz (2001b) for a detailed discussion on the scopes of each of these approaches and the benefits they bring in.

3 Inductive learning of typed unification grammars with *ilp*LIGHT

Inductive logic programming (ILP) (Muggleton and De Raedt, 1994) is a machine learning approach where the goal is to induce hypotheses starting from sample data (examples of one or more concepts) and background knowledge. In its most common form, ILP uses the representation formalism of logic programming, i.e., a subset of first-order logic.

Given two sets of positive examples E^+ and negative examples E^- and a background theory H, the following conditions have to hold for the hypothesis H:

$$\textit{Prior Satisfiability} \quad B \wedge E^- \not\models \square \tag{1}$$

$$\textit{Posterior Satisfiability} \quad B \wedge H \wedge E^- \not\models \square \tag{2}$$

$$\textit{Prior Necessity} \quad B \not\models E^+ \tag{3}$$

$$\textit{Posterior Sufficiency} \quad B \wedge H \models E^+ \tag{4}$$

When 4 holds, H is said to be *complete*, and when 2 holds, *consistent*.

The *ilp*LIGHT system (Ciortuz, 2002) consists of three components: the Expander, the ABC LIGHT[3] parser, which is a combination of a head corner parser

[3] ABC stands for Active Bottom-up Chart-based parser. LIGHT stands for Logic, Inheritance, Grammars, Heads and Types. It was developed at the Language Technology Lab of the German Research Centre for Artificial Intelligence (DFKI), Saarbrücken.

(ABC), LIGHT, the OSF abstract machine for feature structure unification, and the *ilp* Learner component.

Note that the OSF formalism used by the ILP Learner to represent the grammar is different from first order logic, traditionally used in ILP. In *ilp*LIGHT the inductive learning methods are adapted to typed unification grammars which are represented in the feature constraint logic. The architecture of the system is summarised in Figure 2.

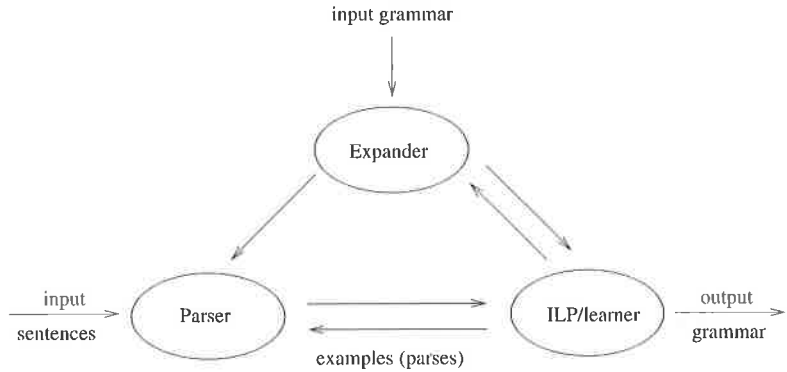


Fig. 2. The *ilp*LIGHT system for learning typed feature structure grammars

```
cons
[ FIRST top,
  REST  list ]

diff_list
[ FIRST_LIST list,
  REST_LIST  list ]

categ_cons
[ FIRST categ,
  REST categ_list ]

phrase_or_word
[ PHON   diff_list
  CAT    categ,
  SUBCAT categ_list ]

phrase
[ HEAD #1:phrase_or_word,
  COMP #2:phrase_or_word,
  ARGS <#1, #2> ]

satisfy_HPSG_principles
[CAT    #1,
 SUBCAT #2,
 HEAD   top
        [ CAT    #1,
          SUBCAT #3|#2 ],
 COMP   top
        [ CAT    #3,
          SUBCAT nil ] ]

det_le
[ CAT    det,
  SUBCAT nil ]

noun_le
[ CAT noun ]

pnoun_le
[ SUBCAT nil ]

cnoun_le
 [ SUBCAT <det> ]

adjective_le
[ CAT    adjective,
  SUBCAT nil ]

iverb_le
 [ CAT    verb,
   SUBCAT <noun> ]

tverb_le
 [ CAT    verb,
   SUBCAT <noun, noun> ]

lH_phrase
 [ PHON      #1!#3,
   HEAD.PHON #1!#2,
   COMP.PHON #2!#3 ]

rH_phrase
 [ PHON      #1!#3,
   HEAD.PHON #2!#3,
   COMP.PHON #1!#2 ]

the [ PHON <!"the"!> ]
girl [ PHON <!"girl"!> ]
john [ PHON <!"john"!> ]
mary [ PHON <!"mary"!> ]
nice [ PHON <!"nice"!> ]
fed [ PHON<!"fed"!> ]
pretty [ PHON <!"pretty"!> ]
met [ PHON <!"met"!> ]
kissed [ PHON <!"kissed!"> ]
is [ PHON   <!"is"!>,
     CAT    verb,
     SUBCAT <adjective, noun> ]
laughs [ PHON <!"laughs"!> ]
kisses [ PHON <!"kisses"!> ]
thinks [ PHON   <!"thinks"!>,
           CAT    verb,
           SUBCAT <verb, noun> ]
meets [ PHON <!"meets"!> ]
feeds [ PHON <!"feeds"!> ]
```

Fig. 3. A sample typed feature structure grammar

The framework can be used to process large-scale HPSG grammars such as LinGO (Copestake, Flickinger and Sag, 1999) developed by the Center for Studies of Language and Information (CSLI) at the University of Stanford.[4] For the purposes of this article we will only be concerned with a small subset of this grammar described by Ciortuz (2002) which is given in Figure 3. The type hierarchy is as in Figure 1.

The grammar is fed to the Expander which expands the types with the features inherited from their ancestors in the type hierarchy. The expanded types are sent to the Parser which uses them to analyse the input sentences. Since the grammar may be incomplete, parsing may produce partial parses which are evaluated by the ILP learner module. This module suggests improved grammar rules that provide a better coverage of the input sentences, that is, to generate or succeed only on positive examples from the set of input sentences and to exclude the negative examples. Alternatively, the learning data may consist of a set of sentences which are only marked for the number of possible parses or of pairs of sentences and parses.

The parser and the expander operate bidirectionally as shown by the arrows. For example, the parser can take as input a parse[5] and then attempt to build a feature structure associated with that parse. If the process fails the grammar is incorrect and the information on failure is passed back to the learner in order to augment the grammar to accept the parse. The expander, on the other hand, can accept as an input an expanded type and a feature path inside this type and returns information on the unexpanded type which introduced that feature path/constraint.

Learning in the *ilp*LIGHT set up is implemented as search through the space of possible generalisations and specialisations of one of the elements of the OSF theory encoding the grammar. Generalisation can be achieved through (1) replacing a sort (unexpanded type) with a more general one, (2) removing an equation unifying two variables (indices) in the OSF type, or (3) removing a feature from that type. Specialisation is based on four refinement operators, which are to a large extent complementary to the generalisation steps. One can (1) replace a sort with another, subsumed by the former, (2) introduce a new equation unifying two indices, or (3) unfold a sort, i.e., replace it with the corresponding

[4] Online resources are available at http://lingo.stanford.edu.

[5] Complete parses are of the form
`(rH_phrase 0 2 (john 0 1 ("john" 0 1)) (laughs 1 2 ("laughs" 1 2))).`

OSF type and propagate the relevant constraints through the root features of this type. The remaining, fourth refinement operator merges two independently produced refinements of the same term if they are unifiable. In the present implementation, the generalisation step considers all elements of the grammar in turn, whereas the specialisation step has to be provided with the grammar item that is to be refined (Ciortuz, 2002).

4 WordNet

WordNet is an on-line lexical database which contains various syntactic and semantic information for a large number of words and idioms. Originally developed for English (Miller *et al.*, 1993), WordNet is also implemented for a number of other languages as EuroWordNet, BalkanWordNet, etc.[6] The central building element of WordNet is called a *synset*, or *lexicographer's entry*. A synset is a set of words or idioms which share a common meaning. For instance: {*(to) shut, (to) close*}. To simplify the internal representation, each synset is assigned a large integer used as a unique identifier. For instance, {*monetary resources, funds*} is Synset 109616555 in WordNet1.6. WordNet uses a set of rules and lists of exceptions to map word-forms to all relevant lexical entries. Figure 4 shows the word-form 'funds' which is recognised by WordNet as corresponding to two lexical entries, 'fund' and 'funds'. The lexical entry 'fund' appears in three synsets: {*store, fund*}, {*fund, monetary fund*}, and {*investment company, fund*}, respectively. The lexical entry 'funds' only appears in the synset {*monetary resource, funds*}. WordNet describes several semantic relations between synsets, such as meronymy (part-of), hypernymy or hyponymy. The latter are shown in Figure 4.

[6] Information on these resources is available at http://www.globalwordnet.org.

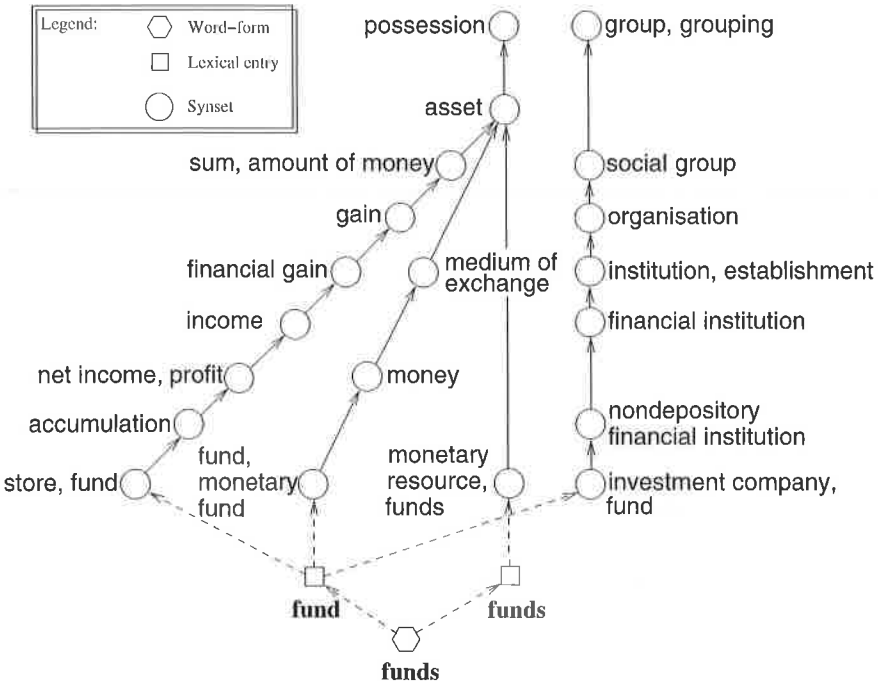


Fig. 4. Mapping from word-forms and lexical entries to synsets and their hypernyms in WordNet

5 Related Work

There are a number of related approaches where grammars or the equivalent parsers have been automatically modified to best represent the properties of a corpus.

Learning Fast LR Parsers Samuelsson (1994) describes a technique for the development of an efficient LR parser based on Explanation-Based Learning (EBL) (Mitchell, 1997) and entropy-related information measures. The method is based on partial lexicalisation of grammar rules and expansion of RHS nonterminals. The new rules do not cover parts of the grammar, which are only marginally represented in the treebank. As a result, the grammar is less ambiguous

at the price of a certain loss of coverage. Using that grammar also results in considerably faster parsing.

Zelle and Mooney (1993) have developed a method for learning semantic grammars from a treebank containing syntactic trees with semantically tagged nonterminal nodes. The treebank is used to construct an over-general shift-reduce parser covering the sentences in the treebank. The parser is then specialised and made deterministic by using ILP. When the semantic tags assigned to nonterminals are not sufficient to remove the ambiguity from the grammar, lexical semantic classes are automatically defined in order to achieve deterministic parsing.

LAPIS (Kazakov, 1999) is a system which builds on Zelle and Mooney's research on the induction of shift-reduce parsers and extends it to learning LR parsers, while changing at the same time the focus of the learning task. The existence of sources of lexical semantic knowledge such as WordNet (Miller *et al.*, 1993) makes the learning of lexical semantic classes done by Zelle and Mooney less attractive. Also, treebanks annotated with phrasal semantic tags are not commonly available. Instead, the system LAPIS constructs LR parsers from treebanks annotated with lexical semantic tags. LAPIS aims at the reduction of nondeterminism in the parsers it creates by the means of lexicalisation and partial unfolding of the underlying grammar rules, in combination with the use of lexical semantic constraints.

Cussens and Pulman (2000) combine ILP with a chart parser to find the missing rules in a grammar that would allow parsing all training sentences provided. When a sentence cannot be parsed, the parser suggests 'needed edges', i.e., those necessary to complete a parse. Examples of these are stored and used to learn a more general pattern corresponding to a grammar rule; additional linguistic constraints may be provided to prune the search for new grammar rules and ensure their plausibility.

6 Learning lexical semantics

This section describes methods of incorporating lexical semantic information from WordNet into type hierarchy and grammar rules.

6.1 Extending the type hierarchy

A type hierarchy is a finite bounded complete partial order (Copestake, 2000). This is also the case for the WordNet hierarchy of synsets, if a dummy bottom ($\perp$) synset that is subsumed by all others is added. Figure 5 represents a simplified WordNet hierarchy for some common and proper nouns.

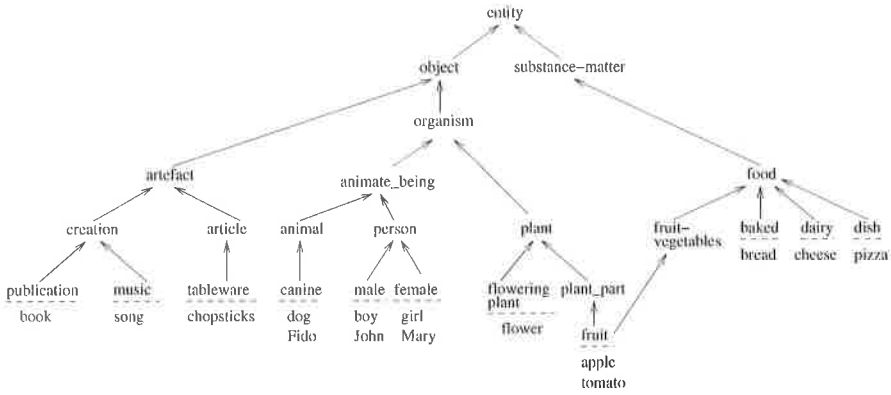


Fig. 5. A simplified WordNet hierarchy of synsets

Remember that WordNet distinguishes between words and lexical entries. Words are strictly not a part of the hierarchy but there exists a function which maps them to the hierarchy members. To distinguish the two we represent words below a dashed line.

The type hierarchy in Figure 1 and the WordNet hierarchy have to be merged so that the types are properly expanded and the lexical information is properly propagated within the same framework.

The lexical information in Figure 5 further restricts the categorial information for a given class of grammatical objects, nouns in our case. For example, most generally, the lexical item 'book' belongs to the category of *nouns*, but also to a number of its lexical sub-categories: it is an *entity*, an *object*, an *artefact*, and a *creation*. These semantic classes are gradual refinements of the two extremes: the grammatical category and the lexical item itself. Thus, the simplest solution to incorporate the lexical hierarchy into the existing type hierarchy is to embed the lexical *top* node *entity* in Figure 5 underneath the *noun* node of Figure 1. The *noun*

node and thus all its daughters are now subsumed by the *categ* node (which is subsumed by the *top* node). This means that the system will now not only allow to infer that a certain lexical item is a noun but also that it is a 'creation-kind-of-noun'.

6.2 Adding lexical semantic types to the grammar

The *categ* type and the types that it subsumes do not introduce any new features that would interact with the existing grammatical constraints. Examining the sample grammar in Figure 3 the categorial information on the noun *word*s is introduced in the type *noun_le*. However, restricting this category with a more specific subtype such as *organism* is of no use since this would mean that our grammar would only be able to deal with 'organism' noun phrases. In our grammar, lexical information is as specific as phonological information and thus must be restricted at the same level (which is also the lowest level). Thus, the 'lexical types' in Figure 3 will be extended to the following unexpanded representation:

```
girl [ PHON <!"girl"!>,
       CAT  female ]

flower [ PHON <!"flower!"!>,
         CAT  flowering-plant ]
```

Note that now the specification of the feature `CAT` in the type *noun_le* is not necessary and can be removed. Its presence does not affect type expansion since a unification of `CAT noun` and `CAT flowering-plant` will always result in `CAT flowering-plant`.

6.3 Lexically enriched grammar: what can be done with it?

We shall discuss some of the possible benefits of introducing a lexically enriched grammar as shown in the examples below.

Learning the semantic restrictions of verbal predicates Allowing a sufficiently large set of positive and negative data and a sufficiently fine-grained hierarchy of lexical relations it is theoretically possible to use *ilp*LIGHT to learn the lexical semantic restrictions of individual verbal predicates. Presently, the verbal

predicates of the grammar in Figure 3 are only restricted in the most general way: they are specified for categorial information:

```
kisses [ PHON <!"kisses"!>,
         CAT  verb,
         SUBCAT <noun, noun> ]
```

Let our training corpus for *ilp*LIGHT consist of the following sentences.

John kissed Mary.
Fido kissed John.
The girl kissed the boy.
**Mary kissed the flower.*
**The book kissed the castle.*
**Mary kissed the book.*

It is possible for the learner component of *ilp*LIGHT to find that the verbal predicate *kisses* requires both arguments to belong to the lexical class *animate being* through a number of specialisation steps.

```
kisses
    [ PHON <!"kisses"!>,
      CAT verb,
      SUBCAT <animate_being, animate_being> ]
```

Lexical semantic ambiguity Words that correspond to more than one meaning (synset) can be handled in this framework by adding separate lexical entries for each of their meanings to the grammar. Similarly, separate entries may have to be learned to cover the various semantic roles of a word where the grammar originally contained one lexical entry only specifying syntactic categorial information. To our knowledge, *ilp*LIGHT cannot handle this without a further modification.

Reducing ambiguity in parsing Lexically enriched grammars can be used to reduce non-determinism in the output of parsers. The grammar in Figure 3 does not distinguish between *complements* of verb and noun phrases and *adjuncts* which here are optional phrases modifying the verb phrase as a whole. The following is a typical set of sentences that introduces parsing ambiguity to any grammar which distinguishes the two.

John eats pizza with cheese.
John eats pizza with chopsticks.

The prepositional phrase *with cheese* is a complement of the noun *pizza* (it further specifies the pizza)[7] and thus should be analysed as its argument. On the other hand, *with chopsticks* is an adjunct of the entire verb phrase *eats pizza*. Because both phrases belong to the same grammatical category the parser would give two parses for each sentence, but only one of which is correct. The choice of the parse depends on the lexical semantics of the complement noun inside the prepositional phrase and important generalisations can be made to reduce the ambiguity. If the noun belongs to the lexical semantic class *food* the entire prepositional phrase is unambiguously analysed as the complement of the noun *pizza* which also belongs to this lexical category (or is subsumed by this type), otherwise the prepositional phrase is an adjunct of the verb phrase. In this case it is an *instrument* further specifying the pizza eating event.

Before setting up a learning task, the grammar in Figure 3 needs a modification since it does not recognise prepositional phrases, nor does it distinguish between complements and adjuncts.

Adding a new grammatical category is straightforward: a new node/type such as *prep* can be added to the type hierarchy so that it is subsumed by the *categ* node. The category of prepositions takes a noun phrase as its complement. We add this constraint in the form of a *prep_le* type subsumed by the *word* node.

```
prep_le
  [ CAT    prep,
    SUBCAT <noun> ]
```

To accommodate noun phrases with prepositional complements we create a new type *cnoun-compl_le* which is subsumed by the *noun_le* node. The *cnoun-compl_le* adds the following constraints:

```
cnoun-comp_le
 [ SUBCAT <prep> ]
```

We treat prepositional adjuncts of verbs as parts of their subcategorisation frame.[8] We add another type *tverb-pp_le* which is subsumed by the *verb_le* node and is specified for the following features:

[7] Notionally there is a very strong distinction between complements and adjuncts, yet no formal definition exists.

[8] This is the standard HPSG treatment (Sag and Wasow, 1999).

```
tverb-pp_le
 [ CAT verb,
   SUBCAT <noun, prep, noun> ]
```

The learning task here is significantly more complex than the previous one since here the learner is trying to specialise phrasal templates rather than words. For example, to learn that a prepositional phrase is a complement of the noun rather than the verb phrase (in which case it is an adjunct), a relation must be established between the head noun (*pizza*) and the complement (*cheese*) of its complement (*with*) which leads to a specialisation of the *categ* value of *pizza* and *cheese* from *noun* to *food.* In the present design of the grammar, which distinguishes between words and phrases, words only contain information stated in the SUBCAT feature on the category of their *immediate* complements but not their internal structures. For example, an expanded type that constrains nouns that take prepositional complements *cnoun-comp_le* is as follows:

```
cnoun-comp_le
 [ PHON   diff_list,
   CAT    noun,
   SUBCAT <prep> ]
```

Therefore, the specialisation must be done at the level of expanded phrases *lH_phrase* and *rH_phrase* where *cnoun-comp_le* is the head of the phrase.[9] The learner should look for phrasal templates such as the following:

```
food-with-food
 [ PHON      #8!#10,
   CAT       #1:food,
   SUBCAT    #2:nil,
   HEAD      #4:cnoun-comp_le
               [ PHON   #8!#9,
                 CAT    #1:food,
                 SUBCAT #3:with|#2:nil ],
   COMP      #5:prep_le
               [ PHON   #9!#10,
                 CAT    #3:with,
                 SUBCAT nil
                 COMP   #6:cnoun_le,
                          [ CAT #7:food ]],
```

[9] Another apparent solution would be to merge the types *phrase_or_word* and *phrase* into a single type. This would result in a grammar which only accounts for phrases which is incorrect. Nouns such as *cheese* in *pizza with cheese* would be left unaccounted for.

```
   ARGS       <#4, #5> ]
```

The template is a specialisation of the *lH_phrase*. It adds the following constraints to the expanded *lH_phrase*:

```
food-with-food
 [ HEAD cnoun-comp_le
        [ CAT #1:food ],
   COMP prep_le
        [ CAT  with,
          COMP cnoun_le,
           [ CAT #1:food ]]
```

It follows that this new specialised type, a partially unfolded and lexicalised template, is added to the type hierarchy so that it is subsumed by the *lH_phrase*.

A similar specialisation is needed for the case where the prepositional phrase is an adjunct.

```
verb-food-with-tableware
 [ PHON    #11!#13,
   CAT     #1:verb,
   SUBCAT  #2:noun,
   HEAD    #4:tverb-pp_le
            [ PHON    #11!#12,
              CAT     #1:verb,
              SUBCAT  #3:prep|#2:noun,
              HEAD    #6: tverb-pp_le
                        [ CAT #1:verb,
                          SUBCAT #7:food|#8:prep,noun,
                          COMP #9:cnoun_le,
                           [ CAT #7:food ]],
   COMP   #5:prep_le
           [ PHON #12!#13,
             CAT  #3:with,
             COMP #10:cnoun_le
                   [ CAT tableware ]]]
   ARGS   <#4,#5>
```

The previous examples show two implications of this approach for the learning procedures. Firstly, due to an increased level of depth at which the search operates (`HEAD.HEAD.COMP.CAT food`) the search space or the time to find a solution will be considerably increased. Secondly, learning does not only specialise or generalise over the existing types but creates new types which must be added to the type hierarchy.

7 Discussion and Future Work

We have described how WordNet, a standard resource of lexical semantic information, could be incorporated in a TFSG without any extension of the formalism used, and how an existing inductive grammar learner may use this information to add semantic constraints to a grammar in order to optimise its coverage of a training corpus. Further work should focus on implementing the approach outlined here in *ilp*LIGHT and testing its performance, and independently, on extending the initial grammar to include additional linguistic phenomena. In a study combining these two dimensions, the issue of pruning the search space of the ILP learner will become central. Here the research should concentrate on a systematic enumeration of the range of possible modifications of the grammar and the elimination of those constraints that are linguistically implausible, cf. (Cussens and Pulman, 2000).

References

Aït-Kaci, Hassan, Andreas Podelski, and Seth Copen Goldstein. 1994. Order-sorted feature theory unification. *The journal of logic programming*, 19(20):1–25.

Carpenter, Bob. 1992. *The logic of typed feature structures: with applications to unification grammars, logic programs and constraint resolution.* Number 32 in Cambridge tracts in theoretical computer science. Cambridge University Press.

Ciortuz, Liviu. 2001a. Expanding feature-based constraint grammars: Experience on a large-scale HPSG grammar for English. In *Proc. of the Workshop on modelling and solving problems with constraints*, Seattle, USA. http://www.lirmm.fr/~bessiere/proc_wsijcai01.html.

Ciortuz, Liviu. 2001b. Light AM — another abstract machine for FS unification. In Stephan Oepen, Daniel Flickinger, Jun-Chi Tsujii, and Hans Uszkoreit, editors, *Efficiency in Unification-Based Processing*. CSLI Publications, Stanford, pages 1–27.

Ciortuz, Liviu. 2002. Towards inductive learning of typed-unification grammars. In the (electronic) *Proceedings of the 17th Workshop on Logic Programming*. Dresden Technical University, Germany, 11–13 December.

Copestake, Ann. 2000. Definitions of typed feature structures. *Natural Language Engineering*, 1(6). Appendix to special issue on efficient processing with HPSG.

Copestake, Ann. 2002. *Implementing typed feature structure grammar.* Number 110 in CSLI Lecture notes. CSLI Publications, Stanford.

Copestake, Ann, Daniel Flickinger, and Ivan Sag. 1999. *A grammar of English in HPSG: Design and implementation.* CSLI Publications, Stanford.

Cussens, James and Sašo Džeroski, editors. 2000. *Learning Language in Logic.* Lecture Notes in Artificial Intelligence. Springer-Verlag.

Cussens, James and Stephen Pulman. 2000. Experiments in inductive chart parsing. In James Cussens and Sašo Džeroski , editors, *Learning Language in Logic*, Lecture Notes in Artificial Intelligence. Springer-Verlag.

Kaplan, Ronald M. and Joan Bresnan. 1982. Lexical-functional grammar: A formal system for grammatical representation. In Joan Bresnan, editor, *The mental representation of grammatical relations*. Cambridge, Mass.: MIT Press, pages 173–381.

Kazakov, Dimitar. 1999. Combining LAPIS and WordNet for the learning of LR parsers with optimal semantic constraints. In Sašo Džeroski and Peter Flach, editors, *9th International Workshop on Inductive Logic Programming ILP-99*, number 1634 in Lecture Notes in Artificial Inteligence, pages 140–151, Bled, Slovenia. Springer-Verlag.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1993. Introduction to WordNet: An on-line lexical database. Technical report, University of Princeton. ftp://ftp.cogsci.princeton.edu/wordnet.

Mitchell, Tom M. 1997. *Machine Learning*. McGraw-Hill.

Muggleton, S. and L. De Raedt. 1994. Inductive logic programming. Theory and methods. *Journal of Logic Programming*, 19(20):629–679.

Pollard, Carl and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: Chicago University Press.

Sag, Ivan A. and Tom Wasow. 1999. *Syntactic theory: a formal introduction.* Number 92 in CSLI Lecture notes. CSLI Publications, Stanford.

Samuelsson, Ch. 1994. *Fast Natural-Language Parsing Using Explanation-Based Learning*. Ph.D. thesis, The Royal Institute of Technology and Stockholm University.

Shieber, Stuart M. 1986. *An introduction to unification-based approaches to grammar*. Number 4 in CSLI Lecture notes. CSLI Publications, Stanford.

Smolka, Gert. 1992. Feature-constraint logics for unification grammars. *Journal of Logic Programming*, 12:51–87.

Uszkoreit, Hans. 1986. Categorial unification grammar. In *International conference on computational linguistics (COLING-92)*, pages 440–446, Nantes, France.

Zelle, John M. and Raymond J. Mooney. 1993. Learning semantic grammars with constructive inductive logic programming. In *Proceedings of AAAI-93*, pages 817–822. AAI Press/MIT Press.

Deriving a domain theory for disambiguation purposes

MARIA LIAKATA

Abstract

A domain is understood as a homogeneous corpus whose texts are related by a common theme. A set of rules describing relations between entities in the domain constitutes a domain theory. Such a set of rules can be used to aid disambiguation tasks (word sense disambiguation, pronoun resolution, the resolution of prepositional phrase attachment). A mechanism for deriving a domain theory automatically from the Wall Street Journal (WSJ) section of the Penn Treebank is explored and some preliminary results are presented. The method used for learning the theory is a combination of Inductive Logic Programming (ILP) and knowledge mining techniques.

Motivation

The inspiration for this work stems from a personal interest in linguistic ambiguity and the well-acknowledged need for disambiguation at various levels of linguistic analysis. Consider the classic example

I saw the man with the telescope.

Here there is ambiguity as to whether the prepositional phrase (PP) should be attached to the verb *saw* or its object *man*. Knowledge of the fact that a telescope is an optical instrument may increase the likelihood of the PP being assigned to the verb rather than the object but ambiguity still remains. This is a problem for natural language systems with no knowledge of the world. Matters get even more complicated when one thinks of lexical ambiguity or anaphora resolution.

John was looking for his toy box. The box was in the pen.
The police refused the students permission to demonstrate because they advocated/feared violence.

The word *pen* has two meanings, *writing implement* and *enclosure*. In the first example, one requires a mechanism for ruling out the *writing implement* sense. In the second example, the pronoun *they* can refer either to the police or to the students and the correct choice depends on the verb (*advocated/feared)* of which *they* is the subject. The sentence as a whole would not make sense if the police were to advocate violence.

Statistical techniques for disambiguation

One way of dealing with ambiguity is with a variant on the approach followed by most statistical systems. This involves training a system on a disambiguated corpus in order to extract *features* of good or bad analyses. These features are usually *n*-tuples of words or grammatical relations. In the case of PP attachment, where the issue is whether the PP should attach to the verb or to its object, the relevant features upon which the final choice is conditioned are the values of V, N_1, P and N_2, where V is the verb, N_1 the head of the object, P is the preposition of the PP in question and N_2 its head noun (Brooks and Collins, 1995). The probability of adverbial or adjectival attachment, $P(A = adv/adj \mid V = v,\ N_1 = n_1,\ P = p,\ N_2 = n_2)$ is estimated based on these four values. Given the earlier example

I saw the man with the bicycle.

it is possible to compare the two attachment probabilities P(*see* + *with bicycle*) < P(*man* + *with bicycle*) and arrive at the correct conclusion.

The number of features considered depends on the model — in the back-off model of Brooks and Collins (1995), the final choice can end up being made with reference to the preposition P alone. Other systems use co-occurrence frequencies originating from unambiguous PP attachments to provide clues for disambiguation (Ratnaparkhi, 1998). Pantel and Lin (2000) use contextually similar words to compute average attachment scores as well as a linear combination of features (such as the 4-tuple above) and finally combine the two to make a disambiguation decision.

In the case of word sense disambiguation, similarity based measures are used to resolve ambiguous meanings of words, depending on their context. The idea is that conceptually similar words are to be combined with the same or conceptually similar arguments, prepositions etc. (Dagan, Lee and Pereira, 1997).

An alternative approach: Domain Theories

Irrespective of the methodology followed, there are several disadvantages in relying on statistical methods for disambiguation purposes. For a start, statistical approaches can only model what is explicit in the data, without allowing the possibility for further inference. For example, consider the sentence

I observed the man with the binoculars.

If there is no instance of *observe with binoculars* in the corpus, the statistical mechanism will have to be fairly sophisticated to first note the similarity between the verbs *observe* and *see* and the nouns *telescope* and *binoculars* and to then base its decision on the observed probability of *see with telescope*. Even such a sophisticated system would be unable to deal with a case like

John said Tom was made redundant. He will receive his compensation tomorrow.

Here there is ambiguity as to whether the pronoun *He* refers to John or to Tom. To adequately resolve this, we require a rule that associates the subject of the verb phrase *make redundant* with the subject of the verb phrase *receive compensation*. Such an association is too long-range for the window of words considered by a statistical system.

Consider the rule

$$make\_redundant(Event1, X1, Y1) \Leftrightarrow receive(Event2, Y1, Y2) \wedge compensation(Y2) \wedge from(Event2, X1)$$

This states that if $Y1$ is made redundant by $X1$ then $Y1$ receives compensation from $X1$. The application of such a rule would allow the correct resolution of the pronoun *He* in the example above to *Tom*. Most statistics-based systems (and for that matter most syntactic algorithms for pronoun resolution) would fail.

Another problem with statistical approaches for disambiguation is that they need to be trained on a large amount of data, which is not always available. Even for a corpus of four million words, such as the Penn Treebank, the data is not statistically significant for obtaining probability distributions, at least not without a certain amount of pre-processing. It is difficult to make statistical systems context-specific without exploding the amount of data required. Additionally, when the results produced by statistical systems are not satisfactory, tracing the problem may be problematic. On the other hand, rules are transparent and can easily be edited.

It is therefore desirable to have some deductive structure — a set of rules and a reasoning mechanism from which we can draw inferences. Consider the following definitions.

1. A *domain* is a corpus whose texts are semantically cohesive.
2. A *domain theory* is a set of rules, often axiom-like, that describe entities and relations between entities in a corpus characterized by a common theme.

Hobbs (1978) describes a system for semantic analysis which performs pronoun resolution as a by-product. This system assumes that all world knowledge is available in the form of axioms written in First Order Logic — it presupposes a global domain theory. Axioms are of the form

$$(\forall y)(\exists z)(p(y) \supset r(z, y))$$

for example,

$$(\forall y)(\exists z)(building(y) \supset roof(z, y))$$

From axioms such as this, one can draw both forward and backward inferences. In the first case, one knows $p(y)$ and infers something about $r(z, y)$, e.g. if y is a building ($building(y)$) then there is an entity z such that z is the roof of y ($roof(z, y)$). Backward inference involves using information from which $p(y)$ can be inferred, e.g. if z is a roof and we know the relationship $roof(z, y)$ then y is a building.

Further axioms may be derived using standard inference rules like resolution. Given, for example, the two axioms stating that every bank is a building and that every building has a roof,

$(\forall y)(bank(y) \supset \text{building}(y))$
$(\forall y)(\exists z)(building(y) \supset (roof(z, y)))$

we may infer a further axiom that every bank has a roof.

$(\forall y)(\exists z)(bank(y) \supset roof(z, y))$

The system performs semantic analysis by means of several *semantic operations* that make inferences selectively from the axioms. These semantic operations look for specific patterns in the input and use the domain theory axioms to extend the information at hand.

In (Hobbs *et al.*, 1993) the semantic analysis performed involves various types of disambiguation tasks, including reference resolution, interpretation of compound nominals, metonymy and resolution of other kinds of syntactic and lexical ambiguity. Again, a large knowledge base in the form of axioms is assumed and the goal is to resolve ambiguities as a consequence of proving the logical form of the input sentence.

The interest in the use of a domain theory for disambiguation purposes is thus not new and its usefulness has already been acknowledged. In the cases above, however, all axioms were obtained manually. Handcrafting a theory is difficult, time-consuming and impossible to replicate in a consistent way. It would be preferable to derive such a theory automatically. It also makes practical sense to aim at more domain-specific theories rather than a global one, since different associations and meanings can underlie the same terms in different thematic contexts. Moreover, the principles and techniques involved in the derivation of one domain theory can easily be recycled in the derivation of another.

Pilot experiment on the ATIS Corpus

Pulman (2000) describes a pilot experiment aiming at inducing domain-specific axioms to aid syntactic disambiguation. The approach is based on the assumption that the amount of world knowledge necessary to perform such a task can be

obtained from previously disambiguated sentences within the domain. The chosen domain was the ATIS (Air Travel Information Service) Corpus, because its content is simple while at the same time being representative of the kind of domains one would be interested in modelling.

The sentence type was further restricted by selecting a subset of 500 sentences, all of which contained the words *flight*, *fare* and *meal/breakfast/dinner*. The sentences were then parsed into quasi-logical forms (QLFs) — logical expressions that correspond to the surface form of a sentence and can easily be converted to first order logic. The QLF corresponding to the sentence *I would like a flight from Washington to Atlanta* is

$$\exists A.flight(A) \wedge from(A, washington) \wedge to(A, atlanta) \wedge like(e78, I, A)$$

A set of disambiguated sentences were used as training material. The corresponding QLFs were converted into first order logic and fed as input to a machine learning paradigm (Inductive Logic Programming) in order to induce general rules about the domain.

Rules for this domain are expressed as

$$meal(A) \wedge flight(B) \rightarrow on(A, B)$$
$$flight(A) \wedge airline(B) \rightarrow on(A, B)$$

(paraphrased as *a meal is on a flight* and *a flight is on an airline* respectively). Negative rules, outlining what does not hold in the domain, were also derived. For example,

$$from(A, B) \wedge to(A, B) \rightarrow false$$

(read as *it is impossible to have a flight that is both from and to the same place*).

The rules derived from the learning mechanism were then used to disambiguate unseen sentences for prepositional phrase attachment. Candidate parses were selected on the basis of which of the corresponding QLFs were consistent with the derived rules. The results from the experiment were encouraging, since out of the QLFs tested, only a few were misclassified as correct. This was because they contained data for which there was no negative evidence.

Scaling-up domain theory extraction

The idea of automatically inducing a set of First Order Logic (FOL) rules for disambiguation purposes is very appealing. Pulman's experiment showed that the extraction of such a theory is feasible for a restricted domain. However, it was only performed on a small scale and with a very restricted set of sentences. It would be desirable to scale the process up to a wider domain, where the type of sentences encountered are more varied in both structure and vocabulary and the resulting rules are more complex, and to establish that the derived theory is still useful for disambiguation tasks like PP attachment and pronoun resolution.

A more extensive domain is provided by the Wall Street Journal (WSJ) section of the Penn Treebank. Even though the WSJ may not constitute as clearly defined a domain as the ATIS corpus, it can still be regarded as a domain for *financial news*. The advantage of the WSJ over other more homogeneous corpora is that it contains the parse trees of each sentence. Additionally, it contains extra semantic information and standard Part of Speech (POS) tags, that can facilitate the construction of logical forms.

Constructing LFs from the Penn Treebank

The first step in producing a domain theory is translating the input text into logical form. QLFs can be viewed as the natural level of sentence representation, resulting from the application of compositional semantic rules to the sentence constituents. They are neutral with respect to quantifier scope, reference resolution and tense (Alshawi and Eijk, 1990) and do not take the context into account.

Extracting such QLFs should be easy for sentences in the Penn Treebank owing to the extra amount of information available. The process is far from trivial, however. Liakata and Pulman (2002) present a method for retrieving predicate-argument structures from the Penn Treebank. The system operates on a flattened, morphologically-enriched version of the corpus. The flattened representation allows access to all levels of the parse tree simultaneously and enables the detection of the main sentence constituents by means of simple template rules. A small number of rules apply to identify the head words of each constituent and these in turn fill in the constituent templates to build the logical forms representative of the predicate-argument structure. The evaluation of the logical

forms showed a success rate of approximately 90%. The output QLFs of this system were used as input to the learning paradigm to induce the domain theory.

The learning paradigm

Having obtained the QLFs, we must choose an appropriate mechanism for learning a first order theory. Pulman (2000) uses Inductive Logic Programming (ILP). ILP is a discipline used to learn theories (hypotheses) from examples in the form of logic programs. ILP has the advantage over classical machine learning techniques for the following reasons:

- ILP allows a more expressive knowledge representation formalism, more suitable than propositional logic (or attribute-value vectors) to many domains of expertise, including natural language.
- Background knowledge can be included to aid the induction task.

The following section provides some definitions from First Order Logic and Logic Programming, which are essential for the understanding of the basic principles of ILP. Most of the definitions below were taken from (Bergadano and Gunetti, 1996) and (Muggleton, 1995).

Definitions

- *Clause:* a disjunction of literals, e.g. $(\neg p_1 \vee \neg p_2 \vee \neg p_3 \vee q)$, also written as $(p_1, p_2, p_3 \rightarrow q)$ or $(q :\text{-} p_1, p_2, p_3)$.
- *Definite Clause:* a clause containing **exactly one** positive literal.
- *Horn Clause:* a clause that contains **at most** one positive literal.
- *Logic Program:* a set of Horn clauses.

A *well-formed formula (wff)* is a logical expression which, when its constituents are interpreted, yields a proposition (with truth value *true* or *false)*. Given a well-formed formula, W,

- the *Herbrand universe* of W is the set of all ground terms composed of function symbols found in W;

- the *Herbrand base* of W, $B(W)$ is the set of all ground atoms composed of predicate and function symbols found in W;
- a *Herbrand interpretation* I of W is a total function from ground atoms to $\{false, true\}$, whose domain is the Herbrand base of W.

Given a well-formed formula, W, and a Herbrand interpretation, I,

- W is true in I if for every variable v in W, the substitution for v of any term t in the Herbrand universe of W yields a well-formed formula which is true in I;
- I is a *Herbrand model* of W iff W is true in I.

Every logic program P has a unique least, or minimal, Herbrand model denoted by $M^{+}(P)$.

ILP basics

The purpose of ILP is to learn a set of general rules which concisely explain or predict the observed input data. The input is split into two logic programs — one (the *examples* or the *evidence*) representing the input data considered specific to the domain, the other (the *background*) representing information which holds independently of the examples. The output is a further logic program called the *hypothesis*. The table below, taken from (Pulman, 2000), shows the data for a typical ILP system.

Input		**Output**
Examples	Background	Hypothesis
% positive *flight*(*sk*80). *from*(*sk*80, *washington*). *to*(*sk*80, *atlanta*). *would_like*(*e*73, *I*, *sk*80). *event*(*e*73). % negative *not*(*to*(*e*73, *atlanta*)).	*city*(*washington*). *city*(*atlanta*).	*from*(*A*, *B*), *to*(*A*, *B*) → *false* *event*(*A*), *to*(*A*, *B*) → *false*

Table 1: Instances of input and output for an ILP system

The ILP algorithm is initialised by interpreting constraints and specifications describing the form of permissible candidate hypotheses. This generates the *hypothesis space*. Hypothesis generation is combined with a search strategy[1] (either specific-to-general or general-to-specific, depending on the implementation) during which some hypotheses are pruned and others are further generalized or refined. The algorithm contains a convergence criterion, which terminates the search process and gives a final result.

There are various frameworks for ILP, depending mainly on (a) how hypothesis construction is performed, and (b) how the accumulation of information provided by examples influences hypothesis formulation (Bergadano and Gunetti, 1996). According to the second criterion, systems can be classified into incremental and non-incremental. For non-incremental systems, the examples are provided at the beginning of the experiment and are not altered in any way during the process. In the case of incremental systems, the examples are input one-by-one by the user.

With respect to how hypothesis construction is performed, one can distinguish between the **normal semantics setting** (otherwise called explanatory ILP or the examples setting) and the **non-monotonic semantics setting** (also referred to as descriptive ILP).

It is important to make a distinction between the two settings of ILP, as different assumptions are made concerning the input and how relevant hypotheses

[1] In actual fact, ILP can be viewed as a search problem.

are induced. Which setting is ulimately employed depends on the special requirements of the application.

Explanatory ILP

In the normal semantics setting, the problem can be formulated as follows. Given background B and prior evidence E (where $E = E^+ \wedge E^-$, meaning that the evidence E consists of both positive evidence E^+ and negative evidence E^-) the aim is to find a hypothesis H such that the following conditions hold (Muggleton and DeRaedt, 1994):

Prior Satisfiability: $B \wedge E^- \not\models \mathit{false}$ (B and E^- are not inconsistent.)[2]
Posterior Satisfiability: $B \wedge H \wedge E^- \not\models \mathit{false}$ (any hypothesis H adopted is consistent with the negative examples.)
Prior Necessity: $B \not\models E^+$ (not all positive examples can be accounted for by the background.)
Posterior Sufficiency: $B \wedge H \models E^+$ (the hypotheses together with the background account for the positive examples.)

The Posterior Sufficiency rule is logically equivalent to

$$B \wedge \neg E^+ \models \neg H \qquad (1)$$

By finding the set of clauses $\neg C$ that are true in every model of $B \wedge \neg E^+$, a hyperset of H has been discovered, since from (1) and the definition of $\neg C$, $\neg C \models \neg H$ and hence, $H \models C$

The process denoted by '$\models$' is a type of generalization, which depending on the approach can be either θ-subsumption, entailment (in which case the paradigm is called Inverse Entailment (Muggleton, 1995)), or resolution (with the corresponding ILP paradigm being Inverse Resolution (Muggleton and DeRaedt, 1994; Mitchell, 1997)).

[2] The symbol '$\wedge$' represents conjunction (logical 'and') rather than intersection.

Descriptive ILP

The problem of inverting deduction adopts a different perspective in the non-monotonic semantics setting, otherwise called *learning from interpretation.* In this case there is no clear distinction between background knowledge and positive evidence. Each is expressed in terms of a set of definite clauses, called a model. The negative evidence is not provided beforehand, but rather derived implicitly by making a closed world assumption. This means that any combination of literals not explicitly present in the positive data is considered to be false. For instance, given the following background and evidence

$$B_1 = \begin{cases} bird(tweety) \leftarrow \\ bird(oliver) \leftarrow \end{cases} \qquad E_1^+ = flies(tweety) \leftarrow \qquad (2)$$

the clause $flies(oliver) \leftarrow$ would be considered as negative evidence. The aim is to find a hypothesis H such that the following conditions are met.

- $\forall e \in E^+$: H is true in $M^+(e \wedge B)$ (H is true in the minimal Herbrand model of e and B).
- $\forall e \in E^-$: H is false in $M^+(e \wedge B)$ (each negative example makes each clause in H false).

Thus, the space of candidate hypotheses is constructed by discovering all clauses true in $M^+(B \wedge E^+)$. A negative theory can be induced by collecting all clauses in $M^+(B \wedge E^-)$.

The above conditions can be contrasted with the posterior sufficiency and posterior satisfiability, which can be rephrased as

- $\forall e \in E^+: H \wedge B \models e$ (e is true in $M^+(B \wedge H)$)
- $\forall e \in E^-: H \wedge B \not\models e$ (e is false in $M^+(B \wedge H)$)

Comparing the conditions for learning from interpretation to the conditions of the explanatory setting, one can see that if $M^+(B \wedge H) = M^+(B \wedge E^+)$ (which holds if the two Herbrand bases are equal, $B(B \wedge H) = B(B \wedge E^+)$ (Muggleton and DeRaedt, 1994, p. 638)), then a hypothesis H in the explanatory setting is equivalent to a hypothesis in the descriptive setting. However, the inverse is not true.

Which flavour of ILP?

It is clear that the final choice of which framework to adopt depends on the application at hand. If one's goal is to induce a hypothesis able to predict the evidence, then the explanatory setting constitutes a wiser choice. On the other hand, if one is interested in capturing the regularities in a database or logic program, then the non-monotonic setting is more appropriate (Muggleton and DeRaedt, 1994, p. 671).

In (Pulman, 2000) the idea is to obtain a domain theory consisting of axiom-like rules able to aid the disambiguation task of PP attachment. Both the input data and the induced rules are specified in terms of what clause relations can occur in the body of a clause whose head is a prepositional phrase:

on(A, B):- meal(A), flight(B).
from(A, B):- flight(A), city(B).

This format may be informative for a disambiguation task resolving PP attachment where there is only ever one PP involved. In this case, resolving PP attachment can be viewed quite naturally as a prediction task. With multiple PPs involved, it would be necessary to have a number of rules, each with one of the PPs as its head. For instance, a special dinner *A*, on a flight *B* to city *C* would have to be described using the two rules

on(*A*, *B*):- *special_dinner*(*A*),*flight*(*B*), *to*(*B*, *C*), *city*(*C*).
to(*B*, *C*):- *flight*(*B*), *city*(*C*), *on*(*A*, *B*), *special_dinner*(*A*).

On the other hand, the rule

true:- *special_dinner*(*A*),*flight*(*B*), *to*(*B*, *C*), *city*(*C*), *on*(*A*, *B*).

would be sufficient since either PP can be the head according to one's needs. When one considers a disambiguation task other than PP attachment, such as pronoun resolution, this becomes more prominent. It is much more difficult to view pronoun resolution as a predictive task, as any NP featuring as subject, object, indirect object or adjunct can be replaced by a pronoun. In the sentence

The government gives benefits to the unemployed.

a pronoun could stand for any of the three nouns *government*, *benefits* or *unemployed.* It does not make sense to allow for the prediction of one semantic role over the other two. What would be most helpful in this case would be a rule such as

true:- *verb*(E_1, *give*), *subj*(E_1, *organizationClass*),
obj(E_1, *moneyClass*), *iobj*(E_1, *personClass*).

With this rule, given the example

The police watched as the employees protested. The company decided to cut down on their salaries.

it is possible to infer that the correct antecedent for *their* should be *the employees* rather than *the police.* Since the aim in the current context is to derive a domain theory for various disambiguation tasks, it makes sense to use the non-monotonic setting.

The systems CLAUDIEN and WARMR

CLAUDIEN (DeRaedt and DeHaspe, 1997) is a system developed in the non-monotonic semantics setting. It discovers frequent clauses by iteratively processing a queue of candidate clauses C that can be generalized further to valid patterns. These clauses C are generated according to the specifications provided by the user in a *DLAB grammar*. For details see (DeHaspe, 1998) and (DeRaedt and DeHaspe, 1997).

In compliance with the induction principle of the non-monotonic setting, the hypothesis H deduced from the observed examples and the background theory, holds for all possible sets of examples. This results in generalization beyond the observations and thus produces very strict and conservative clauses. This is due to the closed world assumption mentioned above. The side effect of the closed world assumption is useful when one is looking to generate a negative theory (rules having *false* as their head and describing what is not possible), since the system will generate a plethora of clauses which can then be filtered. However, when it

comes to a positive theory, CLAUDIEN is too strict as to what clauses it allows as solutions, which for sparse evidence can be problematic. Considering example (2), CLAUDIEN does not allow *flies*(X) ← *bird*(X) as a valid hypothesis, since there is no instance of *flies*(*oliver*). This is counter-intuitive and leaves no room for any ability to predict.

It would be useful if we could get rid of the closed world assumption and allow more flexible solutions. Such an approach is implemented in the system WARMR. WARMR considers many more solutions than CLAUDIEN, since instead of discovering full clauses as useful knowledge patterns, it discovers association rules, which have a much higher frequency than clauses. WARMR discovers *frequent queries* instead of frequent clauses. *Queries* are existentially quantified conjunctions $\exists(l_1 \wedge ... \wedge l_n)$ as opposed to clauses, which are universally quantified disjunctions $\forall(l_1 \vee ... \vee l_n)$ (equivalently, $\forall(h_1 \vee ... \vee h_k \leftarrow b_1 \wedge ... \wedge b_n)$). A query can be viewed as the negation of a clause with an empty head, since $\exists(l_1 \wedge ... \wedge l_n) \equiv \forall(\leftarrow l_1 \wedge ... \wedge l_n)$. For instance, *bird*(*tweety*) ← is a clause whereas *bird*(*tweety*) → is a query. WARMR has a mechanism for computing the frequency of queries in the input query queue and then computes specializations of queries frequent enough to pass a certain threshold. It adds these queries to the queue and moves infrequent queries out of the queue. The process is repeated until there are no more queries left in the queue (no more valid specializations can be made).

A Preliminary Experiment

A preliminary experiment was conducted using both CLAUDIEN and WARMR. The underlying idea was to learn from weaknesses and problems encountered and prepare the ground for a larger scale experiment. The goal of the initial experiment was the extraction of a set of rules (both associations as well as CLAUDIEN-style axioms) from a sample of 31 sentences in the WSJ. The domain rules of interest were of the nature of

true:- *verb*(E_1, *give*), *subj*(E_1, X_1, *organizationClass*),
obj(E_1, X_2, *moneyClass*), *iobj*(E_1, X_3, *personClass*).

This rule states that it is possible for an entity X_1 of class *organization* to give an entity X_3 of class *person* an entity X_2 of class *money*. One of the aims, therefore,

was to derive rules capturing associations between verbs and their arguments. Another interesting pattern would be one linking two verbs and their respective arguments, such as

true:- *verb*(E_1, *board*), *subj*(E_1, X_1, *personClass*), *obj*(E_1, X_2, *planeClass*),
verb(E_2, *fly*), *subj*(E_2, X_1, *personClass*).

According to this, it is possible for an entity of class *person* to board an entity X_2 of class *plane* and fly.

Finally, another type of information sought in this experiment was the mutual information between the arguments of a verb. This can be shown in the rule

true:- *subj*(E_1, X_1, *organizationClass*), *obj*(E_1, X_2, *moneyClass*).

This says that if an entity X_1 is a subject of class *organization*, then the corresponding object X_2 can be of class *money* (and vice versa). This type of rule is useful in the case where one wants to make the link between two arguments (e.g. agent and cause) and is not interested in the verb providing the link (i.e. when the verb is not frequent).

Description of Input

The 31 sentences mentioned formed the input evidence to the systems. The sentences were chosen so that they all included the verb *emphasize*, in the hope that they would constitute a concise domain, which would facilitate the derivation of useful relevant patterns. The data originated in the logical forms extracted in (Liakata and Pulman, 2002), which had to be converted to a format compatible with the requirements of CLAUDIEN and WARMR. In addition to this, the logical forms had to be translated into relations between fixed predicates, such as verb, subject, object, indirect object, adjuncts and modifiers.

Each sentence corresponding to a logical form is now mapped to a model, which is a Prolog program consisting of facts. An example of such a model is

```
begin(model('0001_1')).
funct_of('Nov._29',x4).
funct_of('Pierre_Vinken',x1).
funct_of(board,x2).
funct_of(director,x3).
mod(x1,old).
adjt(e1,as,x3).
adjt(e1,temporal,x4).
verb(1,main,e1,join).
subj(e1,x1).
obj(e1,x2).
end(model('0001_1')).
```

This corresponds to the logical form for the sentence *Pierre Vinken will join the board as a director, Nov. 29*,

and(*join*(e_1, '*Pierre_Vinken*', *the*(*board*)), *as*(e_1, *a*(*director*)), '*TMP*'(e_1, '*Nov*.29'))

The predicates *verb*/4, *subj*/2, *obj*/2, *adjt*/3 are self-explanatory. *funct_of*/2 assigns its second argument the property described by its first. Thus, *funct_of*(*director*, x_3) means that entity x_3 has the property of being a director.

CLAUDIEN and WARMR both accept evidence in this format. Systems like these take as input some amount of evidence (also known as a knowledge base), which may or may not be distinct from the background knowledge. In this case, the background knowledge consists of facts classifying properties of entities occurring as arguments of verbs into groups[3]. An instance of such a fact is

```
class(lawyer, 6, -15.110988, A):- funct_of(lawyer, A).
```

This states that an entity *A* described by the property *lawyer* belongs to class 6. The number in the third argument stands for the logarithm of the quantity

$$\frac{frequency\_in\_class6\_of(lawyer)}{total\_frequency(lawyer) \times total\_frequency\_of\_words(class6)}$$

[3] Details on how these groups were obtained are available in the explanation of the clustering method employed in (Liakata, forthcoming).

If this quantity exceeds a certain threshold, as in this case, the word *lawyer* is assigned to class 6. A suitable semantic label for class 6 is *enterprise-company-person.*

Another important aspect of the input is the specifications, which consist of semantic constraints regulating the form of the clauses or queries and the type of generalizations and refinements allowed. These restrictions are formulated in terms of a *DLAB grammar* in CLAUDIEN and *type* and *mode* declarations in WARMR.

A DLAB grammar defines a finite set of literal lists. It consists of a set of DLAB templates and a set of DLAB variables. A DLAB template (DeHaspe, 1998, Ch.4) is of one of two forms:

1. $p(t_1, ..., t_n)$, where p is a predicate symbol followed by a bracketed n-tuple of DLAB terms t_i;
2. $Min..Max : L$, where Min and Max are integers with $0 \leq Min \leq Max \leq length(L)$ and L a list of DLAB templates.

A *DLAB term* is a variable, or of the form $f(t_1, ..., t_n)$ (where f is a function symbol and the t_i are DLAB terms), or of the form $Min..Max : L$ (where $0 \leq Min \leq Max \leq length(L)$ and L is a list of DLAB terms).

The following is a well-formed DLAB template, taken from the input specifications for CLAUDIEN.

```
dlab_template('true

<--

2-len:[verb(Num, Main, Event, verbvalue1),
       len-len:[subj(Event,Id1), class(Word1, class1, Log1,
Id1)],
       len-len:[obj(Event, Id2), class(Word2, class2, Log2,
Id2)]
       ]')
```

The meaning of this template is that the clause to be derived should have *true* as its head and either a verb and a subject, with its corresponding class as its body, or a verb and an object with its corresponding class or a subject and an object each with its corresponding class, with or without a verb.

In WARMR the corresponding information is expressed by means of type and mode declarations, called *rmode*s. An *rmode* declaration corresponds to a possibly infinite set of literal lists. *Mode* constraints declare whether a variable is an input variable (+), an output variable (−) or both (±). *Type* constraints make sure that variables shared between different predicates are of the same type. Below is an example of an *rmode* declaration.

```
type(class(word, class, log, id)).
type(verb(sentence_id, subordination, event, vname)).
type(subj(event, id)).

rmode(1 : #((Class) :
    (verb(-N, -S, \Event,_),
     subj(Event, -ID),
     class(-Word, Class, _, ID)))).
```

This says that the combination of *verb*/3, *subj*/3, *class*/3 should be included in one run of the query specialization, where the verb is required to be a new instance, not one already encountered, and the class of the subject is a constant generated from the data.

Even though the two formalisms may appear completely distinct, their differences are not crucial. DLAB allows the user to give more fine-grained definitions of useful patterns whereas *rmode*s are easier for designing quick implementations. They are also preferable when a lot of variables are expected, even though they may initially result in the matching of meaningless patterns against the database.

Discussion of Results

The rules obtained did not manage to capture all the associations sought. The initial runs were only capable of finding sortal information between a verb and one of its arguments and mutual information between the arguments of a single verb. Instances of such rules are

CLAUDIEN:

true:- *verb*(*A*, *B*, *C*, *provide*), *subj*(*C*, *D*), *class*(*E*, *'enterprise-company'*, *F*, *D*).
true:- *verb*(*A*, *B*, *C*, *provide*), *adjt*(*C*, *with*, *D*), *class*(*E*, *'financial_action-person-company-cost-cut-share'*, *F*, *D*).
false:- *verb*(*A*, *B*, *C*, *trace*), *adjt*(*C*, *D*, *E*), *class*(*F*, *'product-company'*, *G*, *E*).
false:- *verb*(*A*, *B*, *C*, *disarm*), *subj*(*C*, *D*), *class*(*E*, *'product-company'*, *F*, *D*).

WARMR:

freq(15, 1, [*sentence*(*A*), *verb*(*A*, *B*, *C*, *D*, *emphasize*), *subj*(*A*, *D*, *E*), *class*(*F*, *'enterprise-company-person'*, *G*, *A*, *E*)], 0.29).
freq(695, 1, [*sentence*(*A*), *verb*(*A*, *B*, *C*, *D*, *risk*), *subj*(*A*, *D*, *E*), *class*(*F*, *'prosecutor-analyst-person-consumption-number-share'*, *G*, *A*, *E*)], 0.032).

CLAUDIEN has produced clauses whereas WARMR has captured patterns with a frequency ratio given by the fourth argument of *freq*/4. The absence from the rules of any other more sophisticated link, such as the association between two verbs and their arguments or even a single verb and its arguments can be attributed to the sparsity of the data relative to its size; there just weren't enough frequent patterns to capture. The frequency threshold had been set to a minimum threshold of two examples but even this was too high an expectation for the size and quality of the data. What is meant by the quality of the data is that, even though the 31 sentences were sampled to include the verb *emphasize* (a moderately frequent verb), they did not constitute a concise domain. A better choice would have been the sentences containing instances of the verb *resign*, as they seem to define the domain of company successions. Work in progress involves an example set of 93 files (a total of 2201 sentences) containing various forms of the verb *resign*.

Another way to remedy the sparseness of data would be to cluster the verbs, so that instead of an association between two verbs and their arguments one would obtain a pattern containing two verb group labels and their corresponding arguments. A hierarchical clustering of the verbs in the corpus which also occur in WordNet has been developed, based on the hypernym relations defined in WordNet and an efficient way to embed them in the WARMR settings is being

explored[4]. A method for improving efficiency in WARMR, so as to allow for fast deep level search is also under way[5].

CLAUDIEN took a very long time to run and was stopped without converging. It was decided that CLAUDIEN would only be used for the derivation of a negative theory (which is not obtainable with WARMR) as the positive rules it produced added nothing to those output by WARMR. Moreover, the number of rules CLAUDIEN produced was much smaller (since it only generates clauses) and it took much longer to derive them. It is hoped that for a larger dataset, the negative rules generated by CLAUDIEN will be more meaningful.

As one can see from the rules, the semantic tags assigned to the argument classes are quite vague and often rather counter-intuitive. One of the tasks currently being undertaken is clustering the arguments under more informative categories by mapping them to the clusters produced by Pantel and Lin (2002).

Future Work

The results obtained for the sample of thirty-one sentences were satisfactory for the scale of the experiment, but showed that there is still room for improvement. Apart from continuing work in progress, the following objectives need to be met in the near future.

- More sophisticated *rmode*s need to be written for WARMR, to allow the specification of the type of dependency relation between verbs or verb groups co-occurring in the same pattern.
- *rmode*s should also be extended to capture the associations between arguments and modifiers, especially prepositional phrase modifiers.
- In order to make use of the rules induced for disambiguation purposes, it is essential to develop a mechanism for probabilistic reasoning.

[4] in conjunction with Jan Struyf and Jan Ramon of the Artificial Intelligence group at the Katholieke Universiteit Leuven

[5] current work by Jan Struyf

References

Alshawi, H. and J. Eijk. 1990. Logical forms in the Core Language Engine. In *Proceedings of 26th Annual Meeting of Association for Computational Linguistics*. Vancouver.

Bergadano, F. and D. Gunetti. 1996. *Inductive Logic Programming: from machine learning to software engineering*. MIT Press.

Brooks, J. and M. Collins. 1995. Prepositional Phrase Attachment through a Backed-off Model. In *Proceedings of the Third Workshop on Very Large Corpora*. Cambridge, Massachusetts.

Dagan, I., L. Lee and F. Pereira. 1997. Similarity-Based Methods for Word Sense Disambiguation. In *Proceedings of Association for Computational Linguistics 35th / EACL 8th*.

DeHaspe, L. and L. DeRaedt. 1996. DLAB: A declarative language bias formalism. In *Proceedings of the International Symposium on Methodologies for Intelligent Systems (ISMIS96)*, volume 1079 of Lecture notes in Artificial Intelligence, pages 613–622. Springer-Verlag.

DeHaspe, L. 1998. *Frequent Pattern Discovery in First-Order Logic*. Ph.D. thesis, Katholieke Universiteit, Leuven.

DeRaedt, L. and L. DeHaspe. 1997. Clausal discovery. *Machine Learning*, 26(2):99–146.

Hobbs, J., M. Stickel, D. Appelt and P. Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, number 63:69–142.

Hobbs, J. 1978. Resolving pronoun references. *Lingua*, number 44:311–338.

Liakata, M. and S.G. Pulman. 2002. From Trees to Predicate-Argument Structures. In *International Conference for Computational Linguistics (COLING)*, Taipei, Taiwan.

Liakata, M. (forthcoming). Inducing domain theories. D.Phil. thesis, University of Oxford.

Mitchell, T. 1997. *Machine Learning*. MIT Press and The McGraw-Hill Companies Inc.

Muggleton, S. and L. DeRaedt. 1994. Inductive Logic Programming: Theory and Methods. *Journal of Logic Programming*, 19,20:629–679.

Muggleton, S. 1995. Inverse Entailment and Progol. *New Generation Computing*, special issue on Inductive Logic Programming, 13(3–4):245–286.

Pantel, P. and D. Lin. 2000. An Unsupervised approach to Prepositional Phrase attachment using contextually similar words. In *Proceedings of Association for Computational Linguistics 2000*, Hong Kong.

Pantel, P. and D. Lin. 2002. Concept Discovery from Text. In *Proceedings of COLING 2002*, Taipei, Taiwan.

Pulman, S.G. 2000. Statistical and logical reasoning in disambiguation. *Philosophical Transactions of the Royal Society*, 358 number 1769:1267–1279.

Pulman, S.G. 2002. Learning domain theories. Slide presentation at Edinburgh Informatics Jamboree.

Ratnaparkhi, A. 1998. Unsupervised Statistical Models for Prepositional Phrase Attachment. In *Proceedings of COLING-ACL98*, Montreal, Canada.

Word-level prominence distinctions in Tamil

ELINOR KEANE

1 Introduction

This paper examines whether Tamil makes any prominence distinctions at the word level. There are no stress-based lexical contrasts in the language, and previous work, which is largely impressionistic, has reached no consensus on the existence of prominence distinctions, or their location within either the word or phrase. Results reported here indicate that word-initial syllables are marked by phonetic correlates associated with vowel reduction, which has frequently been implicated in prominence distinctions cross-linguistically. Vowels in initial syllables are significantly longer and less centralized in quality than those of word-medial or word-final syllables. Evidence is also presented suggesting that this pattern of initial prominence is independent of phrase-level effects.

2 Background

There are no lexical distinctions based on stress in Tamil and native speakers have no clear intuitions about its placement. There is no consensus in the literature on whether stress exists in the language, or on its location, although most descriptions favour either fixed initial stress or a quantity-sensitive dynamic system. These claims are largely based on impressionistic observations, but what little experimental work has been conducted has also failed to find any consistent phonetic correlates of non-emphatic stress (Balasubramanian 1972, 1980). This suggests that Tamil may not mark prominence at all at the word-level, as has been claimed, for instance, for Indonesian (Goedemans & van Zanten forthcoming).

Another line of evidence points in a different direction: it has been observed that the vowels /i/, /a/ and /u/ have centralized quality and reduced duration in non-initial syllables (Christdas 1988, Schiffman 1999: 17). The initial syllable is also singled out by Schiffman (1999: 23) as the only position in which /ai/ tokens are

fully produced as diphthongs. Monophthongization to [ɛ] or [a], depending on dialect, is also reported by Asher (1982: 219), but only for final syllables of polysyllabic words. Differences in duration and vowel quality are well-established correlates of stress cross-linguistically, so this indicates that the language may mark prominence.

This study was designed to distinguish between the two possibilities – presence or absence of prominence at the word-level – by considering evidence of vowel duration and quality.

3 Method

Test words were selected containing the target vowels /i/, /a/ and /u/ and the diphthong /ai/ in different syllable positions – initial, medial and final. All were presented in the context of simple sentences, with no indication of emphasis. The immediate phonetic environment of the vowels was balanced as far as possible, as shown in Table 1, which contains all the /a/ vowels. The test sentences were recorded five times each by three native speakers, two male and one female. All were long-term residents of Pondicherry, a town in south-east India, where the recordings were made.

Tamil has both a colloquial spoken variety and a formal variety used for writing, and the two are sufficiently divergent for the language to be classed as diglossic (Britto 1986). Written representations of colloquial Tamil are relatively rare, and there are no standards, official or unofficial, for spelling. In preparing the stimuli for orthographic presentation, therefore, colloquial forms were written as phonetically as possible within the limits of the writing system. Subjects were encouraged to speak in a colloquial fashion and encountered no difficulties in reading the sentences fluently. Nevertheless, it is possible that their production was influenced to some degree by spelling pronunciations. None of the speakers was sufficiently proficient in English to be comfortable translating into Tamil on the spot, which would have avoided the danger completely.

The data were recorded using a DAT recorder and lapel lavalier microphones (Audio-Technica AT803b), and subsequently digitized at a rate of 16 kHz (16 bit resolution). In a few cases tokens were discarded because of interference from background noise, accidental slips by the speaker or absence of identifiable formants. The total number of vowel tokens measured was 1139 monophthongs and 162 diphthongs. For each token cursors were placed at the onset and offset

using the ESPS/*xwaves*™ software. The duration of the vowel was measured, and also the first and second formant frequencies at its midpoint. For the diphthongs F1 and F2 were additionally measured at points one quarter and three quarters of the way through the vocalic portion of the signal.

Linear predictive coding analysis was used to track the formants automatically, applying autocorrelation within a pitch period to identify broad spectral peaks corresponding to vocal tract resonances. Twelve coefficients were used in the linear predictive equation, and the window duration was 49 milliseconds, with 5 millisecond steps between analysis frames. The tracking was originally set to locate three formants, and this worked reasonably well for the two male speakers. Some difficulties were encountered for the female speaker, but these were largely resolved by resetting the tracking to search for four formants during measurement of her tokens.

Context	**Initial**	**Medial**	**Final**
v_n	*vantaa*	*avanukku*	*avan*
v_	*varaar*	*koɳʈuvaraatee*	*kiɻavan*
l_	*laɳʈan*	*uḷḷaŋkai*	*koolam*
_n	*anpu*	*mookanin*	*paiyan*
r_	*rattam*	*paattirattai*	*citamparam*
t/ʈ_	*talai*	*paarttatu*	*eŋkiʈʈa*
t_m	*tampi*	*citamparam*	*rattam*
p_ra	*parantatu*	*citamparam*	
_nta	*anta*	*parantatu*	
p_ʈi	*paʈiccaan*	*ippaʈi*	
k_	*kattatu*		*uʈaikka*
_ll/#	*nallatalla*	*nallatalla*	*aɻakaana*
n_		*enakku*	*ena*
_n		*paiyan-ŋŋu*	*mookan*
ʈ_		*paʈattai*	

Table 1: Words containing /a/ tokens: target vowels are emboldened.

Statistical analysis was performed using analyses of variance in which the independent variables were syllable position and speaker. In cases where the Levene test indicated that the variance of the data was not homogenous, the non-parametric Mann-Whitney test was employed.

3. Results

Consistent and significant differences between speakers were found for the formant frequency values, both for the data set as a whole and for the individual vowels. Duration, however, was relatively unaffected by interspeaker variation: a significant difference was found only for the /a/ vowels ($p < .001$).

Analysis of the /a/ tokens revealed that their duration was significantly affected according to whether they appeared in an initial syllable or not ($p < .0005$). Boxplots illustrating the duration values are displayed in figure 1, which indicates that the primary distinction is between longer initial and shorter non-initial syllables, rather than being a gradient effect. This was confirmed by a repeated measures analysis of variance which showed the three-way distinction in syllable position not to be significant.

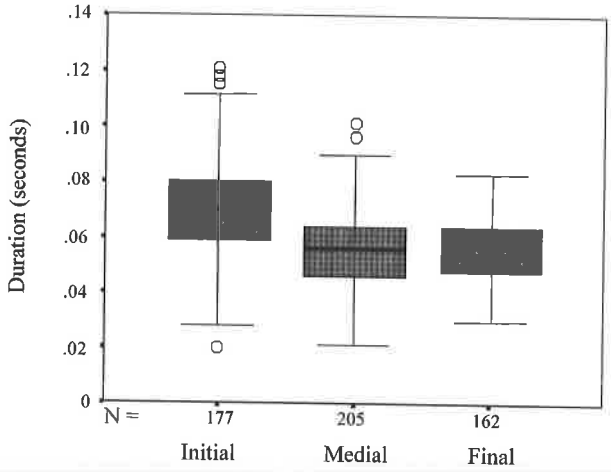


Figure 1: Boxplots of duration of /a/ tokens against syllable position for all speakers.

Syllable position also had a significant effect on the formant frequency values for /a/, F1 being higher ($p < .0005$) and F2 lower ($p < .001$) in initial than non-initial syllables. These results all suggest that the vowels of non-initial syllables are reduced in duration and centralized in quality.

Results for the /i/ tokens from the original data set proved inconclusive: values for duration in particular were confounded by a strong effect of final lengthening. A more controlled data set was therefore used, consisting of four reduplicated expressive words in which /i/ vowels appeared in the first and third syllables e.g. *kirukiruppu* 'giddiness'. Syllable position was found to have a significant effect on duration ($p < .013$), with vowels in initial syllables being longer. They were also characterized by significantly lower F1 and higher F2 ($p < .0005$ for both), indicating a more peripheral vowel quality.

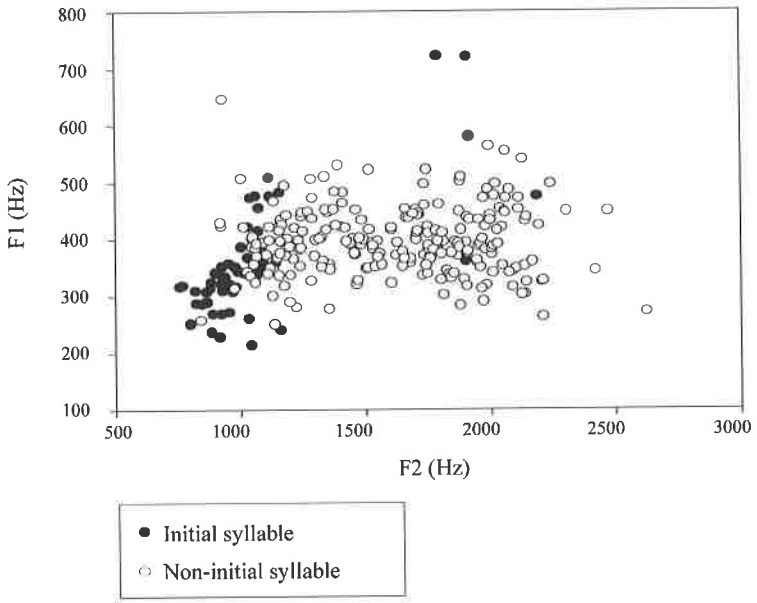


Figure 2: Scatterplot of F1 against F2 for /u/ tokens in initial and non-initial syllables for all speakers.

The same pattern emerged clearly from analysis of the /u/ tokens: significantly longer duration in initial than non-initial syllables ($p < .004$) and highly significant differences in the formant frequencies correlated with syllable position ($p < .0005$), as illustrated in figure 2. The higher values for both F1 and F2 in non-initial syllables would again be consistent with a more centralized articulation but might also reflect less lip-rounding. (The five initial tokens that are not clustered with the rest were all contained in repetitions of a single test sentence by the female speaker: why these results diverge so markedly from the rest is not clear.) As with the /a/ tokens, a repeated measures analysis of variance confirmed that the significant distinction is between initial and non-initial syllables, rather than a three-way gradient effect.

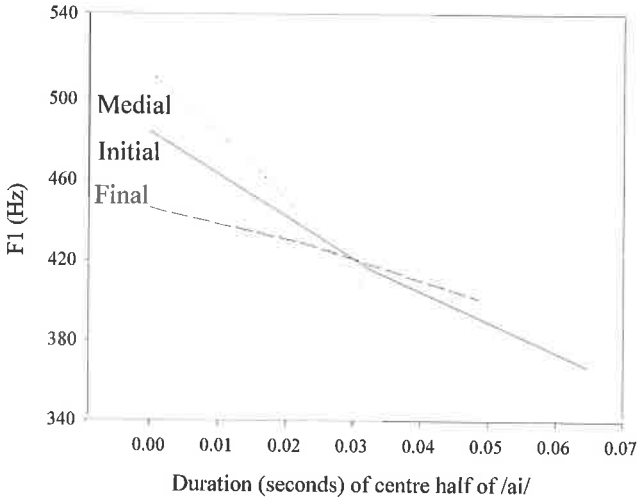


Figure 3: Line chart showing the change in F1 over the centre half of /ai/ in initial, medial and final syllables for all speakers.

The duration results for the /ai/ diphthongs are in line with those for the monophthongs, i.e. the overall duration of the vocalic portion of an initial syllable is significantly longer than those of non-initial syllables ($p < .0005$), and there is no significant interspeaker variation. The change in formant frequencies over the central half of the diphthong was also affected by syllable position, the degree of change being significantly higher in initial than non-initial syllables ($p < .0005$ for

both F1 and F2). Comparison of the rates of change revealed no marked differences: the gradients in figures 3 and 4 are not strongly differentiated. This suggests that the reported monophthongization is primarily a matter of reduced duration, rather than a contrast between a dynamic and a steady-state segment. However, it would be unwise to draw any firm conclusions, given a possible effect from spelling pronunciations.

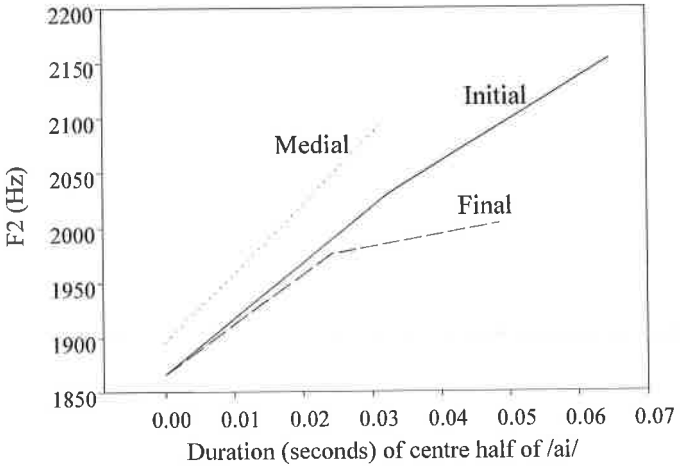


Figure 4: Line chart showing the change in F2 over the centre half of /ai/ in initial, medial and final syllables for all speakers.

No effort has been made thus far to control for the potentially confounding effects of accent, so the possibility that these results reflect prominence at the phrase rather than the word level should be considered. There is evidence for Dutch that accent increases the durations of both vowels (van Bergem 1993) and syllables (Sluijter & van Heuven 1996), and also affects spectral quality, although to a lesser degree than word stress. Moreover, the effects of accent are not limited to the specific syllable with which the main pitch movement is associated but are found in all syllables of the accented word. In order to control for any such effects

in Tamil, tokens of vowels in unaccented words were analysed to see whether the same pattern of results would be found.

As with word-level prominence, there is no agreement on the location of accent in Tamil. Word order is consistently SOV and informal inspection of F0 contours suggests that the main accent does not fall on the rightmost constituent. When two lexical phrases precede the verb it is marked by a slight fall in the initial syllable to a level plateau. This typically extends to the final syllable, where there may be a further fall in declarative sentences. By contrast, pronounced peaks are associated with the preceding constituents. Tokens of vowels in sentence-final constituents were therefore assumed to be unaccented.

Eight /a/ vowels, all in verbs associated with the intonation contour described above, were chosen, four in initial and four in non-initial syllables. No significant differences in duration were found, which may be due to syllable structure, there being a disproportionate number of closed syllables in the initial condition. However, differences in vowel quality precisely paralleled those in the larger data set, i.e. significantly higher F1 and lower F2 in initial than non-initial syllables ($p <$ .0005 for both). Figure 5 illustrates how the tokens broadly separate into two clusters according to whether they occur in initial or non-initial syllables.

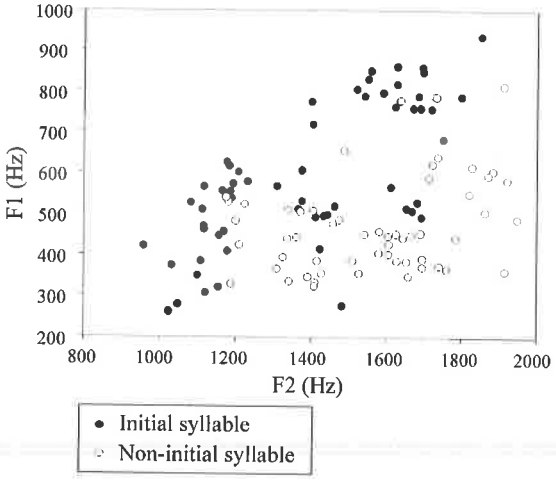


Figure 5: Scatterplot of F1 against F2 for unaccented /a/ tokens in initial and non-initial syllables for all speakers.

Discussion

These results are evidence that the three Tamil monophthongs considered undergo reduction in both duration and quality in non-initial syllables. Moreover, these same effects are found in /a/ tokens controlled for the presence of accent, suggesting that they are independent of any phrase-level prominence. /ai/ tokens in initial syllables are also distinguished by greater length, although the rate of formant change is not significantly affected by syllable position. In conclusion, it appears that there are phonetic correlates of prominence at the word-level in Tamil, and that they are consistently associated with an initial syllable.

The phonetic parameters analysed in this study are limited, and it is possible that they are complemented by other correlates. Candidates for future investigation include the intensity of vowel tokens: unfortunately the recording conditions meant that the level of background noise was too high for any such analysis in this study. Spectral balance should also be considered: this is reported to be a more reliable correlate of word-level prominence than overall intensity in Dutch (Sluijter & van Heuven 1996). Another avenue for future research, largely unexplored at present, is the interaction between pitch and prominence in Tamil. Given the results reported here, the expectation would be that significant pitch excursions are associated with initial syllables. Informal inspection supports this hypothesis and it is being tested in work in progress.

A further line of investigation suggested by these results concerns the /ai/ diphthongs. In environments where monophthongization has been reported there was no firm evidence that formant structures are flattened. Since spelling pronunciations may be responsible, a priority would be to determine whether the same pattern is found in utterances where such an effect can be ruled out. Furthermore, as the structure of diphthongs is known to vary cross-linguistically (Lindau, Norlin & Svantesson 1990), it would be interesting to establish the salient characteristics for Tamil, and in particular the relative importance of a dynamic structure, compared to having particular formant frequency values at the onset and offset.

A final question concerns the function of prominence in Tamil. The phonetic properties by which it is marked match those of classic stress-accent languages such as English and Dutch. However, there seems to be a difference of degree, or at least of perceptual salience, given that native speakers (even those with phonetic training), apparently find it so hard to judge where prominence is placed. One

possibility is that its main role is to mark word boundaries: certainly giving prominence to initial syllables accords well with the morphological structure of Tamil. There is little or no prefixation in the language, and inflectional morphology involves the concatenation of one or more suffixes to a lexical root. A final, speculative suggestion would therefore be that the phonetic prominence of initial syllables revealed in this study guides the listener to the lexical content of an utterance.

References

Asher, R.E. (1982). *Tamil*. Amsterdam: North-Holland.

Balasubramanian, T. (1972). *The phonetics of colloquial Tamil*, Ph.D. dissertation, University of Edinburgh.

Balasubramanian, T. (1980). Timing in Tamil. *Journal of Phonetics* 8. 449–467.

Britto, F. (1986). *Diglossia: a study of the theory with application to Tamil*. Washington DC: Georgetown University Press.

Christdas, P. (1988). *The phonology and morphology of Tamil*. Ph.D. dissertation, Cornell University.

Goedemans, R.W.N. & E. van Zanten. (To appear). Stress and accent in Indonesian. In D. Gil (ed.) *Malay / Indonesian Linguistics*. London: Curzon Press.

Lindau, M., K. Norlin & J.-O. Svantesson. (1990). Some cross-linguistic differences in diphthongs. *Journal of the International Phonetic Association* 20.10–14.

Schiffman, H.F. (1999). *A reference grammar of spoken Tamil*. Cambridge: Cambridge University Press.

Sluijter, A.M.C. & V. van Heuven. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America* 100.2471–2485.

van Bergem, D.R. (1993). Acoustic vowel reduction as a function of sentence accent, word stress and word class. *Speech Communication* 12.1–23.

How to tell beans from farmers: cues to the perception of pitch accent in whispered Norwegian

HANNELE NICHOLSON AND ANDREAS HILMO TEIG

1 Introduction

Though quite distinct in physical form, 'beans' and 'farmers' are distinguishable lexically only by a difference in pitch accent in spoken Norwegian. In the absence of pitch information, Fintoft (1970) investigates whether or not listeners are able to discern between the two tokens in whispered speech – without the assistance of context. Though his findings suggest that listeners may rely upon the presence of an additional cue present in the whispered speech stream, the possibility that context could aid the listeners in detecting the appropriate pitch accent is not discussed. This paper presents the results of an experiment conducted to assess listeners' ability in determining which pitch accent word token best fits into a whispered ambiguous utterance in spoken Norwegian. The results confirm that context is not a reliable cue to assist in lexical selection and concur with Fintoft (1970) in suggesting that listeners utilise a separate prosodic cue, possibly syllable duration or intensity, to make the pitch accent distinction in whispered speech.

2 The Problem

East Norwegian employs pitch accent contours in order to make lexical distinctions. For example, the words /[1]bøner/ '*farmers*' and /[2]bøner/ '*beans*'[1] consist of identical phonemes. There are two ways of distinguishing between these words in isolation: to rely on a) the context within which the utterance is created or b) the pitch accent accompanying the word. This paper seeks to investigate whether there is enough information present in the speech signal for the speakers not to have to rely on any of these, say during whisper.

[1] Accents 1 and 2 are normally marked with a superscript 1 and 2 in Norwegian.

From Figure 1, we can see that pitch accent 1 is associated with a simple lexical L*, here followed by a H% boundary tone[2]. Pitch accent 2, on the other hand, consists of a lexical H* and a L, here occurring with a H% boundary tone. There are some 3000 pairs of words in the language, which are disambiguated by pitch accent in this way (Fintoft 1970).

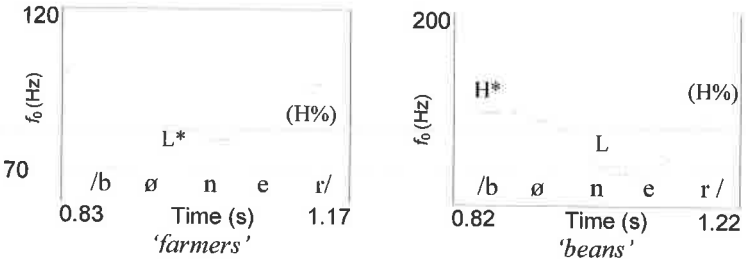


Figure 1. f_0 traces with typical intonation contours for two words distinguished only by accent 1 (left) and accent 2 (right).

The second part of an explanation of East Norwegian pitch accent is best expressed in phonetic terms. During whispered phonation, the vocal folds do not vibrate (see e.g. Catford 1977). This is due to properties of the glottis, whereby in order for whisper phonation to be produced, the glottis is considerably narrowed at the anterior end. This narrowing allows a turbulent air stream to pass through the glottis without the generation of voicing. Figure 2, from Saunders (2002), portrays a comparative picture of the glottis during voiced and whispered speech.

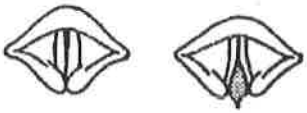


Figure 2. A view of the glottis during voiced speech and whispered speech.

[2] This follows the system of Norwegian intonation called the Trondheim model, described in e.g. Nilsen 1992.

During voice production, it has long been known that the vocal folds must vibrate in order to produce pitch (cf. e.g. Lass 1996). The vocal folds open and close at a periodic rate, known as fundamental frequency, or f_0 (see e.g. Ladefoged 2001). Therefore, it is physically impossible for a speaker to produce whispered speech if the vocal folds are vibrating. In this context, there can be no perception of f_0, hence the listener must rely on either the linguistic or extra-linguistic context or some other acoustic cue to distinguish between pitch accent contours in Norwegian. One of the important questions for our research in this paper is whether or not such other acoustic cues are present. In his seminal study of Norwegian pitch accents, Fintoft (1970) investigated the issue for whispered Norwegian. However, his study concerned only identification of pitch accent in single word tokens spoken in isolation.

2.2 Is accent perceptible in whispered speech?

The fact that pitch information is dependent upon f_0 information leads one to wonder how languages that consistently employ tone make lexical distinctions function in the absence of pitch information, such as in whispered speech. This question has been researched before for Chinese, Thai, Norwegian and Swedish (Giet 1956). No matter what language is used as a means of investigating the issue, the conclusion is the same. Pitch is perceptible in whispered speech. There is much debate however, over the mechanism, whether contextual or prosodic, that may act as a substitute for voicing and pitch.

Meyer-Eppler (1957) suggests that intonational information is understood via a spectral shift in whispered speech. To research this matter, Meyer-Eppler (1957) conducted an acoustic analysis of German vowels that had been "sung" without voicing. For front vowels, a second formant (F2) was detectable at 2000 Hz. The first formant (F1) of back vowels exhibited a value much higher than that of a normal voiced vowel. The validity of Meyer-Eppler's results is somewhat contestable, in that requesting subjects to "sing" without voicing is a rather unnatural, non-speech like act.

A second observation, also made on the basis of evidence from whispered vowels, suggest that listeners perceive a pitch in the F2 range (Thomas 1969). Three musically trained listeners were presented with a variety of whispered stimuli and asked to adjust an oscillator dial to within 10 Hz of the perceived pitch. This task was repeated twice for each of the three listeners and listeners

consistently reproduced the same responses on the second occasion. This could be used to suggest that listeners have no difficulty in participating in a pitch perception task where no f_0 information is provided. It also suggests that there is a clear association between the formant frequencies and the perceived pitches of the whispered vowels (Thomas 1969).

A more recent general study asserts that listeners are able to perform voice recognition tasks by comparing a whispered vowel with voiced speech from a variety of speakers (Tartter 1991). Tartter makes no particular claim for which prosodic cue is possibly present during a whispered token. However, subjects consistently mistook tense and lax vowels for one another (e.g. [ɪ] vs. [i]; [ʊ] vs. [u]) that differed in information present in the F1/F2 space.

From the studies outlined above, one might conclude that f_0 is not a necessary cue to the perception of pitch in whispered speech. The F1/F2 space seems to be an important and consistent finding amongst the reports summarised here. However, this issue lies outside the scope of the present paper and for future work to pursue.

3 Whisper in Norwegian

The aforementioned work conducted by Fintoft (1970) reported on a focused study on the dynamics of whisper and perception in Norwegian speech. His study was largely focused on the variation between dialects in voiced speech (an issue that is not pursued in the present study), but he also addressed the same issues with respect to whisper. To test a listener's ability to distinguish between voiced and whispered speech, Fintoft presented subjects with the pair *live* ('(a)live') – *livet* ('the life'), without any surrounding syntactic or semantic context. Subjects were able to make a reliable distinction, though the ability varied across dialect groups. Speakers of East Norwegian from the Trondheim area had the most difficulty in perceiving the difference, though by and large were able to complete the task in a satisfactory manner. Fintoft finds that listeners are likely to rely on other parameters apart from f_0. Intensity and syllable duration are potential candidate cues to the pitch accent distinction in whispered Norwegian speech.

Findings from Hadding-Koch (1961, 1962) corroborate the claim that intensity may be one such parameter with observations made from Swedish. She observed that subjects' judgements for two evenly spoken syllables in a continuous fashion with no pauses corresponded to pitch accent one. On the other hand, two evenly

spoken syllables that were produced with low intensity or a small pause in between were often judged as pitch accent two.

Though Fintoft presents a thorough investigation into the potential prosodic cues for whispered pitch accent, there are some issues regarding his experiments that merit attention today. First, his study only tested subjects for accent perception on single-word utterances. The difficulty stemming from this is that a subject is given a single task, accent perception, on which all attention may be focussed. If this were avoided, by inserting the crucial words in an otherwise identical context, or by telling subjects to identify non-related differences in meaning, one is more likely to achieve results that are realistic, and speech-like. The second, more important factor is the technical development that has taken place since Fintoft's work. With improved equipment and updated methods, such as speech synthesizers, we expect that the difference in importance between context and phonetic cues, and the issue of phonetic cues themselves, may be addressed with a hope of stronger conclusions. Fintoft himself proposed that future work should incorporate synthetic stimuli into the whisper experiment so as to compare subject reactions to natural speech. We therefore set forth to revisit Fintoft's findings with respect to these issues.

4 The Experiment

Our focus in this paper, then, is the status of phonetic cues in the perception of whispered speech. In order to investigate this question, we devised a perceptual listening experiment. This test was designed principally to investigate whether pitch accent can be perceived in whispered speech without the assistance of context, i.e. just on the basis of phonetic cues. As has been shown previously, there has long been a disagreement and uncertainty on this issue, and empirical research presented from previous work has not addressed this issue from the perspective of a real-life communication situation. This feature makes them less interesting, for the reasons mentioned above.

Our experiment presented seven pairs of incomplete utterances, coupled with two possible completions. For each incomplete utterance, there was one critical word, occurring at the end of the utterance, which provided the pitch information necessary to obtain the only possible semantic disambiguation. Only incomplete utterances were included in the auditory stimuli. Our subjects were instructed to select one of two possible completions for each stimulus, and their choice of

completion would be determined by their interpretation of the word with pitch accent. For example, in (1), the incomplete utterance is phonologically identical for both senses of the word (despite the orthographic difference). The intonation contour on the final word would be the only cue available to help disambiguation and the listener would choose the appropriate completion, either a) or b) based on which contour they think they have heard.

(1) Jeg hørte at noen bønder/ bønner…
I heard that some farmers/ beans…
'I heard that some farmers/beans…'

(a) har reist til Oslo for å protestere.
have travelled to Oslo for to-inf protest.
'have travelled to Oslo to protest'

(b) har blitt trukket tilbake av Rema fordi de var forgiftet.
have been drawn back by Rema because they were poisoned.
'have been withdrawn by Rema because they were harmful.'

The two possible completions in (1a) and (1b) are designed so that neither is felicitous with more than one pitch contour. In this example, it is evident that only farmers (accent 1 variant) can travel to Oslo to protest, whereas it is only beans (accent 2 variant) that can be withdrawn because of harmfulness. A complete list of the utterances used can be found in the appendix.

In order to address the issue of perception during whisper, we devised three test conditions for our stimuli. The conditions were designed to test a listener's ability in perceiving pitch accent in three separate acoustic environments. The first acoustic condition, the baseline condition, consisted of normal speech, recorded in a normal reading voice by a native Norwegian speaker (the second author) in a sound-proofed recording studio. The second condition had the same 7 utterance pairs, only this time recorded in a whisper. The third acoustic condition was resynthesised from the original voiced baseline condition. This resynthesis was performed with an application available in the Oxford Phonetics Laboratory, which splits the speech signal of its input, removes voicing information, and recombines spectral information and frication, to produce a 'synthetic' whisper. The purpose of the resynthesised condition was to show beyond a reasonable doubt whether

there were additional phonetic cues involved in naturally whispered speech which may assist a speaker to distinguish one pitch accent from another. The implementation of the third condition, if subjects proved able to correctly distinguish pitch accents 1 and 2, allows us to say with certainty that any extra cues for pitch accent - beyond pitch information - must be present already in normal speech.

For each of the three acoustic conditions, the seven pairs were randomised and repeated five times, giving the subjects 70 utterances to judge for each condition. Each condition was presented separately, which made it easier for subjects to concentrate on each condition in turn[3]. There were nine subjects, all native speakers of an East Norwegian dialect. The stimuli were presented in a sound-proofed room via loudspeakers in randomised order, and the subjects were asked to indicate which completion they preferred. For this purpose they were provided with a written text of all the stimuli, similar to (1) above, and could follow the written text while selecting the appropriate completions.

Four subjects completed the test in the Phonetics Laboratory, and five subjects did the test online[4]. Because of the nature of this study, and in order to secure similar acoustic conditions for all subjects, we had to select and monitor closely who participated in the online part of the study. Detailed instructions were provided for the subjects doing the test online with regards to achieving reliable results, and for preparing them for what to expect. We particularly stressed the need for using earphones and quiet conditions. The design of the online application made it necessary to do the entire test in one sitting, mimicking laboratory conditions. We found that performing the listening test online was a very helpful way of enabling subjects to participate who were not able to come to the laboratory. Even though this is not the traditional way of performing perception tests, it is becoming more common as access to internet technology is more widely available (cf. e.g. Caspers 2000).

[3] This was particularly important with the resynthesised stimuli. Our reviewer noted that the three conditions should have been counterbalanced across subjects, but we were not aware of this at the time of the experiment design.

[4] Available at http://teig.nvg.org/lyttetest/

5 Findings and discussion

The results show that, as might be expected, subjects were overwhelmingly successful in disambiguating of the meaning in the incomplete utterances. As there were no pre-test effects of sentence and tone, a univariate ANOVA[5] with acoustic condition, subject, and correct perception as factors was applied to the data. This test gauged the subjects' ability to correctly perceive pitch accents (regardless of whether they were accent 1 or 2) in each of the acoustic conditions. While the acoustic variable was not independently significant, a significant interaction was obtained between correct perception and the acoustic variable ($F(2,48) = 139.72$, $p<.001$). This result enables us to deny the null hypothesis. A subject's ability to discriminate correctly between possible completions is dependent upon the acoustic variable.

Table 1 demonstrates how well subjects were able to perceive each accent individually in each of the three acoustic conditions. Hardly any errors occurred in the responses to the baseline condition stimuli, for both accent 1 and 2 stimuli, with 98 percent correct and 96 percent correct identifications respectively.

Voicing Condition	Percentage of Correctly Perceived Accents	
	Accent 1	Accent 2
Normal voiced	0.983	0.959
Natural whisper	0.632	0.597
Resynthesised whisper	0.680	0.489

Table 1. The Percentage of correctly perceived accents for each acoustic condition.

On the other hand, the two other test conditions show interesting disambiguation results. In natural whisper, subjects were slightly better at perceiving accent 1 correctly (0.63 correct), and were only a trifle worse at perceiving accent 2 (0.59 correct). The resynthesised whisper condition reveals some interesting results. Subjects were able to correctly identify accent 1 in 68 percent of cases, while the same subjects were only able to identify accent 2 tokens 48 percent of the time.

[5] ANOVA tests were conducted upon the advice of our reviewer.

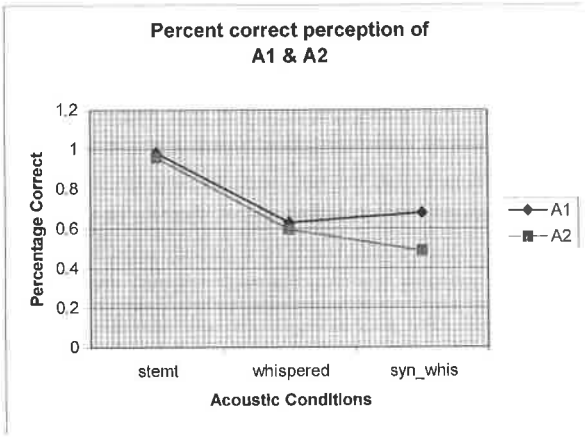


Figure 3. The percentage of correct perception of accent 1 and 2 in each of the three acoustic conditions.

As shown in Figure 3, there is a definite trend for subjects' perceptual ability to decline between conditions. However, as Figure 3 also depicts, it appears that subjects recognized accents 1 and 2 with the same degree of ability in the normal voiced and natural whisper conditions but that they found accent 1 tokens easier to recognise in the resynthesised condition than they did accent 2. Why this might appear to be the case is uncertain but the topic will be revisited below. Importantly, results with respect to specific accent types show that subjects seem to encounter general difficulty when trying to perceive either of the impoverished acoustic conditions.

To further answer some of our questions with a more specific measure, we investigated whether subjects can actually be said to discriminate correctly between the two pitch accents, and to see whether the condition variable influenced the degree of correct perception of the intended meaning. To this end we employed a d' test, which is based on the z-score of the normal distribution and measures the effect of the independent variable on the dependent variable. The test results show that subjects were able to make the correct identifications in both whispered conditions, with $d' = 2.02$ for natural whisper, and $d' = 2.8$ for resynthesised whisper ($p < 0.05$, with $p = 0.05$ when $d' = 1.96$). As was mentioned above, this is an issue about which previous research (including Fintoft's) does not

provide firm conclusions. A Student's t-test showed that discrimination was better in voiced speech than in natural whisper ($p < 0.05$), and better in natural than in resynthesised whisper ($p < 0.001$). With these results we can confirm that speakers of East Norwegian do indeed discriminate between tokens that are traditionally seen as differing only in pitch accent, even when f_0 is demonstrably absent.

We have now confirmed with statistical evidence the hitherto intuitive claim that speakers of East Norwegian do not need to rely on context to disambiguate pitch accent in the absence of pitch information, that is, in whispered speech. The statistical significance of our results leads to a further conclusion that there must be other phonetic cues present in the speech signal which are utilised for disambiguating between two possible lexical meanings in place of prosodic information. This is an important finding, and should guide further research in this field.

An interesting finding of this study, identified in the tests above, regards the resynthesised whisper results. We saw that the correct identification rate for these stimuli was significant, and more often correctly identified than the results for the naturally whispered stimuli. This result suggests that we should not expect to find that speakers are dependent on *additional* cues in whispered speech that are not also present in voiced speech. There is simply no evidence for the dependence of such extra cues. Moreover, if these extra cues had been present and actively utilised in utterance interpretation, the perception of pitch accent differences in resynthesised whisper should be impossible. On the contrary, this is fully possible, demonstrating that there is sufficient information other than f_0 in voiced speech. This is obviously not evidence against the presence of extra cues in whispered speech. What we have shown is that speakers do not depend crucially on such cues. In order to firmly decide the matter, our entourage of acoustic cues should include at least two further conditions. By researching subjects' perception in both an additional purely synthetic voiced and synthetic whisper condition we would be able to determine the nature of accent structure with regards to potential acoustic cues. This task is left for future research.

Finally, as mentioned above and revealed in Figure 3, the percentage of correctly recognised accent one tokens increased in the resynthesised condition. Fintoft (1970) reports that whispered accent two tokens that had been spoken by a Trondheim speaker[6] were frequently identified as toneme one. While Fintoft

[6] The second author, who provided the stimuli for this experiment, is a native speaker of the Trondheim dialect.

(1970) does not provide any conclusive explanations for this outcome, Hadding-Koch (1962) suggests that certain dialect speakers may less accurately substitute other acoustic cues when f_0 was lacking. This finding is in line with the proposal for future research previously mentioned in which a detailed analysis of the durational, intensity and stress related components of each stimulus in each acoustic condition precedes experimentation.

6 Conclusion

This paper has addressed the issue of speaker reliance on phonetic cues for utterance disambiguation when otherwise necessary pitch information is absent, such as during whispered speech. Results from a perception test of speakers of East Norwegian show conclusively that speakers can discriminate between pairs of words traditionally described in terms of purely intonational difference, also in the absence of intonational information. This suggests that there may be phonetic cues other than f_0 present in the speech signal which speakers may rely upon. Furthermore, tests using resynthesised whisper, demonstrate that speakers do not require extra phonetic cues present in whispered speech to compensate for the loss of f_0 information. Rather, adequate cues for correct interpretation during whisper must be present in the voiced speech signal.

Acknowledgements

The authors would like to thank Dr. Henning Reetz, Dr. John Coleman and Dr. Esther Grabe for advice and assistance during the revision process.

Appendix

Stimuli presented in the perception test. Each first line represents an incomplete utterance, and the two following lines represent the possible completions.

La oss håpe at gangen…
Let us hope that the corridor/the gait...

er stor nok til at vi kan henge fra oss frakkene der.
is big enough for us to leave our coats there.
avslører tyven på videoopptaket.
will reveal the identity of the thief on the CCTV.

Du må fortelle meg om tanken…
Could you tell me whether the tank/the thought…
kom plutselig, eller om du hadde vurdert det lenge.
came suddenly, or whether you had been considering it for a while.
kan fylles med vanlig bensin.
can be filled with ordinary petrol.

Det er helt sikkert at [1]hakket/[2]hakke…
It is certain that the cut/a pick-axe…
ble laget med en skarpt gjenstand.
was made with a sharp object.
bør være i enhver brevandrers oppakning.
should be part of the luggage of everyone crossing on a glacier.

Jeg håper noen finner…
I hope that someone will find…/some Fins…
kan ta med litt vodka til festen.
can bring some vodka to the party.
klokka som forsvant.
the-watch that disappeared.

Turistbrosjyra ga oss informasjon om [1]faret/[2]fare…
The leaflet gave information about the track/the dangers…
forbundet med fjellklatring.
connected with mountaineering.
som gikk mellom Tydal og Storerikvollen.
connecting Tydal and Storerikvollen.

Jeg hørte at noen [1]bønder/[2]bønner…
I heard that some farmers/beans…
har reist til Oslo for å protestere.
have gone to Oslo to protest.

har blitt trukket tilbake av Rema fordi de var forgiftet.
have been withdrawn by Rema because they were harmful.

Vi er helt avhengige av at en eller annen skriver...
We are fully dependent on someone writing/that some registrar...
er tilstede under seremonien.
is present during the ceremony.
et brev om dette til avisa.
a letter about this to the newspaper.

References

Caspers, J. 2000. Experiments on the meaning of four types of single-accent intonation patterns in Dutch. *Language and Speech* 43, 127-161.

Catford, J. 1977. *Fundamental Problems in Phonetics*. Edinburgh, Edinburgh University Press.

Fintoft, K. 1970. *Acoustical Analysis and Perception of Tonemes in Some Norwegian Dialects*. Oslo, Universitetsforlaget.

Hadding-Koch, K. 1961. *Acoustico-Phonetic Studies in the Intonation of Southern Swedish*. Lund, Gleerup.

Hadding-Koch, K. 1962. 'Notes on Swedish Word Tones'. *Proceedings of the International Conference of Phonetic Sciences*. Helsinki, The Hague, Mouton: 630-638.

Ladefoged, P. 2001. *A Course in Phonetic*. New York, Harcourt Brace.

Lass, N. J. 1996. *Principles of Experimental Phonetics*. St Louis, Mo, Mosby Inc.

Meyer-Eppler, W. 1957. Realization of Prosodic Features in Whispered Speech. *Journal of the Acoustical Society of America* 29: 104-106.

Nilsen, R. A. 1992. *Intonasjon i interaksjon: sentrale spørsmål i norsk intonologi,* unpublished dr.art. dissertation, University of Trondheim, Norway.

Saunders, R. 2002. Online materials for the phonetics course at the Simon Fraser University, Burnaby, British Columbia, Canada, http://www.sfu.ca/~saunders/l33098/L4/L4_6.html

Tartter, V. 1991. Identifiability of vowels speakers of whispered syllables. *Perception and Psychophysics* 49: 365-372.

Thomas, I.B. 1969. Perceived pitch of whispered vowels. *Journal of the Acoustical Society of America* 46: 468-470.

French phrasing and accentuation in different speaking styles

BRECHTJE POST

1 Introduction

A number of interacting factors determine which syllables are accented in an utterance, such as word grouping, grammatical category, and speaking style. Two questions are addressed in this paper: (1) how does speaking style affect phrasing and accentuation in French, and (2) can the account proposed in Post (2000), which uses partial ranking to model prosodic variation, adequately describe these data? Recordings of two Map Tasks were analysed auditorily and acoustically, and compared with earlier findings for read speech. The results support the account, and show that speakers produced roughly equal numbers of phrases, but considerably fewer accents. The findings not only allow us to evaluate the explanatory power of partial ranking, but also have implications for a phonological account of French intonation. If clear predictions can be made about the locations of pitch movements in the utterance, the number of intonation contours that can be realised is restricted.

1.1 Prosodic variation and partial ranking

Post (2000) proposes a constraint-based account of the complex interaction between phrasing and accentuation in French, in which a number of universal well-formedness constraints are ranked relative to each other in the grammar (e.g. NoClash: two immediately adjacent accented syllables are prohibited, and RightmostPWd: Prosodic Words must have final accents). The surface form that best satisfies the higher-ranked constraints is the one that is selected (Prince and Smolensky 1993). For instance, the constraint hierarchy would select *de JOlis AIRS* 'pretty tunes' rather than *de joLIS AIRS* with a clash.

The hierarchy proposed in Post (2000) describes a 'default' situation, that is, the patterns of phrasing and accentuation that are most frequent in careful speech

which is produced at a normal rate (e.g. Delais 1995, Verluyten 1982). However, a wide range of alternative realisations is possible. In Post (1999), for instance, potential clash items such as *de jolies airs* were sometimes realised with only one accent (*de jolies AIRS)*, instead of the 'default' two (*de JOlis AIRS*). Variation in the realisation of accents and phrases depends on a number of factors such as speaking rate and style, but little is known about this kind of variation in French (but see e.g. Bruce and Touati 1990a,b, Lacheret-Dujour 2002, and Martin 1999).

Post (2000) proposes to account for prosodic variation in French by means of partial ranking, which allows small subsets of constraints to be unranked relative to each other, while they are still crucially ranked relative to the other constraints in the hierarchy (Anttilla 1997). The unranked constraints are the ones which regulate, for instance, the widening of a domain of application of a post-lexical process at a higher speaking rate. For example, instead of aligning with an X' projection, a Phonological Phrase could align with an X" projection (giving [*un verger vert*]PP instead of [*un verger*]PP [*vert*]PP 'a green orchard'; cf. Selkirk 1995, Delais 1995). A postlexical process that has the phonological phrase as its domain of application, such as Clash Resolution, will now apply within this enlarged domain (giving [*un VERger VERT*]PP instead of [*un verGER*]PP [*VERT*]PP where the two accents are separated by a PP boundary; Post 1999). This can be captured by reranking the two constraints regulating Phonological Phrase alignment, which are immediately adjacent in the constraint hierarchy, as is shown in Tableaux 1 and 2 (only the constraints that are relevant to this example have been given in the tableaux).

Tableau 1 represents the ranking of the constraints for the 'default' situation, where monosyllabic complements form a Phonological Phrase with the preceding X' head (i.e. *vert* groups with *verger* in this example).

|un verger vert| 'a green orchard'	AlignX"	AlignX'	NoClash	RightPWd
☞a. [un (verger)(vert)]PP		*		*
b. [un (verger)]PP [(vert)]PP	*!		*	
c. [un (verger)(vert)]PP		*	*!	

Tableau 1: Align X" is ranked above Align X', giving [*un VERger VERT*]PP (indicated by ☞).

Here, output candidate (b), in which both X' heads project a Phonological Phrase, falls at the first hurdle, because it violates the constraint AlignX". That is, AlignX" requires the alignment of a Phonological Phrase with (all and only) X" projections. The 'fatal violation' is indicated by '*!' in the tableau. When we move on to the next column, which represents the next constraint down in the hierarchy in this ranking, we see that candidates (a) and (c) both violate AlignX', since *un verger* does not form a PP on its own. This means that the constraint AlignX' cannot distinguish between these candidates. In fact, it is the immediately adjacent accents on *-ger* and *vert* which rule out candidate (c), as can be seen in the third column with the constraint NoClash, which prohibits clashes between adjacent accents. Thus, even though candidate (a) violates as many constraints as candidates (b) and (c), only one candidate is selected as the optimal form, because the constraints are crucially ranked relative to each other.

|un verger vert| 'a green orchard'	AlignX'	AlignX"	NoClash	RightPWd
a. [un (verger)(vert)]PP	*!			*
☞b. [un (verger)]PP [(vert)]PP		*	*	
c. [un (verger)(vert)]PP	*!		*	

Tableau 2: Align X' is ranked above Align X'',
giving [*un verGER*]PP [*VERT*]PP

In Tableau 2, the ranking of the partially ranked alignment constraints is reversed, and this represents a speaking style in which all X' heads form phrases

on their own, resulting in more phrases for the same lexical material. As a consequence of the reranking, candidate (a) violates the topmost constraint instead of (b), and candidate (c) is also immediately eliminated. Tableau 2 shows that when AlignX' outranks AlignX", and therefore applies in full force, all X' words are aligned with a Phonological Phrase boundary on their right-hand side. This leads to the selection of the output form for *un verger vert* that has a Phonological Phrase break between the lexical words and two word-final accents.

This minimal adjustment of the constraint hierarchy does not affect the selection of the optimal candidates for structures such as *de jolis airs* and *des hivers autres qu'en Afrique* 'winters other than (those) in Africa', that is, for structures with different numbers of syllables and X' or X" heads. In other words, reranking within the subset of the partially ranked alignment constraints only changes the output for a small and clearly defined subset of cases (see Post (2000) for a full description of the account, which also describes variation in accentuation which is independent of variation in phrasing, as outlined in (4) below). Therefore, the model has the advantage that it gives a unified account of the data, making clear predictions about the forms that surface in 'default' cases and in cases of prosodic variation, while it still excludes ungrammatical forms. However, since little is known about this type of prosodic variation, the question is whether the predictions of the model are correct.

The aim of this paper is therefore to establish (1) how prosodic phrasing and accentuation vary as a function of speaking style in French, and (2) whether the account of prosodic variation proposed in Post (2000) is indeed adequate to describe the variation we observe.

2 Methods

Recordings of two Map Tasks (Brown et al. 1984) were analysed auditorily and acoustically, and compared with earlier findings for read speech (Post 1999).

2.1 Materials

The Map Tasks were specifically developed to test the interaction between phrasing and accentuation in (semi-)spontaneous speech in French. More specifically, a number of landmarks were included which allowed a comparison

with the patterns that have been observed in careful, usually read speech. These 'default' patterns can be summarised as follows:

(1) The 'default' accentual patterns
- (a) All Phonological Phrases (PP) are obligatorily marked by a final pitch accent ([*des verGERS*]PP; Verluyten 1982, Post 1999, 2000)
- (b) PPs align with X' heads ([*des vergers*]PP; Nespor and Vogel 1986), except for
  - heads followed by a monosyllabic direct complement (align with X" [*des vergers verts*]PP; Post 1999, 2000)
  - prenominal adjectives (align PP with following noun [*de jolis airs*]PP; (Nespor and Vogel 1986, Selkirk 1986, Verluyten 1982)
- (c) PPs have additional accents
  - on the final syllable of each Prosodic Word ([*de peTITS enFANTS*]PP 'small children'; Delais 1995), unless this creates a clash ([*de JOlis AIRS*]PP; Post 1999)
  - Prosodic Words with more than two syllables are also accented on the first syllable (*la NEcessiTE*; Pasdeloup 1992, Di Cristo and Hirst 1997).[1]

The starting point of the account is that the Phonological Phrase (PP) is the domain of pitch accent distribution in French, regardless of factors such as speaking rate or style (Verluyten 1982, Post 1999, 2000). That is, the pitch accent at the right edge of the PP is obligatory and other pitch accents are optional. Since Phonological Phrases are derived from the syntactic constituent structure by means of an algorithm (Nespor and Vogel 1986, Selkirk 1986), the claim that every PP must end in an accent can be verified (see Post 1999 for a discussion).

Post (1999) tested the application of Clash Resolution in domains that formed a Small PP (derived from an X' head) as in (2), or two Small PPs which together can form a Maximal PPs (derived from an X" head) as in (3) (cf. Selkirk 1986, 1995). The formation of MPPs is optional, depending on, for instance, rate of speech. Therefore, Clash Resolution is only obligatory in these contexts when an MPP has actually been formed.

[1] When words with more than three syllables start in a vowel, the accent tends to be realised on the second syllable instead of the first, e.g. *l'impossibilité* (Pasdeloup 1992). Since there was only one example of this pattern in the present study, it will not be discussed here.

(2) X': [*de JOlis AIRS*]SPP
(3) X": [*de VERgers VERTS*]MPP or
X' and X': [*de verGERS*]SPP [*VERTS*]SPP

Since acoustic measurements confirmed the auditory judgements of accentuation and phrase formation (i.e. which SPPs had been grouped into an MPP), we can conclude that Clash Resolution is a good indicator of phonological phrasing. In order to verify whether the algorithm for phrase formation holds for spontaneous speech, and whether PPs indeed have an obligatory final accent, a number of items with these structures were included in the Map Tasks, thus testing claims (1a) and (1b) above.

These items were complemented with landmarks that did not have a clash context (as in (1c) above), so that variation in accentual patterns as a function of speaking style could be examined. Similarly, landmarks with longer X" projections were included, because the formation of MPPs appears to be sensitive to factors such as the number of syllables in the complement of the first X' category word, e.g. plurisyllabic *balnéaire* in *la station balnéaire* 'the seaside resort' (Post (1999, 2000) found that MPP formation does not depend on whether the complement is branching, as Nespor and Vogel (1986) propose).

Variants of the 'default' patterns that I expected to observe are given in (4).

(4) Variation on the 'default' patterns
  (a) PPs only align with X' heads, never with X" alone [*des vergers*]PP [*verts*]PP; e.g. Post 1999)
  (b) PPs also form when the complement contains more than one syllable ([*la station balnéaire*]PP; Post 1999)
  (c) a 'hammock' pattern: additional accents are realised on the first syllable of the first Prosodic Word in the PP, rather than on word-final syllables ([*de PEtits enFANTS*]PP ; e.g. Hirst and Di Cristo 1984, Mertens 1992, Pasdeloup 1992, Delais 1994, Di Cristo 2000)
  (d) all non-final accents are omitted ([*de petits enFANTS*]PP; e.g. Post 1999).

2.2 Subjects and procedure

Two native speakers in their early twenties, both from Paris, took turns as Instruction Giver and Instruction Follower, so that a reasonably large speech sample could be obtained for each subject. The recordings were made on DAT-tape in the sound-treated studio of the Department of Language and Speech at the University of Nijmegen, and digitised at 16KHz.

2.3 Auditory and acoustic analysis

136 speech samples with structures that were appropriate to test the patterns described in (1) and (4) were selected; most, but not all, were realisations of the landmarks. They were divided into different groups, depending on their morpho-syntactic structure (potential MPP or SPP), and the number of Prosodic Words and syllables (giving contexts for the realisation of additional accents or Clash Resolution). Then, the samples were analysed auditorily to establish the location of accents and phrase boundaries. Judgements such as these were shown to be highly reliable in Post (1999).

The data were analysed acoustically by means of the PRAAT signal processing package (Boersma and Weenink 1996) to obtain independent evidence to support the judgements. The measurements were taken in a wide-band spectrogram in three vowels, as exemplified in (5): (V1) the first vowel of the first Prosodic Word in the PP, (V2) the final vowel of the prefinal Prosodic Word in the PP, and (V3) the final vowel of the PP (the syllables containing the three vowels are underlined). The second formant was used to identify the start and end points of the vowels.

(5)(a) [*la station balnéaire*]PP or [*la station*]PP [*balnéaire*]PP
V1 V2 V3 V1 V2 V3

(b) [*un petit peu tortueux*]PP or [*un petit peu*]PP [*tortueux*]PP
V1 V2 V3 V1 V2 V3

'a little bit winding'

Duration measurements were taken to verify whether pre-final lengthening supports the judgements of PP boundaries. Fundamental frequency measurements were taken to verify whether the auditory judgements of the accentual patterns were reflected in the pitch trace. Since non-final accents are generally assumed to

be characterised by a high accent in French (H*; e.g. Delais 1995, Jun and Fougeron 1995, Post 2000, 2002), fundamental frequency should normally be higher for accented than for unaccented syllables (exceptions in which V1 and V2 were located on a falling slope from a preceding higher accent were excluded).

71 of the 136 speech samples could not be analysed acoustically, because the PP-final accent was low instead of high (before an utterance boundary), or because one of the vowels could not be measured. In *un petit peu*, for instance, the schwa of *pe* was often omitted altogether when it was not accented, and therefore, the sample did not in fact match the structures of interest described in (1) and (4) above.

3 Results

3.1 The algorithm for PP formation and PP final accents

The auditory judgements showed that of the 51 speech samples that contained an X" projection, the Phonological Phrase aligned with X" in 15 cases (forming 15 MPPs), and with both X' categories in 36 cases (resulting in 62 SPPs). In the remaining 85 samples that contained only one X' category word, the SPP aligned with X', as expected. Except for one realisation of *les torrents*, where the initial and final syllables sounded equally prominent, all PPs had a final accent.

A comparison of the durations measured for the 90 PP-final vowels and the 119 non-final vowels of all PPs in the sample supported the judgements of phrasing (means: 90ms vs. 163ms.; $T(120.44)=-10.05$, $p<0.001$). However, there are several confounding factors that might bias the result in the direction of our hypotheses. Different vowels have intrinsically different durations, and if 'short' vowels happen to predominate in the non-final group, this would bias the results. However, I did not find any evidence to support this. More importantly, all final vowels were judged to be accented (except one), but only 42 of the 119 non-final vowels that were measured were accented, confounding the lengthening effect of an accent with that of a PP boundary. Therefore, all non-accented vowels were excluded in a second analysis. The difference in duration between non-final and final vowels was still highly significant (102ms. vs. 171ms. $T(119.3)=-7.81$, $p<0.001$).

As a second check on the judgements of PP boundaries, the application of Clash Resolution (CR) was investigated, which had proven to be a good indicator for phonological phrasing in a previous study (Post 1999). 62 of the 136 speech samples contained a CR context, where two word-final accents would be immediately adjacent within the PP unless the first accent is omitted or realised elsewhere.

50 of those 62 samples were of the *de jolis airs* type, which had only one X' head that can align with an SPP. As predicted, CR always applied in these cases. The remaining 12 cases, of the *des vergers verts* type, could form two SPPs or one MPP. Only in the latter case, CR would apply; in the former, two PP final accents would be required. With one possible exception, this was indeed what was found. The mean fundamental frequency measured in V1, V2 and V3 for these samples is given in Figure 1. Fundamental frequency is averaged over samples in which (1) V1 and V2 were judged to be equally prominent, (2) V1 was more prominent than V2 (initial accent), or (3) V2 was more prominent than V1 (final accent; only 5 cases). The values measured in Post (1999) are given in the right-hand panel for comparison.

The most striking difference is that the mean fundamental frequency values in the Map Task data are higher and cover a wider range, but this is largely attributable to the fact that unlike the Map Task data, the read speech data contain measurements from both male and female speakers. Also, only two speakers provided the data for the Map Task measurements, so further data would be needed before any generalisations can be made. Another difference is that in the read speech data, the judgements are clearly reflected in the fundamental frequencies measured for V1 and V2, whereas in the Map Task data, the values for V1 and V2 are much closer together. Again, it is difficult to generalise, but we can verify whether V1 and V2 are significantly different for the subset in which V1 sounds more prominent than V2 (labelled 'initial accent' in the graph).

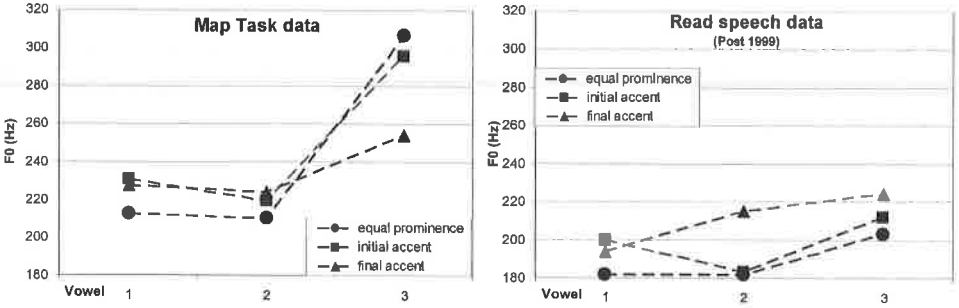


Figure 1: Fundamental frequency measurements taken in the Map Task data (left panel) and read speech data (right panel) for samples with equal prominence on V1 and V2 (circles), an initial accent on V1 (squares), and a final accent on V2 (triangles).

A paired samples t-test confirmed the CR judgements (mean V1 231Hz. and mean V2 219Hz. with $T(16)=3.00$, $p<0.01$ for initial accent cases; when equal prominence and initial accent cases are collapsed: mean V1 227Hz. and mean V2 217Hz.; $T(20)=2.95$, $p<0.01$). Unfortunately, there were not enough cases with a final accent for which f0 measurements could be taken to make a meaningful statistical comparison between cases in which CR had applied and those in which it had not (only 6 cases).

3.2 Variation in Phonological Phrase formation

The aim of the second part of the analysis was to establish whether speaking style has an effect on phonological phrasing. The results are given in Figure 2. As mentioned, 51 speech samples contained two X' projections which could optionally form an MPP. In 41 of those the complement was plurisyllabic (*la station balnéaire*), and in 10 it was monosyllabic (*le verger vert*). Monosyllabic complements phrased with the preceding material more often than plurisyllabic ones (50% versus 25% of cases). Whether the complement was branching (contained more than one word) or not did not have an effect (25% and 23% formed MPPs, respectively).

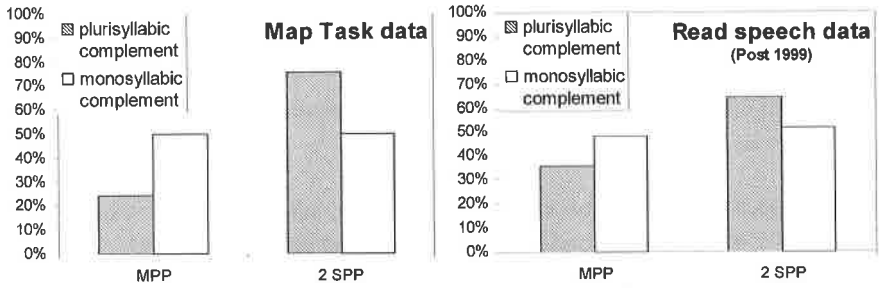


Figure 2: Phonological Phrasing in the Map Task data (left panel) and the read speech data (right panel). MPP: one PP is aligned with an X" projection containing two X' heads; 2 SPP: two PPs are aligned with an X" projection containing two X' heads (i.e. alignment with X' projections).

These findings are quite similar to those reported for read speech in Post (1999, 2000). Here, 37% of plurisyllabic and 48% of monosyllabic complements grouped with the preceding material. It should be noted, though, that these data are not directly comparable, since they were collected in a production experiment that tested the application of CR and Liaison.

3.3 Variation in accentuation

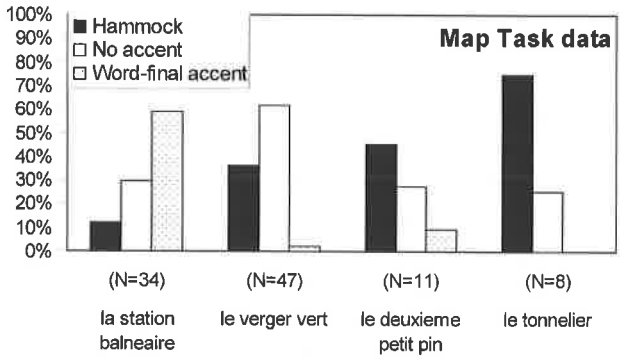
The final set of analyses investigated the realisation of accentual patterns in the Map Task data. 100 PPs were selected which contained enough material to accommodate two accents without necessarily creating a clash, as specified in (6).

(6) Structure of the samples testing accentual variation
- (a) two plurisyllabic Prosodic Words (e.g. [*la station balnéaire*]PP)
- (b) a plurisyllabic word followed by a monosyllabic word (e.g. [*le verger vert*]PP)
- (c) more than two words (e.g. [*le deuxième petit pin*]PP 'the second small tree')
- (d) one word with more than two syllables (e.g. [*le tonnelier*]PP 'the cooper')

The 'default' pattern for the structures in (6a) and (6c) is a sequence of word-final accents, while initial accents would be realised in (6b) and (6d) (as specified in (1c) above). Accentual patterns such as these have been observed to vary in two ways. Either, the PP is marked by an initial and a final accent only ('hammock'), or all non-final accents are omitted (see (4) above). The results are shown in Figure 3.

Although a total of 32 of the 100 Map Task samples showed a 'hammock' pattern, only 9 of those occurred in contexts (6a) and (6c), which means that in the vast majority of cases, hammocks occur in a context in which there is only one position available for the second accent in the phrase. This means that when there is a choice, the second accent tends to occur in word-final position (21 of the 45 cases in (6a) and (6c), and 22 cases in all).

The alternative is not to produce any non-final accents at all (44 cases in all). This is the preferred option in clash contexts like (6b) (29 cases). For comparison, the non-final accent was omitted in only 14% of all CR cases in the read speech data of Post (1999). Even in structures where there is more material (6a and 6c), the omission of the non-final accents is quite popular (13 of 45 cases).



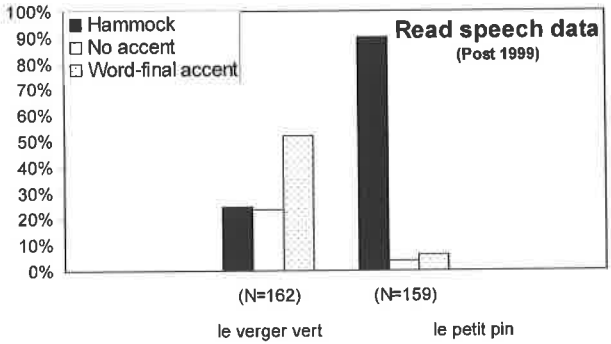


Figure 3: Non PP-final accents observed in four types of structures in the Map Task data (top), and similar structures in the read speech data (bottom). Black bar: accent on initial syllable of first prosodic word (hammock), white bar: no non-final accent, and dotted bar: accent on final syllable of first prosodic word.

4 Conclusion

The aims of this study were to investigate (1) how speaking style affects prosodic phrasing and accentuation in French, and (2) whether the constraint-based account of prosodic variation proposed in Post (2000) is indeed adequate to describe this variation.

The results show that all Phonological Phrases are marked by a final pitch accent, regardless of speaking style, and that PPs either align with X' or X" projections. This finding confirms the starting points of Post (2000). The results also show that speakers produced roughly the same number of phonological phrases as were observed for comparable data in read speech, and they produced considerably fewer accents in non-final position in the PP, where they are optional.

These findings can easily be accommodated in the model of Post (2000). Partial ranking has the advantage that prosodic variation that arises within the language system of one speaker is accounted for by essentially the same, minimally modified, grammar. However, it is not clear what mechanism should trigger a reranking of constraints. Although differences in the number of occurrences of a particular prosodic pattern were found in the speaking styles investigated here, all patterns were realised in both styles; a change in speaking style in itself does not seem to concur with a reranking of a particular pair of constraints.

Clearly, further studies are required with more speakers and a wider variety of structures, in which different prosodic domains and factors other than speaking style are also taken into consideration (cf. Lacheret-Dujour 2002, Martin 1999).

References

Anttila, A. 1997. Deriving variation from grammar. In Hinskens, F., R. van Hout, and L. Wetzels (eds), *Variation, change and phonological theory*. Amsterdam: Benjamins: 35-68.

Boersma, P. and D. Weenink. 1996. PRAAT: A system for doing phonetics by computer. *Report of the Institute of Phonetic Sciences, University of Amsterdam* 132.

Brown, G., A. Anderson, D. Shillcock, and G. Yule. 1984. *Teaching Talk. Strategies for production and assessment.* Cambridge: CUP

Bruce, G. and P. Touati. 1990a. Auditory and acoustic analysis of dialogue prosody in Swedish and French. *Phonum* 1: 27-30. Umeå University, Department of Phonetics.

Bruce, G. and P. Touati. 1990b. On the analysis of prosody in spontaneous dialogue. *Working Papers* 36. Lund, Department of Linguistics: 37-55.

Delais, E. 1994. Rythme et structure prosodique en français. In C. Lyche (ed) *French generative phonology: Retrospective and perspectives*: 131-150. Salford: AFLS/ESRI.

Delais, E. 1995. *Pour une approche parallèle de la structure prosodique.* Doctoral dissertation, Université de Toulouse-Le Mirail.

Di Cristo, A. 2000. Vers une modélisation de l'accentuation du français. *Journal of French Language Studies* 10: 27-44.

Di Cristo, A. and D. Hirst. 1997. L'accent non-emphatique en français: Stratégies et paramètres. In Perrot, J. (ed) *Hommages I. Fónagy*: 71-101. Paris: l'Harmattan.

Hirst, D. and A. Di Cristo. 1984. French intonation: A parametric approach. *Die Neueren Sprachen* **83** (5): 554-569.

Jun, S.-A. and C. Fougeron. 1995. The accentual phrase and the prosodic structure of French. *Proceedings International Congress of the Phonetic Sciences* 13 (2): 722-725.

Lacheret-Dujour, A. 2002 . The intonational marking of topical salience in spontaneous speech: Evidence from spoken French'. *Proceedings Speech Prosody 2002*, Aix-en-Provence. 2002.

Martin, P. 1999. Intonation of spontaneous speech in French. *Proceedings ICPhS 1999*: 17-20.

Mertens, P. 1986. L'accentuation de syllabes contiguës. *I.T.L.* 95/96: 145-163.

Nespor, M. and I. Vogel. 1986. *Prosodic phonology*. Dordrecht: Foris.

Selkirk, E. 1986. On derived domains in sentence phonology. *Phonology Yearbook* 3: 371-405.

Selkirk, E. 1995. The prosodic structure of function words. In Beckman, J., L. Walsh Dickey & S. Urbanczyk (eds) *Papers in Optimality Theory*: 439-469. Amherst: GLSA.

Pasdeloup, V. 1992. A prosodic model for French text-to-speech synthesis: A psycholinguistic approach. In Bailly, G., C. Benoit and T. Sawallis (eds) *Talking machines: Theories, models, and designs*: 335-348. Amsterdam: Elsevier Science Publishers.

Post, B. 1999. Restructured Phonological Phrases in French. Evidence from Clash Resolution. *Linguistics* 37 (1), 41-63,

Post, B. 2000. *Tonal and phrasal structures in French intonation*. Doctoral dissertation. The Hague: Holland Academic Graphics.

Post, B. 2002. French tonal structures. In B. Bel and I. Marlien (eds), *Proceedings of the Speech Prosody 2002 conference*. Aix-en-Provence: Laboratoire Parole et Langage: 583-586.

Prince, A. and P. Smolensky. 1993. *Optimality Theory: Constraint Interaction in Generative Grammar*. Cambridge, Mass.: MIT Press.

Verluyten, S. 1982. *Recherches sur la prosodie et la métrique du français*. Doctoral dissertation, Universiteit Antwerpen.

Prosody, Memory Load, and Memory for Speech

BURTON S. ROSNER, ESTHER GRABE, HANNELE NICHOLSON, KEITH OWEN AND ELINOR KEANE

---

1 Introduction

Previous experiments indicate that prosody may facilitate memory for spoken utterances. Memory load, however, has not been systematically manipulated and may interact with such facilitating effects. Therefore, subjects were asked to listen to connected prose passages between one to five intonation phrases long and to repeat as much of each narrative as possible. Stimuli were normal productions or else had been deprived of intonational variation, pauses, or both. Performance on both the normal and the prosodically altered passages decreased as passage length grew. Errors were overwhelmingly omissions and occurred mainly in the middle of a passage. The prosodic manipulations affected memory for the longest passages. These results alongside previous findings suggest that prosody facilitates memory for spoken stimuli that are relatively difficult to process. Lengthy material, infrequent words, unrelated items, citation form prosody, or thematic or grammatical anomalies may create such difficulties.

2 Background

Numerous experiments show that prosody affects spoken word recognition, the computation of syntactic relationships, and the processing of discourse structure (see Cutler, Dahan, & van Donselaar, 1997, for a review). Other work indicates that disruption of prosody can interfere with memory for spoken language (Darwin, 1975; Wingfield, 1975a, 1975b; Wingfield & Klein, 1971; Stine & Wingfield, 1987; Wingfield, Lahar, & Stine, 1989; Paris, Thomas, Gilson, & Kincaid, 2000) In these studies, however, memory load was not systematically manipulated. Since speech contains numerous redundant cues, prosody may be unimportant for remembering relatively simple, short utterances. As processing

demands grow, however, prosody may increasingly affect memory for connected prose. To test this hypothesis, we examined the influence of intonation and of pauses on memory for continuously spoken passages of differing lengths.

Leonard (1974), O'Connell, Turner, and Onuska (1968) and Zurif and Mendelsohn (1972) all reported that monotone speech produced decrements in memory for spoken items. We therefore eliminated all intonational variation in the first of our three conditiona. Using a resynthesis technique, F0 was held constant throughout an entire narrative.

The results of Huttenlocher and Burke (1972), Frankish (1985, 1989), Saito (1998), and Martin (1968) indicate that pauses, appropriate or otherwise, do affect memory for speech. Furthermore, manuals on public speaking (e.g., Mandel, 1993; Berry, 1994) discuss the importance of putting pauses of the right lengths at the right places, in order to help the listener's comprehension. Presumably, comprehension of running speech would call on mnemonic resources. For our second experimental condition, then, we removed all naturally occurring pauses from continuously spoken prose. In a third condition, we eliminated both intonational variation and pauses from spoken narratives. This procedure seemed to promise the greatest disruption of memory for prose.

3 Method

Stimuli. The stimuli were based on six-word utterances. Each basic utterance comprised one intonation phrase (IP). A complete passage contained between one and five such basic items and formed a connected narrative about a single topic. One six-word constituent of a passage was drawn from the revised Harvard Sentences (Institute of Electrical and Electronic Engineers, 1965). We added whatever further narrative material was necessary. When a passage contained two or more basic utterances, the insertion of an initial 'and' made the final part seven words in length. The 1-IP passages therefore had six words, the 2-IP passages had 13 words, the 3-IP passages 19 words, the 4-IP passages 25 words, and the 5-IP passages 31 words.

Two sets of passages, labelled set I and set II, were constructed. Each set was ordered to begin with two 1-IP sentences. They were followed consecutively by one 2-IP narrative, a 3-IP narrative, a 4-IP narrative, two 5-IP narratives, one 4-IP narrative, a 3-IP narrative, a 2-IP narrative, and finally two 1-IP sentences. Each set of 12 narratives therefore contained four 1-IP passages and two of each of the

2-IP, 3-IP, 4-IP, and 5-IP types. Example sentences, with accented words in upper case, are: 1-IP) Men STRIVE but seldom get RICH; 2-IP) The GIRL gave no clear RESPONSE, and the MEDICS went straight to WORK; 3-IP) SOME ads serve to cheat BUYERS; read EACH with VERY great care, and look CLOSELY at any SMALL print; 4-IP) Pink CLOUDS floated with the BREEZE, the SUN was setting at NIGHTFALL, the SKY slowly turned deep BLUE, and EVERY street light began to GLOW; 5-IP) The MAP shows where we ARE; the BAG contains something to EAT; your CLOTHES are in the SUITCASE; the TENT is in the BOOT; and the CAR is full of PETROL. After the experiment was under way, we found that one 3-IP passage in set II contained only 18 rather than 19 words, while a 4-IP passage in that set contained 26 rather than 25 words.

A male native speaker of Southern British English produced all 24 passages, which were recorded on a CD. In the Appendix, words accented during recording appear in capitals. (See Ladd, 1996, Ch. 6, for a discussion of accentuation.) For recording, we used an Audio-Technica AT4031 cardioid microphone and a Symetrix SX202 phantom power supply and preamplifier. The preamplifier output drove a HHB Communications CDR-850 compact disk recorder.

The Cool96 editing program (Syntrillium Software Corporation, 1996) running on a PC was used to digitize each spoken passage at 22050 16-bit samples per second and to normalize its maximum amplitude to 100 per cent. Each normal passage was then stored to the computer hard disk as a separate .WAV file. Next, we employed PRAAT 3.8 (Boersma & Weenink, 1996), a speech processing program, to make three variants of each normal passage: monotone, pause-free, and monotone-pause-free. The monotone variant was produced by PSOLA resynthesis (Carpentier & Moulines, 1990) of the passage, with F0 fixed throughout at 120 Hz. To produce the pause-free variant, we manually edited out all pauses from the normal passage. Finally, the pause-free variant was subjected to a PSOLA resynthesis with F0 again fixed at 120 Hz, generating the monotone-pause-free variant. At the end of these procedures, we had 96 .WAV files, with a normal, a monotone, a pause-free, and a monotone-pause-free version of each of the 24 spoken passages.

Subjects. Subjects were 24 undergraduates or graduate students at the University of Oxford. All were native speakers of Southern British English. Nine were male and 15 female. They were recruited by the experimenters and were paid for participating in one experimental session of about 45 min duration.

Design. Each subject heard two sets of passages. One had normal prosody, while the other was one of the three variants. Half the subjects heard the normal passages first, and the other half heard either the monotone or the pause-free or the monotone-pause-free passages first. This counterbalanced the normal/variant order. Of the 12 subjects in each of those two subgroups, six heard the normal passages in set I and variant passages in Set II, and the other six subjects heard the normal passages in set II and the variant passages in Set I. This counterbalanced normal and variant passages across sets. Altogether, the design yielded 12 conditions: two orders x three variants x two distributions of sets I and II. Consequently, two subjects underwent a given condition. One of those subjects heard the narratives in each set in the order shown in the Appendix. The other subject heard those narratives in the reverse order.

Procedure. Listeners were tested individually in a sound-attenuated recording booth. The subject faced a VDU and held a keyboard on his or her lap. A PC outside the booth controlled the experiment through a program written in C++. The subject was instructed that the experiment would be split into two parts, with a short rest period between them. In one part, a series of normally spoken passages would be heard. In the other, passages altered by a computer would be heard. The subject was asked to listen to each passage and, after the passage had finished, to repeat aloud what had been heard, as accurately as possible. The subject was encouraged to get as many words as possible, even if some words were felt to have been missed. Stimuli were presented at approximately 65 dBA over Sennheiser HD-320 earphones.

At the start of each part of the experiment, the subject's name and date of birth were entered into a data file, along with the current date and time. The filenames of the passages that the subject would hear were read in their desired order from an input file and were written to the data file. Then a message on the VDU told the subject to signal readiness by hitting the 'Enter' key on the keyboard. Upon sensing that action, the PC sent a new message to the VDU. This asked the subject to initiate a passage by pressing the 'Enter' key. When the subject did so, the VDU went blank and a passage was played after a 500 ms pause. At the end of the passage, a message on the VDU asked the subject to repeat the passage and to hit the 'F' key on the keyboard when finished. The subject's spoken response was recorded on a CD, using the system described above. After the subject pressed the 'F' key, a pause of 1 s occurred before the beginning of the next trial. The passages within each half of the experiment increased and then decreased in length, as

shown in the Appendix. This seemed to make the subject feel comfortable. At the start or the end of the experiment, the WAIS-R digit span subtest was administered to the subject.

Scoring. Using the data file for a given subject, two experimenters scored the responses recorded on CD. The number of words reproduced in the correct order was counted first, even when some intervening words had been missed. Then words reproduced out of order were identified. Finally, incorrect intrusions or substitutions were noted. The two experimenters had to agree on the scoring for each passage. This often required them to listen several times to the subject's attempted repetition of a given passage, particularly when the passage was long and the repetition became hesitant.

4 Results

Intrusions or substitutions were infrequent. They will receive no further attention. Two scores were determined for a subject's response to each passage. One measure, $P(C_O)$, was the proportion of words reproduced in the correct order and represented order information. The other measure was the proportion of words reproduced in total without regard to order, $P(C_I)$, representing item information.

Relatively few items were actually reproduced out of order. Consequently, the two measures differed very little and were often identical. Nonetheless, we subjected both measures to the same statistical treatments. Apart from one analysis, no differences emerged. We therefore report only the results on $P(C_O)$, the proportion of words reproduced in the correct order, except for the single analysis where the two measures differed somewhat. For some analyses, we combined data across the monotone, pause-free, and montone-pause-free conditions. Otherwise, comparisons would have depended on a small number of degrees of freedom, opening the way to errors of type II.

4.1 Shorter (1-IP and 2-IP) passages

The $P(C_O)$ scores for normal passages showed that the 1-IP stimuli were reproduced with no errors. The 2-IP normal passages occasioned a total of just three errors, each due to a different subject. The geometric mean for $P(C_O)$ was .995.

Across all 96 attempts at reproducing prosodically altered 1-IP narratives, the geometric mean for $P(C_O)$ was .968. One subject completely failed to reproduce one altered 1-IP passage, generating six errors, and made an error on another altered 1-IP narrative. Twelve more errors were distributed across six subjects, giving a total of 19 errors. The altered 2-IP passages again produced just three errors, distributed over two subjects, and gave a geometric mean of .995 for $P(C_O)$.

In short, the data show virtually perfect memory for 6-word and 13-word passages of connected prose. Results on both the normal and the prosodically altered 1-IP and 2-IP passages therefore were dropped from further analysis, due to lack of variance.

4.2 Longer (3-IP, 4-IP, and 5-IP) stimuli

One subject failed to reproduce anything after hearing a normal 4-IP passage and a normal 5-IP passage in set I. Another subject dried up similarly on a normal 4-IP passage in set II. Finally, one subject gave no response to a 5-IP pause-free passage. We counted each of these four failures as a response of zero. Non-zero responses were therefore made to practically all the normal and the prosodically altered stimuli.

Overall, perfect performances occurred on about 15 per cent of the 3-IP and 4-IP passages. None occurred on the 5-IP stimuli. About two-thirds of the partial recalls of the 3-IP, 4-IP, and 5-IP passages arose from omission of words in the middle of a narrative. The remaining minority of imperfect performances were largely due to omission of material either at the beginning or at the end of a passage. Rarely did a listener recall words exclusively from the middle of a passage.

Inspection of the data from both the normal and the combined prosodically altered conditions revealed some exponentially shaped or bimodal samples. No monotone transformation could render those results Gaussian. Therefore, Monte Carlo nonparametric statistics (500,000 samples per test) were used for analyses of differences. We set α at a conservative .025 for all statistical tests.

4.3 Aberrant passages and condition orders

As stated earlier, set II proved to have one 3-IP passage containing only 18 rather than 19 words and a 4-IP passage containing 26 rather than 25 words. We compared performance on each aberrant narrative against performance on its counterpart in set II with the same number of intonation phrases. This yielded two one-tailed Wilcoxon tests for $P(C_O)$ for the normal passages and two more for data combined across prosodically altered passages. None of the four tests (N=12 in each case) was significant. Accordingly, we pooled data across both 3-IP passages and across both 4-IP passages within set II as well as within set I.

Data on $P(C_O)$ were compared across condition orders (normal passages first or altered passages first) within each passage size and within each set, for the normal passages and separately for the prosodically altered stimuli combined. This resulted in 12 two-tailed Mann-Whitney tests (N=12 in total in each case) over the 3 passage sizes, 2 sets, and normal or altered passage type. None of the 12 tests proved significant, so we pooled data across condition orders.

4.4 An unexpected result: differences between sets

Box plots of $P(C_O)$ indicated that the stimuli in set I were harder to remember than those in set II. The upper panel in Figure 1 contains the data for normal 3-IP, 4-IP, and 5-IP passages, and the lower panel shows data pooled across altered passages. Given our within-subjects experimental design, we had envisioned comparing results across the two sets in order to measure the effects of the prosodic manipulations. Before making any such comparisons, however, we first had to detour into an examination of the apparent differences between sets I and II.

Two-tailed Mann-Whitney tests (N =24 for each test) showed that $P(C_O)$ for the normal 3-IP and 4-IP passages differed significantly between sets, $z = 2.410$, $p < .025$, and $z = 3.820$, $p < .001$, respectively. For data combined over the prosodically altered stimuli, the 3-IP and 4-IP passages also yielded significant differences between sets, $z = 2.410$, $p < .025$, and $z = 3.820$, $p < .001$, respectively. In keeping with these four significant differences, perfect performances occurred over 20 per cent of the time on the 3-IP and 4-IP passages of set II, normal or altered. Only some 3 per cent of the 3-IP and 4-IP passages of set I produced perfect performances. The difference between sets was not significant for the normal or for the altered 5-IP stimuli.

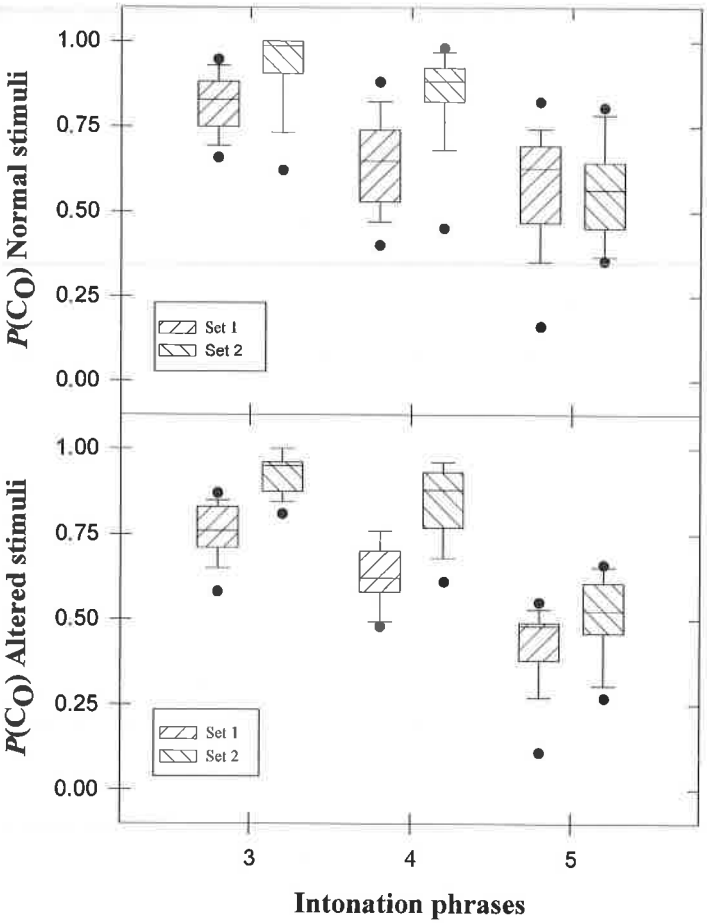


Fig. 1. Box plots of $P(C_O)$ for normal passages (upper panel) and for prosodically altered passages (lower panel) with different numbers of intonation phrases. Results are shown separately for stimuli in set I and in set II. Each box runs from the 25th to the 75th percentile. The line through the box shows the median. Whiskers delineate the 10th and 90th percentiles, and black points indicate outliers.

Figure 1 also shows that longer passages imposed a greater memory load, as expected. Confirmation came from four Monte-Carlo Friedman one-way ANOVAs on $P(C_O)$, carried out for each set of normal stimuli and each set of prosodically altered stimuli combined. The differences between sets forced the use of separate ANOVAs. Each ANOVA ($N = 12$) had three levels (3-IP, 4-IP, and 5-IP passages). For normal sets I and II, the ANOVAs gave significant results, $\chi^2(2) = 15.167$, $p < .001$, and $\chi^2(2) = 18.167$, $p < .001$, respectively. Across the prosodically altered stimuli in sets I and II, the ANOVAs yielded $\chi^2(2) = 18.167$, $p < .001$, and $\chi^2(2) = 20.667$, $p < .001$, respectively.

The subjects who had heard the normal passages of set I heard the altered passages of set II and vice-versa. Nonetheless, the 3-IP and 4-IP stimuli of set I, normal or altered, were always more difficult than those of set II. These differences between the two sets therefore arose from a discrepancy in inherent difficulty and not from differences between subjects.

The syntactic structures of the narratives in sets I and II provided no obvious reasons for the better recall of set II. Within each IP, structures were principally SVO or VO. The only embedded clause occurred in a 4-IP narrative in set II. Each set had a 4-IP passage with a dependent connection between two successive intonation phrases. One VSO structure occurred in set I. If syntactic complexity affected recall of the stimuli, then this more complex 4-IP passage should have been harder than its counterpart in set I. The normal passage with the VSO structure, however, yielded a geometric mean of .90 for $P(C_O)$, while its counterpart produced a value of .48. Across the altered prosody conditions, the two passages gave geometric means of .61 and .64, respectively.

Word frequency, however, did contribute to the difference between sets. We used a frequency dictionary (Kilgarriff, 1997) based on the 'demographic' spoken part of the British National Corpus. Over 4.2 million spoken tokens had been tagged grammatically and counted to make the frequency dictionary. It lacked two items that occurred in our stimuli ('botany' and 'unburnt'). Each was assigned a count of 0.5, and the total number of items in the corpus was increased by 1. For each word in the individual 3-IP, 4-IP, and 5-IP passages, we obtained a frequency count from the dictionary and converted it to a log-probability [$\log(p)$]. Minimum $\log(p)$, which was intended to indicate how unusual or surprising a passage might be, was the measure that we found best related to performance. It was lower for longer passages and for the stimuli in set I.

4.5 Effects of prosodic alterations.

Eight subjects heard the stimuli with a given type of prosodic alteration: monotone, pause-free, and monotone and pause-free combined. For a given subject, subtracting $P(C_O)$ for the altered passages of a given type and length from $P(C_O)$ for equally long normal passages would presumably quantify the effect of the prosodic manipulation. The difference in difficulty between sets I and II, however, would confound the results for the 3-IP and for the 4-IP passages.

To try to overcome this confound, we compensated for the unequal difficulty of sets I and II before testing the effects of the prosodic manipulations. Consider first the 3-IP passages in sets I and II. To equate for difficulty, we increased each $P(C_O)$ for the relatively hard 3-IP passages in set I by multiplying it by a factor greater than unity. The multiplier was simply the ratio of the geometric mean $P(C_O)$ across subjects and passages for the easier 3-IP stimuli in set II to the geometric mean score for the harder 3-IP stimuli in set I. A separate multiplier was calculated for the normal passages and for the prosodically altered conditions as a whole. This increased all individual $P(C_O)$ scores for set I. Similar adjustments were carried out with the data for the 4-IP passages of the two sets. The 5-IP passages in the two sets showed no signs of unequal difficulty, so data on them needed no adjustment.

After the adjustments, we employed one-tailed Monte Carlo Wilcoxon tests ($N = 8$ in each instance) to evaluate the differences between scores on the normal passages and on each type of prosodically manipulated passage. For each type of manipulation, this resulted in 3 tests, one for each passage length. We therefore calculated 9 tests in all.

Two of the 9 tests were significant. The 3-IP pause-free passages produced a significant decrease in $P(C_O)$, $z = 2.366$, $p < .01$. The scores for the 5-IP passages were significantly lower for the monotone than for the normal stimuli, $z = 1.965$, $p < .025$. In addition, the Wilcoxon test for the 5-IP pause-free passages produced a marginally significant result, $z = 1.823$, $p < .05$.

Figure 2 shows the data. The dependent variable is the adjusted difference between performances on the normal and altered passages. Each individual plot contains individual results for 8 subjects. Performance differences are shown for the 3-IP, 4-IP, and 5-IP passages under each type of prosodic manipulation. Each horizontal dashed line indicates a null difference. Results that yielded significant Wilcoxon tests are marked with an asterisk ($p < .025$) or a double asterisk ($p < .01$); a question mark indicates a marginally significant effect. Figure 2 makes

apparent the lack of *consistent* significant differences between performances on the normal and the prosodically altered 3-IP and 4-IP passages. Indeed, the monotone-pause-free condition should have been the most difficult but yielded no significant differences whatsoever. Two of the three effects marked in Figure 2, however, arose from the 5-IP passages. To examine this further, one-tailed Monte Carlo Mann-Whitney tests (N=12 in each case) were applied to the data of Figure 1. For each set and each passage length, we evaluated the between-subjects difference between control data and results combined across prosodic manipulations. Of the six resulting tests, that for the 5-IP passages of set I proved significant, ($U = 29.5$, $p < .01$). Performance was poorer on the prosodically manipulated 5-IP stimuli.

The results in Figures 1 and 2 indicate that interference with prosody produced somewhat poorer recall of the 5-IP passages. The within-subjects and the between-subjects analyses for shorter passages did not yield similarly dependable effects.

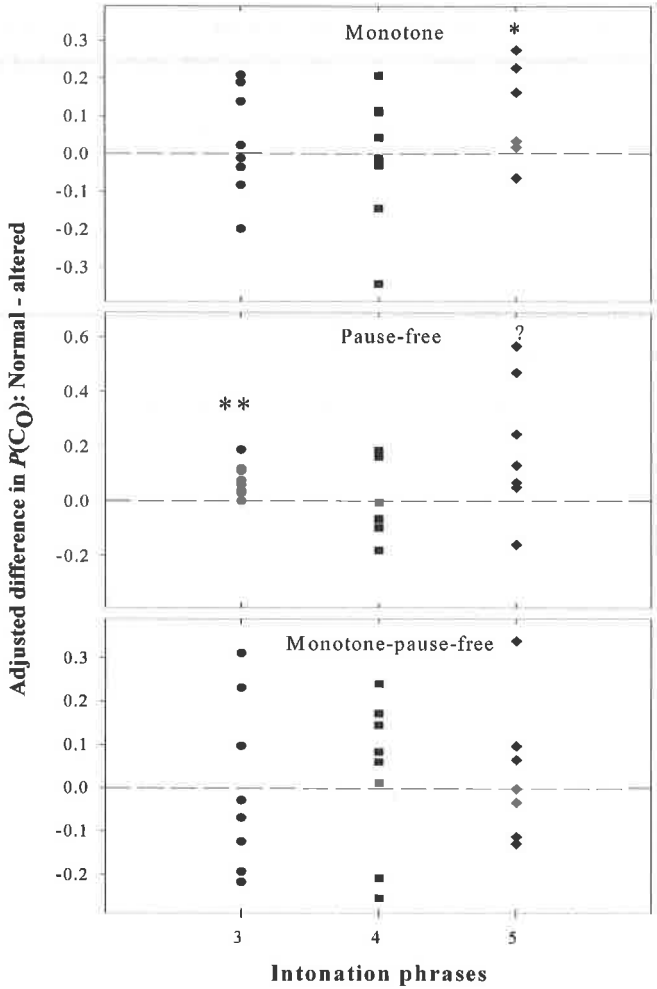


Fig. 2. Differences in $P(C_O)$ between normal and prosodically altered stimuli separated by passage length and type of prosodic manipulation, after adjusting for greater difficulty of set 1. Each data set shows results for eight individual subjects. Significant differences identified by one asterisk ($p < .025$) or two asterisks ($p < .01$); question mark indicates marginally significant difference ($p < .05$).

5 Discussion

Memory for spoken prose was virtually perfect for passages containing up to 13 words. Performance deteriorated progressively as passage length increased beyond 13 words. Almost one-third of performances were perfect, however, on stimuli with 18 or 19 words. Passages with 25 or 26 words even drew a few perfect performances, but none occurred on the 31-word stimuli.

These results go well beyond the memory span measured with lists of independent items. The larger memory span for prose could result from two influences. First, due to its syntactic and semantic properties, prose may enable more efficient 'chunking' (Miller, 1956) of material in immediate memory. Second, prose may be transferred more efficiently out of immediate memory into a longer-term store.

The majority of imperfect performances on 3-IP, 4-IP, and 5-IP passages arose from omitting words in the middle of the passage. Transfer of the initial words from immediate into a longer-term memory could explain retention of the initial words in a passage. The final words would be held in immediate memory. Omission of words in the middle of a passage would then result.

Word frequency affected performance on the 3-IP, 4-IP, and 5-IP narratives. The difference in difficulty between sets I and II, normal or altered, sprang at least partly from differences in word frequency. The probability of the least frequent word in a passage clearly affected performance on that passage.

5.1 Effects of prosodic manipulations.

We measured the effects of each of three prosodic manipulations on memory for prose: removal of F0 variation, removal of pauses, and removal of both. After adjusting for the greater difficulty of set I, only the 3-IP pause-free passages demonstrated significantly worse performance than did the normal 3-IP passages. The prosodic manipulations had no other effects on memory for the 3-IP and the 4-IP passages. The one positive result on the 3-IP monotone passages may be an artefact of our method of compensating for differences in difficulty between sets I and II.

The 5-IP monotone passages, however, proved significantly harder than the 5-IP normal passages, and the 5-IP pause-free passages were marginally harder than the normal narratives. In line with these results, between-subjects tests showed that

the combined 5-IP altered passages of set I were harder to recall than were the normal stimuli. No other similar between-subjects comparisons were significant. Differences between subgroups of listeners exposed to the different types of prosodic manipulation may explain why performance was no different from normal for the 5-IP monotone-pause-free stimuli.

Our generally negative results on the prosodically altered 3-IP and 4-IP passages agree with the findings of Stine and Wingfield (1987). They disagree, however, with previous positive reports on mnemonic effects of intonation (Leonard, 1974; O'Connell, Turner, & Onuska, 1968; Zurif & Mendelsohn, 1972; Paris, Thomas, Gilson, & Kincaid, 2000). All those experiments, however, used relatively short grammatically anomalous sentences, nonsense strings containing English articles and bound morphemes, or utterances with sudden changes of theme. Our straightforward narrative stimuli were very different. The negative findings on pause-free and monotone-pause-free narratives in our experiment also diverge from the results reported by Huttenlocher and Burke (1972), Frankish (1985, 1989), and Martin (1968). All these experimenters found that pauses affect memory for speech. Huttenlocher and Burke and Frankish, however, used lists of random digits, while Martin found that unusually long pauses after principal words decreased recall of grammatical, anomalous, and scrambled utterances.

The differences between our negative results and previous positive findings on the effects of prosody on memory for speech seem to depend on the use of different kinds of stimuli. We employed continuous prose with a relatively simple structure. The previous positive experiments, however, used thematically disjoined or anomalous utterances, lists of unrelated items, nonsense strings, long pauses, or high speech rates. Such conditions are unusual in daily life. They seem to bring out effects of intonation or pauses on memory. When speech contains routine properties and material, however, prosody seems of little importance for memory. The redundancy that other factors afford readily overcomes any effects of interference with prosody. When the going gets difficult, however, prosody may affect memory for speech.

A variety of facts supports this argument. First, increased speech rates facilitate demonstrations of effects of prosody on memory for prose (Stine & Wingfield, 1987; Wingfield, 1975). Second, Paris, Thomas, Gilson, & Kincaid (2000) reported that their variant of citation form prosody interfered with memory for passages of 15-20 words that contained sudden shifts in theme. In contrast, Stine and Wingfield (1987) found little such effect for continuous 16-word prose

passages with no thematic disjunctions. Third, cross-sentence splicing experiments (Darwin, 1975; Wingfield, 1975; Wingfield & Klein, 1971) create unusual stimuli. Such experiments bring out a role for prosody in memory for prose. Fourth, prosody facilitates syntactic disambiguation (see Cutler, Dahan, & van Donselaar, 1997), which necessarily involves material that is hard to process. Fifth, only our longest, 5-IP stimuli yielded positive findings that were consistent in both within- and between-subjects analyses.

The data in Figure 3 give further support to our argument. Total counts of words in correct order were obtained across subjects for each normal passage. We also did this for each passage combining over the three altered conditions. This yielded $P(C_{OG})$, the grand proportion of ordered words correct for a given passage under either normal or pooled altered conditions. In Figure 3, $\log[P(C_{OG})]$ for prosodically altered passages is plotted on the vertical axis against $\log[P(C_{OG})]$ for normal passages. Points that fall around the diagonal line indicate no essential effect of altered prosody. Four points clearly depart from the diagonal. All have relatively low values of $\log[P(C_{OG})]$ for normal passages, compared to the six points clustered at the upper right of the figure. The four normal narratives that yielded these data therefore tended to be relatively difficult even prior to any prosodic changes. The prosodic alterations made them even harder to remember. Three of the four passages contained five intonation phrases.

The most puzzling result in Figure 3 is the poor performance on a 3-IP passage (filled circle). Nothing seems to differentiate it from the other 3-IP passages. In addition, performance was equally good on the normal and altered versions of one 5-IP passage. It seems quite comparable to the other 5-IP passages that yielded reduced performance when altered. Factors other than minimum $\log(p)$ and passage length seem to underlie these two discrepancies.

In summary, prosody may facilitate memory for speech only as processing becomes increasingly difficult. Previous research has shown that difficult conditions include high speech rates, ambiguous or even abnormal syntax, thematic discontinuity, and unstructured sequences of items. Our findings now add two additional factors to this list: increasing utterance length and lower word frequency. As long as processing demands are sufficiently low and redundancy is sufficiently high, however, prosody seems unnecessary for remembering spoken material.

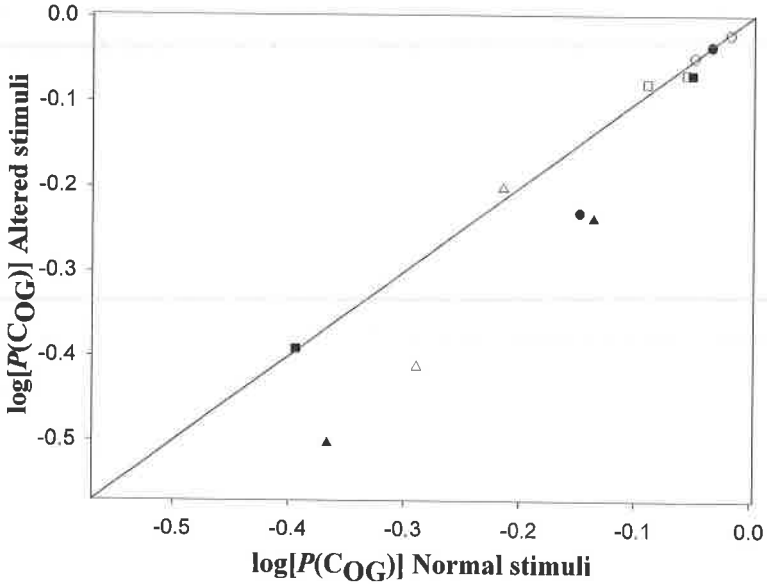


Fig. 3. Log[$P(C_O)$] for prosodically altered passages plotted against log[$P(C_O)$] for normal passages. Filled symbols: set I; open symbols: set II. Circles, squares, and triangles: 3-IP, 4-IP, and 5–IP stimuli, respectively.

References

Berry, C. 1994. *Your voice and how to use it*. London: Virgin Books: 125-126.

Boersma, P. and Weenink, D. 1996. PRAAT: A system for doing phonetics by computer. Report of the Institute of Phonetic Sciences, University of Amsterdam 132.

Carpentier, F. and Moulines, E. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication, 9*: 453-457.

Cutler, A., Dahan, D., and van Donselaar, W. 1997. Prosody in the comprehension of spoken language: A critical review. *Language and Speech 40*: 141-201.

Darwin, C. J. 1975. On the dynamic use of prosody in speech perception. In A. Cohen and S. G. Nooteboom Eds., *Structure and process in speech perception.* Berlin: Springer-Verlag: 178-193.

Frankish, C. 1985. Modality-specific grouping effects in short-term memory. *Journal of Memory and Language 24*: 200-209.

Frankish, C. 1989. Perceptual organization and precategorical acoustic storage. *Journal of Experimental Psychology: Learning, Memory and Cognition, 15*: 469-479.

Huttenlocher, J. and Burke, D. 1976. Why does memory span increase with age? *Cognitive Psychology, 8*: 1-31.

Institute of Electrical and Electronics Engineers. 1969. *IEEE recommended practice for speech quality measurements* IEEE No. 297. New York: Author.

Kilgarriff, A. 1997. Putting frequencies in the dictionary. *International Journal of Lexicography 10*: 135-55.

Ladd, D. R. 1996. *Intonational phonology*. Cambridge: Cambridge University Press.

Leonard, L. B. 1974. The role of intonation in the recall of various linguistic stimuli. *Language and Speech, 16*: 327-335.

Mandel, S. 1993. *Effective presentation skills* Revised ed.. Menlo Park, CA: Crisp Publications.

Martin, J. G. 1968. Temporal word spacing and the perception of ordinary, anomalous, and scrambled strings. *Journal of Verbal Learning and Verbal Behavior, 7*: 154-157.

Miller, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review 63*: 81-97.

O'Connell, D. C., Turner, E. A. and Onuska, L. A. 1968. Intonation, grammatical structure, and contextual association in immediate recall. *Journal of Verbal Learning and Verbal Behavior, 7*, 110-116.

Paris, C. R., Thomas, M. H., Gilson, R. D. and Kincaid, J. P. 2000. Linguistic cues and memory for synthetic and natural speech. *Human Factors, 42*, 421-431.

Robinson, G. M. 1977. Rhythmic organization in speech processing. *Journal of Experimental Psychology: Human Perception and Performance 3*: 83-91.

Saito, S. 1998. Effects of articulatory suppression on immediate serial recall of temporally grouped and intonated lists. *Psychologia, 41*: 95-101.

Speer, S. R., Crowder, R. G. andThomas, L. M. 1993. Prosodic structure and sentence recognition. *Journal of Memory and Language, 32*, 336-358.

Stine, E. L., & Wingfield, A. 1987. Process and strategy in memory for speech among younger and older adults. *Psychology and Aging, 2*: 272-279.

Syntrillium Software Corporation, 1996. *Cool96.*

Wechsler, D. 1981. *Wechsler adult intelligence scales revised.* San Antonio TX: Psychological Corporation and Harcourt Brace Jovanovich.

Wingfield, A. 1975a. Acoustic redundancy and the perception of time-compressed speech. *Journal of Speech and Hearing Research 18*: 96-104.

Wingfield, A. 1975b. The intonation-syntax interaction: prosodic features in the perceptual processing of sentences. In A. Cohen and S. G. Nooteboom Eds., *Structure and process in speech perception.* Berlin: Springer-Verlag: 148-156.

Wingfield, A. and Klein, J. F. 1971. Syntactic structure and acoustic pattern in speech recognition. *Perception and Psychophysics 9*: 23-25.

Wingfield, A., Lahar, C. J., and Stine, E. A. L. 1989. Age and decision strategies in running memory for speech: Effects of prosody and linguistic structure. *Journal of Gerontology: PSYCHOLOGICAL SCIENCES 44*: 106-113.

Zurif, E. B. and Mendelsohn, M. 1972. Hemispheric specialization for the perception of speech sounds: The influence of intonation and structure. *Perception and Psychophysics 72*: 329-332.

Cross-linguistic study of the effect of suprasegmental features conditioning the development of nasal vowels

IAN WATSON AND JOHN HAJEK

1 Introduction

Certain suprasegmental patterns have been shown to condition the perception of vowel nasality and hence to favour the historical development of nasalised vowels. However, experiments showing this have typically used only anglophone subjects. The present experiment examines whether these effects generalise to speakers of French, a language with significantly different suprasegmental properties from English. We repeated with 10 French subjects perceptual tests already used with English speakers and compared results for the two groups with respect to three parameters known to be potentially salient in conditioning nasal percepts on vowels. Results show that the responses of the French subjects differ significantly from those of the English subjects for all the parameters, and in ways related to the differences between French and English phonology. Nevertheless, both groups showed at least some minimal sensitivity to all the effects studied, so that we are unable to rule out the possibility of their having a universal basis.

2 Background

In previous work (Hajek and Watson, 1998; Watson and Hajek, 1999), we have investigated the role of factors known to condition the development of vowel nasality using the techniques of experimental phonology. Hajek, 1997 showed that in a range of Romance dialects, four suprasegmental factors seemed historically to privilege the development of nasal vowels. He named these (i) the Vowel Length parameter, (ii) the Stress parameter, (iii) the Extended stress parameter and (iv) the Foot parameter. These parameters express, respectively, the preferential nasalisation of (i) long rather than short vowels; (ii) vowels in stressed rather than unstressed syllables; (iii) unstressed vowels in pre-tonic rather than post-tonic

position; and (iv) of stressed vowels in oxytonic (weak-strong) rather than paroxytonic (strong-weak) patterns. In Hajek and Watson, 1998 and Watson and Hajek, 1999 we sought to establish whether these conditioning factors had a perceptual basis, of the type first advocated for diachronic phenomena by Ohala (see, e.g., Ohala, 1975). Evidence of such a basis for the Vowel Length parameter was first provided by Whalen and Beddor (1989) in a study using American subjects. In our previous studies we used English subjects. In Hajek and Watson (1998) we confirmed Whalen and Beddor's findings, and found evidence of a perceptual basis for the Stress parameter. In Watson and Hajek (1999) we further showed a similar perceptual basis for the rhythm parameter, but not the Extended stress parameter.

However, no inference of universality may be drawn from the experimental studies carried out thus far; they are limited by the shared characteristic that their subjects were native speakers of English. Many perceptual phenomena vary cross-linguistically (see, e.g., Repp, 1984). Languages, furthermore, exhibit considerable diversity with in their use and realisations of suprasegmental phenomena such as those in question here (Lehiste, 1970). There is no *a priori* reason to assume that the previous results hold for any languages other than those which resemble English in at least three relevant respects: English has no phonemically nasal vowels, but does have contextual vowel nasality. English has marked and phonologically contrastive vowel-length differences. English has non-fixed, contrastive stress. In the present study, we have therefore sought to address the empirical question of the effect of native language on the perceptual phenomena investigated in our previous work, and thus to gain an initial insight into the possible universal status of the vowel length, stress and foot parameters. The perceptual tests as in Watson and Hajek (1999) were repeated with French-speaking subjects. Their responses were compared to those of the English subjects reported previously, to test the null hypothesis that a change in native language will have no effect on the outcome of the perceptual tests.

2.2 French vs. English

The use of French subjects permits constitutes a harsh test of the null hypothesis. French and English differ both prosodically and segmentally in potentially crucial ways. The phonology of French prosody has been hotly debated, but, observationally, it differs from English in that (i) it has fixed main stress; (ii)

this is located on the final syllable of intonation units, whereas English is predominantly trochaic (Cutler and Carter, 1987); (iii) secondary stress is variable in both position and realisation (Fónagy, Léon and Carton 1979). Segmentally, French has in recent times lost the last vestiges of distinctive vowel length, although vowel lengthening is an obligatory allophonic process in certain types of stressed (and therefore final) syllables. Above all, standard French has 3 or 4 distinctively nasal vowels. These are typically produced with longer durations than oral vowels, and have been analysed as diphthongs, starting with an oral portion followed by a movement to a nasal target (Linthorst, 1973).

The French subjects could thus present a number of potential challenges to the null hypothesis. The technique used in the experimental studies cited above involves subjects making graded judgements of the degree of nasality in vowels; as French phonology uses vowel nasality categorically, French subjects may have difficulty making such judgements. Furthermore, responses to each of Hajek's parameters could be affected by the prosodic structure of French. As stress is non-distinctive, its presence or absence may play a lesser conditioning role on other phenomena than in English. Alternatively the fixed nature of French stress may constrain francophones to attend to stress-related phenomena only if these are situated in phrase-final position; in this case we would observe, relative to the anglophone subjects, a restriction on the stress parameter, but an intensification of the preference for oxytonic patterns in the Foot parameter.

3. Experimental design

3.1 Stimuli

As described in Watson and Hajek (1999), a series of twenty-four disyllabic synthetic stimuli was created using the HLSYN pseudo-articulatory synthesizer. This synthesizer drives an implementation of the KLYSN formant synthesizer using 10 articulatory variables specified by the experimenter. Nasality was varied using the velopharyngeal port opening (VPO) parameter, specified in mm^2. The stimuli made up two continua; in one the first vowel varied in its degree of nasality, going from [asa] to [ãsa] in three steps, specified by setting VPO for the first vowel to 0, 16.8 and 36 mm^2. In the second continuum, it was the vowel in the second syllable whose nasality was varied, in a similar fashion, giving an [asa]

to [asă] continuum. For both continua, for each degree of VPO, two basic length settings were used, 250 ms and 150 ms; to test the effect of prominence, the relation between the vowels in the two syllables was varied. In the stressed condition, the target vowel (i.e. that whose nasality was being varied, so the first vowel for continuum 1, the second for continuum 2) had a higher intensity than the non-target vowel (sub-glottal pressure of 8.5 vs. 6 cm H_20), was 100 ms longer than the non-target, and was marked by a major F0 fall of 55 Hz, while the non-target had level pitch. In the unstressed condition, the non-target vowel was correspondingly louder, longer and pitch prominent.

This gives a total of 3 (VPO settings) x 2 (length settings) x 2 (prominence conditions), i.e. 12 stimuli per continuum, so 24 stimuli in all. These were recorded six times in pseudo-random order onto a tape, with an inter-stimulus interval of 3 seconds, the two continua being recorder separately. Our null hypothesis predicts that all of the factors varied here should impinge on nasality judgements for French subjects as they did for the English subjects in Watson and Hajek (1999). Specifically (i) stimuli with higher VPO settings should be perceived as more nasal than those with lower settings; (ii) for the same degree of VPO, longer vowels should be perceived as more nasal than shorter vowels; (iii) for the same VPO and vowel length, stressed vowels should be perceived as more nasal than unstressed vowels; and (iv) final stressed vowels should be perceived as more nasal than non-final stressed vowels of the same length and VPO. It should be noted that, whereas for the English subjects in our previous study, the two stress patterns synthesized here are both normal both in conversational speech and in terms of the patterns found in the lexicon, for the French subjects, only the oxytonic pattern has that status. In French, both lexical items and phonological phrases have final stress. This default pattern may, however, be modified under certain pragmatic conditions, such as disambiguation: 'J'ai dit que c'est un **pê**cheur, pas un **pé**cheur.' Such a pattern will not, therefore, be completely unnatural for the French subjects.

3.2 Subjects And Procedure

Subjects were ten native speakers of metropolitan French, all staff or students of the Institut de la Communication Parlée in Grenoble. None was a competent speaker of any foreign language, but all had received some phonetic training. They were compared to the subjects described in Watson and Hajek (1999), who were

speakers of British English, all students at Oxford University. Subjects were asked to respond to each stimulus by marking on a pre-prepared sheet how nasalised they considered the first vowel to be, on a scale 1 (least nasal) - 5 (most nasal). Subjects were run individually, or in very small groups, the experiment being preceded by a short practice session. Each subject heard the stimuli from continuum 1 in one block, followed by those for continuum 2. For each continuum, the first twelve responses in the experiment itself were discarded for each subject, leaving five responses to each stimulus.

The ratings from these were then input to a repeated measures ANOVA, with dependent variable nasality judgement, between-subjects factor language, and within-subjects factors VPO, vowel length, subject, syllable position of nasal vowel (first or second) and prominence of nasal vowel (stressed or unstressed). Further tests were carried out on the results from the French subjects only, matching those for the English subjects previously reported. In particular, planned comparisons were made between the two prominence conditions, and the interactions between prominence and VPO, prominence and syllable position, and the three-way interaction prominence*syllable position*VPO.

4 Results

4.1 Velopharyngeal opening

The degree of velopharyngeal opening was highly significant for the French subjects ($F[2,8]= 21.4$, $p < .001$), as for the English ([1,2]); see Figure 1 below, which shows responses against degree of VPO for each language. The French subjects were therefore able to make judgements concerning the relative degree of nasality on vowels.

4.2 Vowel length parameter

To facilitate cross-linguistic comparisons for this parameter, Figure 2 presents the vowel length results for both languages in terms of the difference in nasality ratings between long and short vowels for each group at each degree of VPO; it thus gives a direct representation of the extent of the effect. This is much smaller in French than English for all VPOs. Analysis of variance shows the two subject

groups to be significantly different ($F[1,18]=5.7$, $p < .028$). Further analysis of just the French subjects shows that for them the vowel length distinction does not quite reach the .05 level of significance ($p < .01$ for the English subjects). However, planned comparisons of the effect at different VPO levels show a significant distinction at the highest level ($p < .005$). There is thus a marginal effect of vowel length in French.

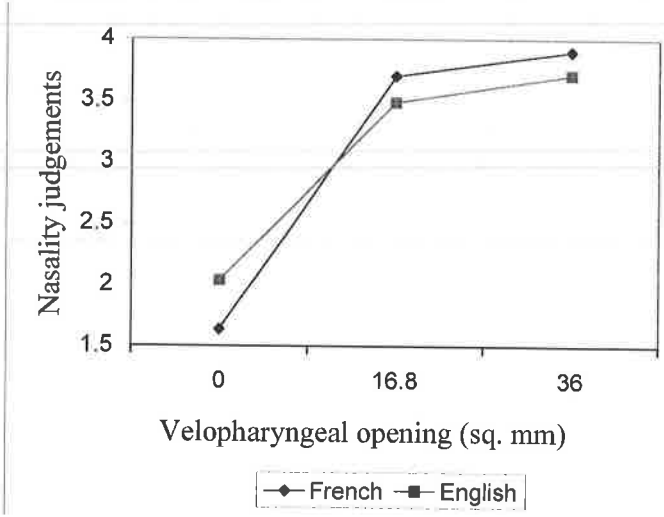


Figure 1. Nasality judgements for both groups

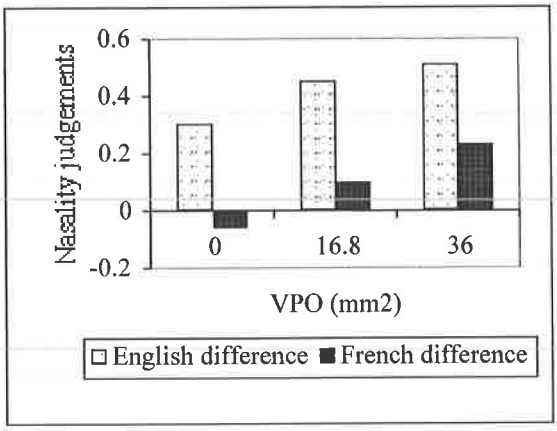


Figure 2. Differences between judgements for long and short vowels

4.3 Stress parameter

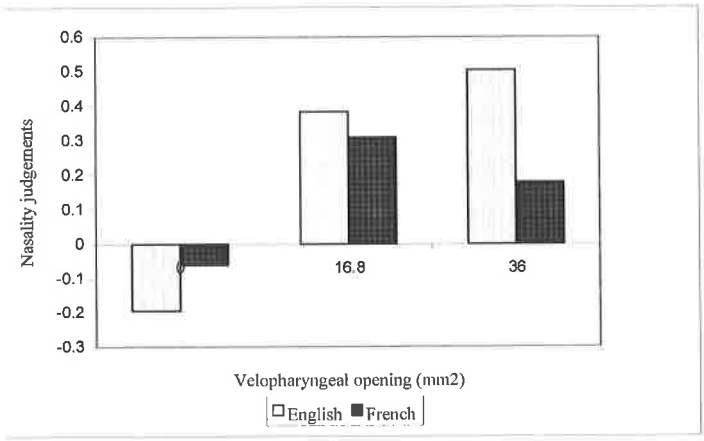


Figure 3 Differences between responses according to stress parameter

In Figure 3, the difference in nasality ratings between stressed and unstressed syllables is given for each group at each degree of VPO. The overall statistical

analysis shows that the two groups differ significantly; the stress*language interaction is not significant, but that between stress*language*VPO is highly so ($F[2,17]=8.1$, $p < .003$). In English, the main effect was significant ($p < .02$), with significant distinctions at the two higher VPO levels. For the French subjects, the main effect of stress is not significant ($p < .184$), but the stress/VPO interaction is significant ($p < .015$). Planned comparisons show that in French stress is significant only for the intermediate VPO value (16.8 mm2; $p < .022$). However, as noted above, main stress usually only occurs in final position in French, and the possibility arises that the mixing of final and pre-final positions may be a confounding factor here. This interpretation was confirmed by detailed analysis: the stress/unstressed difference was significant in French only in final position.

4.4 Foot parameter

As suggested by the above, the French subjects responded to this parameter more strongly than the English (see Figure 4) except at the VPO setting of 0 mm2, at which they showed no effect at all. ($p < .05$). For the English subjects the distinction is weakly significant overall, and highly significant at each VPO level ($p < .01$), whereas for the French it reaches significance only for VPO settings of 16.8 mm2 ($p < .002$) and 36 mm2 ($p < .0002$).

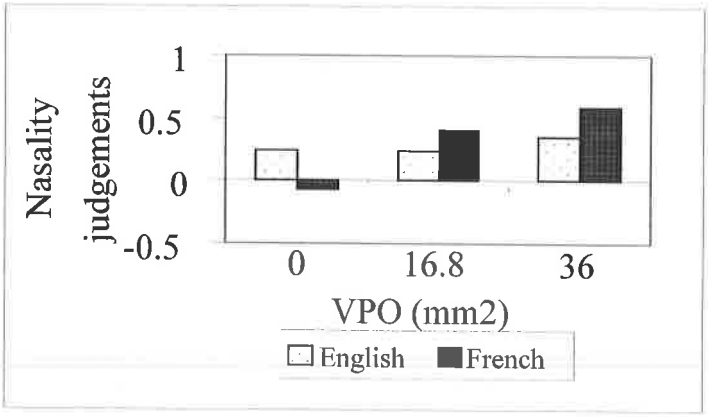


Figure 4 Difference in nasal responses conditioned by rhythm parameter

5 Discussion

For the three parameters examined here, the French subjects differ significantly from their English counterparts, while still showing sensitivity to the effects investigated. In each case, the differences may be attributed to subjects' perceptual behaviour being attuned to the role played by the suprasegmental phenomena concerned in their native language. The two groups show particular disparities with respect to the vowel length and stress parameters. The effect of vowel length was highly significant for speakers of English, in which vowel length contributes to phonemic distinctions, but weak in French, where it does not. Future work should investigate whether there are languages in whose phonologies vowel length plays a role so small that their speakers are completely insensitive to this parameter: it will be recalled that French phonology does have non-distinctive lengthening, and that French nasal vowels have been described as long.

The results for the stress parameter require more complex interpretation. The main effect of stress was non-significant, as might be predicted for subjects for whom main stress, being fixed and non-contrastive, is relatively uninformative. However, we found the stress/VPO interaction was significant, as was the stressed/unstressed contrast for the middle degree of VPO, indicating that the subjects do have some sensitivity to this parameter. Furthermore, the results for the rhythm parameter show that the perception of nasality on stressed syllables is highly conditioned in French by the relative position of those syllables (final vs. penultimate), more so than in English. Indeed, if one focuses only on those positions where main stress may normally be found in each language (either position in English, final in French), there is little difference between the two.

The rhythm parameter was significant in both languages, with, overall, a stronger effect in French. The French subjects' greater preference for oxytones reflects the normal patterning of the language. Non-final stresses in French are either secondary and weak (and these are not normally penultimate) or highly marked. More complex experimentation would be needed to determine whether, in the latter case, nasality is as easily perceived as on stressed final syllables.

6 Conclusion

The null hypothesis tested here was that French subjects would not differ from anglophones in their responses to three factors previously shown to condition the

perception of vowel nasality. That hypothesis must be rejected. There are differences, largely attributable to the different phonological patterns of the two languages. However, all of the three factors were shown to have some effect in both languages. The vowel length effect was very weak in French, leaving open the possibility that this is not of universal applicability, and that a language could be found whose speakers would not respond to this parameter at all. The Stress and Rhythm parameters were salient for both groups of subjects, although they are subject to language-specific variations in implementation.

References

Cutler, A. and Carter, D. M. 1987. The predominance of strong initial syllables in the English vocabulary. Computer Speech and Language 2: 133-142.

Fónagy, I, Léon P & Carton, F. 1979. *L'accent en français contemporain* . Ottawa: Didier

Hajek, J. 1997. *Universals of sound change in nasalization.* Oxford: Blackwell.

Hajek, J. and Watson, I.M.C. 1998. More evidence for the perceptual basis of sound change? Suprasegmental effects in the development of distinctive nasalisation. *Proceedings of the International Congress on Spoken Language Processing.* Sydney; Causal: 1763-1765.

Lehiste, I. 1970. *Suprasegmentals.* MIT Press: Cambridge MA.

Linthorst, P. 1973. Les voyelles nasales du français; étude phonétique et phonologique. Groningen : V. R. B.

Ohala, J.J. 1975. Phonetic explanations for nasal sound patterns. In Ferguson, C.A., Hyman, L. & Ohala, J.J. *Nasálfest.* Stanford: Department of Linguistics, Stanford University: 289-316.

Repp, B.H. 1984 Categorical Perception: Issues, methods and Findings. In N.J. Lass (ed.) Speech and Language: Advances in basic research and practice (Vol. 10). New York, Academic Press

Watson, I.M.C. and Hajek, J. 1999. A perceptual basis for the Foot Parameter in the Development of Distinctive Nasalization. *Proceedings of the XIVth International Congress of Phonetic Sciences*. San Francisco.

Whalen, D.S. and Beddor, P.S. 1989. Connections between nasality and vowel duration and height: Elucidation of the Eastern Algonquian intrusive nasal. *Language* 6.

Automated Assessment of Examination Scripts

STEPHEN G. PULMAN AND JANA Z. SUKKARIEH[1]

Abstract

In this paper we report on progress on a project to build a system for the automated assessment of free text examination answers. We describe some existing approaches to the assessment of free text answers and essays, and describe two experiments in which we have applied information extraction and document classification techniques to GCSE biology scripts, national examinations taken by 16 year old children.

Introduction

The idea of automatic marking (grading) of examinations is not new. It dates to the early 60's (Page, 1994), but has been mostly restricted to multiple choice questions, and other schema specifically designed for computer processing. Such systems invite the criticism that they are inflexible and inhibiting for the student, as well as limited in their applicability. Free text assessment is a hard challenge, since it seems to presuppose an advanced level of performance in automated natural language understanding. Nevertheless, the payoff would be high if such systems were successful, in terms of cost (expert examiners are expensive), speed (marking is a slow process), and consistency (even experts disagree among themselves, and the same examiner will give different grades to the same paper on different occasions).

Free text assessment is being revisited for all the above reasons, because of advances in natural language processing (NLP) techniques and also because of the wide availability of computers in schools and centres for standardized tests. Already, there are a few systems that mark essays directly which are being used in 'real' settings. The three most prominent of these are now described briefly.

[1] Order of authors is alphabetical.

Latent Semantic Analysis

The Intelligent Essay Assessor (IEA) was developed by Knowledge Analysis Technologies (KAT), Colorado (Foltz, Laham and Landauer, 2003). It uses Latent Semantic Analysis for scoring answers of students as part of a tutoring domain. Latent Semantic Analysis is based on word-document co-occurrence statistics in the training corpus represented as a matrix, which is subsequently decomposed, and then subjected to a dimensionality reduction technique. LSA is used to compare students' answers to model answers by calculating the distance between their corresponding vector projections (Graesser *et al.*, 2000). IEA has been tested in different ways, namely, comparing essays to ones that have been previously graded, to an ideal essay or *gold standard* (Wolfe *et al.*, 1998), to portions of the original text, or to sub-components of texts or essays (Foltz, 1996; Foltz, Britt and Perfetti, 1996). In blind testing, agreement with human examiners is high, between 85 and 91 percent.

The LSA technique evaluates content via the choice of words and does not take into account any syntactic information — it is a 'bag-of-words' approach and can be fooled. "It has no way of knowing the difference between *The Germans bombed the British cities* and *The British bombed the German cities*" (Charles Perfetti)[2]. It cannot deal with any of the favourite list of difficult phenomena for NLP systems, like negation, attachment, binding, predication, modification, scope ambiguities and so on. Researchers have tried to improve the performance of LSA by adding some syntactic and semantic information; for example, adding a part-of-speech (POS) to the given word (Wiemar-Hastings and Zipitria, 2001) or adding a part-of-speech to the previous word (Kanejiya, Kumar and Prasad, 2003). The results do not show a significant improvement over the basic technique.

A Hybrid Approach

E-rater[3] (or Essay-rater) is a system developed by the Education Testing Service (ETS) (Burstein *et al.*, 1998a; Burstein, Leacock and Swartz, 2001; Burstein *et al.*, 1998b). It has been used to rate GMAT (a business school admission test) and TWE essays (Test of Written English) for prospective university students. The

[2] *Teachers of Tomorrow?* http://www.wired.com/news/technology/0,1282,16009,00.html
[3] http://www.ets.org/research/erater.html

system uses shallow parsing techniques to identify syntactic and discourse features. Content is checked by vectors of weighted content words. An essay that stays on the topic, is coherent as evidenced by use of discourse structures, and has a good vocabulary and varied syntactic structure is to have a higher grade. E-rater uses both NLP and statistical tools to model the decision of a human marker, and achieves impressive agreement with human markers when tested on unseen data (84–91%).

Another hybrid approach is described by Rosé *et al.* (2003) at the University of Pittsburgh. Their system evaluates qualitative physics questions using a hybrid approach between machine learning classification methods using features extracted from a linguistic analysis of a text and a naive Bayesian classification.

Information Extraction

In the UK, Intelligent Assessment Technologies have developed an automatic essay assessor called Automark[4] (Mitchell *et al.*, 2002). The system uses information extraction techniques in the sense that the content of a correct answer is specified in the form of a number of mark scheme templates. The essay to be marked is fed into a parser (they use the Link Grammar parser (Sleator and Temperley, 1991; Sleator and Temperley, 1993)) and the parsed text is then compared to the already-defined templates or mark scheme. Mitchell *et al.* (2002) claim about 95% agreement with human markers in blind testing.

Callear, Jerrams-Smith and Soh (2001) at the University of Portsmouth also use pattern-matching techniques to mark short answers in programming languages, psychology and biology-related fields.

The UCLES Application

At the time we began this project, we were not aware of the Intelligent Assessment Technologies work. However, we had already decided on the basis of reading about E-rater that information extraction techniques were a likely candidate for our application, since they do not require complete and accurate

[4] You can find demos at http://www.intelligentassessment.com/demonstration.htm for English Comprehension and at http://examonline1.nsl.co.uk/ExamOnline/Jsp/index.jsp for Key Stage 2 Science National Test Questions.

parsing, they are relatively robust in the face of ungrammatical and incomplete sentences (of which GCSE scripts provide a plethora of examples), and they are fairly easy to implement quickly.

After an initial look at a sample of different GCSE and A-level examination papers we decided to begin with GCSE Biology exams where the answers are restricted to about 5 lines and deal with facts rather than subjective opinions or interpretation. Our sponsors, the University of Cambridge Local Examinations Syndicate (UCLES), arranged to have a sample of these scripts transcribed from the original hand-written versions into machine readable text. About 25% of the answers were retained by UCLES to use as unseen test data for our system.

Here are some example questions along with their answer keys — short, often very terse, descriptions of acceptable answers provided by examiners.

1 Write down two things about asexual reproduction in plants which is different from sexual reproduction.

Key:

Can be done at any time
Needs no flowers
Does not need 2 gametes/parents
No fertilisation
No meiosis involved
No genetic variation/clones/identical/same as parent plant

2 Explain what causes two twins to be identical.

Key:

Formed from the same bundle of cells/
same fertilised egg/same embryo /
formed from one egg and sperm /
mitosis forms identical cells
so genetic information the same/
same genes/same DNA/same chromosomes

3 Where could you detect the pulse and what causes it?

Key:

Found at wrist/temple/neck /
where an artery close to the skin can be pressed against

a bone or ankle for infants;
heart beating / blood surging in an artery /
wave down artery wall

Our starting point was 202 × 6 marked student answers for training, and approximately 60 answers per question unseen held-out data. Each answer is associated with the score given by expert examiners, and this score is either 0 (incorrect), 1 (partially correct or incomplete) or 2 (correct and complete). We used a Hidden Markov Model part-of-speech tagger trained on the Penn Treebank corpus, and a Noun Phrase and Verb Group finite state machine (FSM) chunker to provide the input to the information extraction pattern matching phase. The NP network was induced from the Penn Treebank, and then tuned by hand. The Verb Group FSM (i.e. the Hallidayean constituent consisting of the verbal cluster without its complements) was written by hand.

Customisation and Shallow Processing

We have assessed the performance of the tagger on the data (students' answers). The following table gives an idea, in figures, about its performance on the test set we started with. A major source of inaccuracy for taggers is the presence of unknown words. The Wall Street Journal section of the Penn Treebank is not particularly rich in biological vocabulary, and so we would expect problems in this respect, even though the tagger includes some heuristics for guessing at unknown words. To factor out the unknown word issue we first ran the tagger on sentences which contained no words unknown to it (although the entry for the word might still not be the correct one, of course). Then we tested the tagger on a large random sample of answers, and finally on that same sample with all unknown words added (and spelling errors corrected).

Students' answers	Performance of the tagger
No Unknown Words in answers	Total # of words 4030 # of words tagged wrongly: 60
Random Answers	Total # of words: 14196 # of words tagged wrongly: 218
After correcting spelling errors (and adding all new words)	Total # of words: 14196 # of words tagged wrongly: 98

Here is a sample of the output of the tagger and chunker:

```
When/WRB [the/DT caterpillars/NNS]/NP [are/VBP feeding/VBG]/VG
on/IN [the/DT tomato/JJ plants/NNS]/NP,/, [a/DT chemical/NN]/NP
[is/VBZ released/VBN]/VG from/IN [the/DT plants/NNS]/NP./.
[This/DT chemical/NN]/NP [attracts/VBZ]/VG [the/DT wasps/NNS]/NP
[which/WDT]/NP [lay/VBP]/VG [eggs/NNS]/NP inside/IN
[the/DT caterpillars/NNS]/NP./.
[These/DT eggs/NNS]/NP [hatch/VB]/VG into/IN [larvae/NN]/NP
[which/WDT]/NP then/RB [become/VBN]/VG [pupae/NN]/NP./.
[The/DT pupae/NN]/NP then/RB [develop/VB]/VG into/IN
[adult/NN wasps/NNS]/NP./.
```

The Pattern-Matcher

Information extraction can only be used if a fairly determinate task specification and a clear criterion for success are given (Appelt and Israel, 1999). Our task is fairly determinate, namely, identify a right GCSE Biology answer and it has a reasonably well-defined criterion for success. Information extraction consists of applying a set of patterns and templates to discover members of a fixed list of *named entities* and relations within the texts, occurring in a specific configuration.

In our first attempt, given a question and answer, we try to identify a chunk of text in the answer that qualifies for a mark. We do this at a fairly concrete level on the basis of particular collections of keywords. In a more refined version, we would try to identify more abstract semantic elements, like relations, properties, etc. but we wanted to get a baseline system up and running as quickly as possible.

Information extraction patterns, the things that get from each answer the information relevant to the particular task, can be either discovered by a human or can be learned with machine learning algorithms, although to date these techniques have not reached human levels of accuracy in the building of IE systems. In this version of the system, we opted for the first way, namely, the knowledge-engineering approach. This also means that the grammar describing the patterns is constructed by hand and only someone who is familiar with the grammar and the system can modify the rules. Moreover, this approach requires a lot of labour, as will become evident below.

Patterns and Grammar

The 3 crucial steps in which to write extraction rules by hand can be found, among other references on information extraction, in (Appelt and Israel, 1999). These, in order, are:

- Determine all the ways in which the target information is expressed in a given corpus.
- Think of all the plausible variants of these ways.
- Write appropriate patterns for those ways.

Clearly the intuition of the linguistic/knowledge engineer plays an important role. For each domain, this requires some training as one is looking for a tightly defined, mostly unambiguous set of patterns that cover precisely the ways the target information is expressed, yet written in a way that captures the linguistic generalisations that would make it unnecessary to enumerate all the possible ways of expressing it. For the biology task we abstracted the patterns over 3 sets of data. First, we fleshed out the compact key answers provided by the examiners. Second, we used our own version of the answers (we sat the exam ourselves and with the help of a recommended GCSE biology book (Jones and Jones, 2001) we provided answers for the questions). The last set of data we abstracted patterns over was the training data that UCLES provided for us. After checking out the variant ways answers could be devised, we devised a simple language in which to write the patterns.

```
Pattern      ->   Word | Word/Cat | Symbol
                  | Variable | Disjunction
                  | Sequence | k(N, Sequence)
                  (N is upper limit of length of Sequence)
                  | k(N, Sequence, Pattern)
                  (Sequence NOT containing Pattern)
Disjunction  ->   {Pattern, ..., Pattern}
Sequence     ->   [Pattern, ..., Pattern]
Word         ->   sequence of characters
Cat          ->   NN | VB | VG ...
Symbol       ->   & | % | $ ...
Variable     ->   X | Y | Z ...
```

It is easy then to build up named macros expressing more complex concepts, such as 'negated verb group', or 'NP headed by word *protein*'.

The following answers exist in the training data as true answers for Question 2, namely, *Explain what causes two twins to be identical.*

1. the egg after fertilisation splits in two
2. the egg was fertilised it split in two
3. one egg fertilised which split into two
4. the fertilised egg has divided into two
5. 1 fertilised egg splits into two
6. one sperm has fertilized an egg... which split into two... etc.

These all imply *It is the same fertilised egg/embryo*, and variants of what is written above could be captured by a pattern like

```
singular_det + <fertilised egg> +
  {<split>; <divide>; <break>} + {in, into} + <two_halves>

singular_det      = {the, one, 1, a, an}
<split>           = {split, splits, splitting, has split, etc.}
<divide>          = {divides, which divide, has gone,
                     being broken...}
<two_halves>      = {two, 2, half, halves}
<fertilised egg> = NP with the content of 'fertilised egg'
etc.
```

Another difficulty we faced when writing patterns is that examiners allow unexpected ways for students to convey a particular scientific concept. Consider

the word *fertilisation*, where examiners seem to accept *the sperm and the egg meet*, *when a sperm meets an egg* or *the sperm reached the egg* instead.

As we said earlier, intuition plays an important role in writing patterns and it needs training for a particular domain. The development cycle will be familiar to anyone with experience of information extraction applications:

1. Write some rules.
2. Repeat until satisfied with results:
   (a) Run the system over a training corpus.
   (b) Examine output.
   (c) See where rules over-generate, under-generate, etc.
   (d) Modify or add rules.

Now that we have some rules written, we implemented a Prolog meta-interpreter that searches for the patterns in a particular given answer. This is the 'system' referred to in 'Run the system over ...' in stage 2(a) of the iterative process above.

The Basic Marking Algorithm

With each question, we associated a set of patterns or rules. The set of rules for a particular question was then divided into bags or equivalence classes where the equivalence relation, *R*, is *convey the same message/info as*. Equivalence classes are represented by one of their members. *X* belongs to a class **[Rep_of_Class]**[5] iff *X* bears *R* to Rep_of_Class. For example, 'only one parent' *R* 'just one parent'; **[only one parent]** = {*Pattern* | *Pattern R* 'only one parent'}. For Question 1, we have 6 equivalence classes, namely, **[only one parent]**, **[clone]**, **[no need to flower]**, **[can be done at any time]**, **[no fertilisation]**, **[no meiosis involved]**.

The key-answers given by the examiners determine the number of equivalence classes we have. Each equivalence class corresponds to 1 mark the examiners give. Assume we have *N* classes for a particular question:

[5] Representational note: we will always write the representative of a class in bold so that there is no confusion between an alternative of a pattern and a class of a pattern.

Class_1 with $\{C_{11}, C_{12}, ..., C_{2t_1}\}$
Class_2 with $\{C_{21}, C_{22}, ..., C_{2t_2}\}$
...
Class_N with $\{C_{N_1}, C_{N_2}, ..., C_{Nt_n}\}$

```
Given an answer A,
Repeat until no more rules/classes are available
   If A match Cik
      Then   Mark-till-Now is Mark-till-Now + 1
                If    Mark-till-Now = Full Mark
                      Then Exit
          %(ignoring Ci+1, ..., Cn i.e. the rest of the classes)
                      Else Ignore Cij for  k <  j  <=  ti
          %(i.e. ignore the rest of the rules in the same bag)
                          See if A matches any rule in Ci+1
          %(i.e. jump to the next Class and repeat process)
```

The procedure *match* takes the patterns as described by the grammar, case by case and handles them accordingly. In the next section, we report the results of the pattern-matcher on both the training data and the testing data, followed by a brief discussion.

Some Anticipated Problems

Information extraction and shallow processing are not full natural language processing methods and so there will be many cases that we handle incorrectly:

- **The need for reasoning and making inferences:** Assume a student answers Question 1, above, with, *we do not have to wait until Spring*. An assessor that fails to infer *it can be done at any time* from the student's answer will give it a 0. Similarly, an answer like *don't have sperm or egg* will get a 0 if there is no mechanism to infer *no fertilisation*.
- **Students tend to use a negation of a negation (for an affirmative):** An answer like *won't be done only at a specific time* is the same as *will be done at any time*. An answer like *it is not formed from more than one egg and sperm* for Question 2, is the same as saying *formed from one egg and sperm*. This category is merely an instance of the need for more general reasoning and inference outlined above. We have given this case a separate category because here, the wording of the answer is not very

different, while in the general case, the wording can be completely different.

- **Contradictory or inconsistent information:** Other than logical contradiction like *needs fertilisation and does not need fertilisation*, an answer for Question 2 like *identical twins have the same chromosomes but different DNA* holds inconsistent scientific information that needs to be detected.

Some of these issues would be reconfirmed in the students' actual answers, as we will see below, and other issues will be flagged as we describe our approach. We start with the first method, namely, the pattern-matching technique.

Results

There were approximately 201 answers used as training data and 65 testing answers available for each question. The results are summarised in the following table:

	Training Data	Testing Data
Hits	88 %	88 %

'Hits' occur where the system's and the examiners' marks match (the percentage includes answers with mark 0). The results are for the first version of the system, and we find them highly encouraging. One would expect the system to be reasonably accurate in the training data, but to find no deterioration when moving to unseen data is very gratifying. This must mean that there is very little variation in the range of answers, and in particular that a training set of the size we have will very likely be enough.

In fact, the results are even a little better than this table indicates. If the number of misses where we believe the examiners got it wrong were deducted then the system is marking 91% of the answers correctly. Given that this was the first version of the system complete enough to test it is clear that we could get some further improvement by putting more work in on the patterns. However, there are two factors relevant here. Firstly, there is the amount of work involved in writing these patterns and the skills required to do it. It would be nice to be able to

automate or at least to de-skill the task of customising to new questions. Secondly, there are several observations about the behaviour of unintelligent pattern matching of this sort which suggest that the cost/benefit ratio may become unfavourable after a short time. Recall that these patterns are not doing full natural language understanding. This means that there will always be a trade-off between high precision (recognising patterns accurately) and high recall (recognising all variants of patterns correctly). It is not guaranteed that for a particular application the right trade-off can be found. This suggests that it is worth experimenting with machine learning techniques in order to help with the process of customisation. If this cannot be achieved fully automatically we could at least investigate what help could be given to the developers via such techniques. For this reason we next turned to a simple machine learning method in order to develop a suitable experimental framework.

Nearest Neighbour Classification

We have already mentioned some suitable techniques: Latent Semantic Analysis is, from one point of view, just a way of classifying texts. Other researchers (Rosé *et al.* (2003) and Larkey (1998)) have also used text-classification techniques in grading essays in subject matters like physics, or law questions where a legal argument is to be expected in the text. While the accuracy of such systems may not be able to exceed that of hand-crafted systems (although this is not a proven fact), they nevertheless have the advantage of being automatically customisable to new domains needing no other expert knowledge than that of a human examiner.

In general all variants of these techniques begin with a set of examples with a known analysis (preferably covering the entire range of types of analysis). The size of the set can be as few as 100, although the larger the better. When this training phase is complete, new examples to be analysed are matched with old ones, or a combination of them, and the closest match determines the appropriate response. In our application, the 'analyses' are scores, and we assign the score of the nearest matching example to the input to be rated. The matching process may be quite complicated — we will have to experiment with different variations. The attraction of such a setup is that customising it to a new exam would be a matter of marking a few hundred scripts by hand to provide the training examples, once the appropriate matching and analysis schemes have been discovered.

We have begun by using almost the simplest posible text-classification method, known as the *k nearest neighbour* technique (Mitchell, 1997). We decompose examples (or new input) into a set of *features*. These can be words, tuples of words, grammatical relations, synonym sets, combinations of these or whatever the linguistic analysis mechanism is capable of finding accurately. In our case, we began with word tokens, thus approximating a crude marking technique of spotting keywords in answers. We discard determiners and a few other function words which have very low discriminatory power, and assign to each content word which appears in the training set a weight, the so-called 'tf-idf' measure. This is *term frequency* (the number of times the term or feature appears in the example) multiplied by *inverse document frequency*, i.e. 1/(the number of times term or feature appears in all examples). Terms which do not distinguish well among examples will carry less weight. It is easy to see how to adapt such a measure to give higher weights to words that are associated with (in)correct answers, and less weight to words that occur in almost all answers.

Each example, and any new input, can be represented as a vector of weighted feature values, ordered and labelled in some canonical way. We can then calculate a cosine or similar distance measure between the training examples and the input to be scored.

To summarise:

- Collect a set of training examples representing the main possible outcomes.
- Represent each training example as a a vector of features, i.e. linguistic properties believed relevant. In the simplest case this will be keywords.

For example, we might categorise a set of answers on the basis of whether or not they contain one of the following key words:

	egg	fertilized	split	two	male	female	sperm
Answer1	1	1	0	0	1	0	1
Answer2	1	1	1	1	0	0	0
Answer3	1	0	1	0	0	0	1

In our case the vector values are numbers between 0 and 1, since they are weighted. For each training example, the score assigned by the examiners is known.

Classifying unseen answers

To classify a new answer we represent it as a vector and find which of the training examples it is nearest to, where the distance measure between two vectors x_n and y_n is defined as:

$$d(x_n, y_n) = \log\left(\sum_{i=1}^{n}(x_i - y_i)^2\right)$$

This method can be generalised to find the k nearest neighbours, and then the most likely score will be that which occurs most frequently among the k neighbours.

Comparison and Discussion of Results

This is an extremely simple classification technique and we would not expect it to work very well. Like all 'bag-of-words' approaches, it completely ignores any higher level linguistic structure and so would represent *wasps lay eggs inside caterpillars* as the same answer as *caterpillars lay eggs inside wasps*, or indeed *eggs lay caterpillars wasps inside*. Nevertheless, it is still doing some work, as the following table shows. We computed a naive *unigram* baseline score, where each answer gets the grade that occurred most frequently in the training set. We give these figures, followed by the KNN score, with the information extraction score for comparison. Note that the information extraction approach was only implemented for 3 out of the 6 test questions, whereas it is no extra effort to do all the questions with the automatic techniques:

Question	12(c)(i)	12(c)(ii)	13(b)(ii)	4(a)	5(a)(ii)	9(c)	Overall
Baseline	56	54	71	67	78	39	60
k=3, +w, +f	75	59	71	72	67	56	67
patterns	n/a	n/a	n/a	94	80	89	88

Note that 'k=3, +w, +f' means that we are using 3 nearest neighbour matching, function words are filtered out and remaining words are weighted by inverted document frequency.

Clearly the pattern-matching method is doing better on the examples it dealt with. This does not mean, however, that the results would be like this for any choice of the feature set. It is possible that if we had more linguistic information represented in the vectors then the results of the KNN technique would improve.

Conclusions

We have demonstrated that information extraction techniques can be successfully used in the task of marking GCSE biology scripts. We have also shown that a relatively naive text classification method can score better than a simple baseline grading technique. There are many refinements to both kinds of approach that can be made: our eventual aim is to try to approach the accuracy of the information extraction method but using completely automatic machine learning techniques.

References

Appelt D. and Israel D. 1999. Introduction to information extraction technology. IJCAI 99 Tutorial.

Burstein J., Kukich K., Wolff S., Chi Lu, Chodorow M., Braden-Harder L. and Harris M.D. 1998a. Automated scoring using a hybrid feature identification technique.

Burstein J., Kukich K., Wolff S., Chi Lu, Chodorow M., Braden-Harder L. and Harris M.D. 1998b. Computer analysis of essays. In *NCME Symposium on Automated Scoring*.

Burstein J., Leacock C. and Swartz R. 2001. Automated evaluation of essays and short answers. In *5th International Computer Assisted Assessment Conference*, Loughborough University.

Callear D., Jerrams-Smith J. and Soh V. 2001. CAA of short non-MCQ answers. In *Proceedings of the 5th International CAA conference*, Loughborough. http://www.lboro.ac.uk/service/ltd/flicaa/conf2001/pdfs/k3.pdf.

Foltz P.W. 1996. Latent semantic analysis for text-based research. *Behavioral Research Methods, Instruments and Computers,* 28(2):197–202.

Foltz P.W., Britt M.A. and Perfetti C.A. 1996. Reasoning from multiple texts: An automatic analysis of readers' situation models. In *Proceedings of the 18th*

Annual Cognitive Science Conference, pages 110–115. Hillsdale, NJ: Lawrence Erlbaum Associates.

Foltz P.W., Laham D. and Landauer T.K. 2003. Automated essay scoring: Applications to educational technology. http://www-psych.nmsu.edu/~pfoltz/reprints/Edmedia99.html. Reprint.

Graesser A.C., Wiemer-Hastings P., Wiemer-Hastings K., Harter D., Person N and the Tutoring Research Group. 2000. Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*, 8(2):87–109.

Jones M. and Jones G. 2001. *Advanced Biology*. Cambridge University Press.

Kanejiya D., Kumar A. and Prasad S. 2003. Automatic evaluation of students' answers using syntactically enhanced LSA. In *Building Educational Applications Using Natural Language Processing, Proc. of the HLT-NAACL 2003 Workshop*, pages 53–60. Association of Computational Linguistics.

Larkey L.S. 1998. Automatic essay grading using text categorisation techniques. In *ACM-SIGIR Inter. Conference on Research and Development inInformation Retrieval*.

Mitchell T. 1997. *Machine Learning*. McGraw Hill.

Mitchell T., Russell T, Broomhead P. and Aldridge N. 2002. Towards robust computerized marking of free-text responses. In *6th International Computer Aided Assessment Conference*. Loughborough.

Page E.B. 1994. Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education*, 62(2):127–142.

Rosé C.P., Roque A., Bhembe D. and VanLehn K. 2003. A hybrid text classification approach for analysis of student essays. In *Building Educational Applications Using Natural LanguageProcessing*, pages 68–75.

Sleator D.K. and Temperley D. 1991. Parsing English with a link grammar. Technical Report. October 1991. CMU-CS-91-196.

Sleator D.K. and Temperley D. 1993. Parsing with a link grammar. In *Third International Workshop on Parsing Technologies*.

Wiemer-Hastings P. and Zipitria I. 2001. Rules for syntax, vectors for semantics. In *Proc. 23rd annual Conf. of the Cognitive Science Society*. Mawhwah, N.J. Erlbaum.

Wolfe M., Schreiner M.E., Rehder B., Laham D., Foltz P.W., Kintsch W. and Landauer T.K. 1998. Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes*, 25(2 and 3):309–336.

Noun-Verb Associations for Noun-Noun Compound Interpretation

DAVID G. S. WRIGHT

---

1. Introduction

This paper describes ongoing work concerning automated interpretation of noun-noun compounds. The approach taken here is based on the application of linguistic knowledge and probabilistic reasoning to produce satisfactory interpretations entirely automatically. A noun-noun compound (e.g. *bread knife*, *concert hall*) consists of two nouns, a modifier followed by a head. In general, and certainly for the purposes of this work, the semantic type of a compound is a subtype of the head noun. The task of interpreting a compound lies in determining the nature of the specialisation. To this end, the modifier noun supplies some semantic cues — a bread knife is a knife somehow associated with bread. For a satisfactory interpretation, we must establish precisely *how* the modifier relates to the head noun. For example, in the compound *bread knife* head and modifier are related by the action of *cutting* or *slicing*. Having established this relation, the interpretation is that of *a knife typically used for the purpose of cutting bread.*

Interpretations for a given compound can vary in their degree of informativeness. This work aims to produce interpretations which are *satisfactory* in their degree of informativeness in that useful inferences may be drawn directly from the interpretation. An interpretation for *bread knife* along the lines of *a knife used for bread* is unsatisfactory, as it does not establish *how* the knife is used for bread and, as such, is only marginally more informative than the compound itself.

1.1 Related work

A long-standing view in the field of noun-noun compound interpretation is that the set of legitimate relations in compounds is small and predetermined. Levi (1978) provides nine semantic labels. Barker and Szpakowicz (1998)

construct their own list of twenty and are able to reduce the interpretation task to a classification exercise based on a similarity measure and a set of preclassified training examples. This approach does restrict the number of relations. Hence, to achieve coverage for all compounds, relations remain vague and are inevitably rather uninformative. Hobbs *et al.* (1993) classify compounds within a more general interpretation scheme applying abductive inference to contextual and world knowledge. The use of knowledge and reasoning in this approach is attractive but is restricted to domain specific tasks owing to the problems of building a comprehensive and unrestricted knowledge base.

In an alternative approach, Johnson and Busa (1996) claim that semantic representations of compounds can be constructed compositionally from lexical feature structures representing the head and the modifier. The relation between head and modifier is identified implicitly in the way the structure of the modifier interacts with the predicates in the 'qualia' structure of the head noun. The literature demonstrates how powerful the architecture is and how detailed (and hence informative) interpretations can be. Lexical approaches however depend heavily on a set of composition rules and the existence of a complete representation of the lexicon, both of which at present must be written by hand.

The work most similar to the framework outlined here, is that of Lapata (2002) who uses a probabilistic model derived from corpus analysis to find interpretations. Lapata restricts herself to nominalisations — a subset of compounds where the head noun is derived from a verb (e.g. *car lover*, *animal behaviour*). The verb from which the head noun is derived, together with the head noun's morphology, determine the relation involved. What remains is to establish which role (subject or object) the modifier plays in this relation. Hence, a car lover is someone who loves cars (object) while animal behaviour is the manner in which animals behave (subject).

1.2 The current approach

The approach outlined in this paper follows Lapata (2002) in that it applies corpus-derived linguistic knowledge and probabilistic reasoning to produce interpretations entirely automatically. The assumption is that the relation between head and modifier is derived from a verb. Unlike Lapata's approach, we do not restrict ourselves to nominalisations where the relational verb is explicit in the form of a deverbal noun. Moreover, where compositional approaches assume that

verbal predicates are given in the head noun's 'qualia' structure, we assume no such formal structures. The verb underlying the interpretation is identified by a less formal semantic notion of *associations* between nouns and verbs. Such associations are formed by observing the co-occurrence of these words in certain syntactic relations (e.g. verb-subject, verb-object) in 'every day' language. We may define the *strength of association* between a noun and a verb based on the frequency with which the noun and the verb appear in a given syntactic relation throughout a suitable corpus. Informally, an interpretation for a noun-noun compound can be made by identifying the verb which is most strongly associated with both head and modifier.

The advantage of this framework over past approaches is that the process relies on no subjective human input and is entirely automatic. The system does not need a hand-crafted structured lexicon and requires no supervised training. Moreover, the system has the advantage of producing fine-grained, informative interpretations and it is not necessary to agree on a set of legitimate relations prior to runtime. What this scheme requires is access to syntactic relational data from a processed corpus. Recent work in corpus linguistics (see next section) makes this possible, providing such a resource through entirely automatic means.

The remainder of this paper is organised as follows. Section 2 describes the syntactic relational data, its acquisition and its processing. Section 3 develops the model for the probabilistic reasoning behind the interpretation procedure. Section 4 illustrates some preliminary results and discusses issues arising from them. Section 5 outlines plans to incorporate conceptual hierarchies to the model and Section 6 concludes with discussion of issues for consideration in further work.

2. The Corpus Data

In order to calculate a measure of strength of association between a noun and a verb, we require frequency counts for the co-occurrence of all words appearing in a corpus in certain grammatical relations (verb-subject, verb-object, etc.). This is precisely the information gleaned from the British National Corpus (BNC) as part of the WASPS project (Kilgarriff and Tugwell, 2001).[1] A shallow parsing pattern-matching technique was applied to a version of the corpus, tagged for part-of-

[1] The author would like to thank Adam Kilgarriff and David Tugwell from the University of Brighton for supplying the data and granting permission to use it.

speech (PoS). Over the entire corpus, frequency statistics were compiled for the words lying in various grammatical relations. For example, the sentence

The dogs slept on the cold ground.[2]

contributes to the count for the noun *dog* featuring as a subject of the verb *sleep*; to the count for the noun *ground* featuring within a PP-attachment (headed by *on*) to the verb *sleep*; and to the count for the adjective *cold* featuring as a modifier of the noun *ground*. Frequency data for the co-occurrence of pairs of individual words in each relation were compiled and recorded as sets of tuples <*Word1*, *PoS*, *Word2*, *Relation*, *Frequency*> where *PoS* refers to the part-of-speech of *Word1*, *Relation* specifies the grammatical relation, and *PoS* and *Relation* together determine the part-of-speech of *Word2*. For the noun-noun compound interpretation task, only grammatical relations involving nouns and verbs were relevant. All other relational information was discarded. What remained was processed into a set of triples of the form <*Noun*, *Verb*, *Relation*> each with an associated frequency count, *Frequency*. Legitimate values for *Relation* were *subj* for subject, *obj* for direct object, and expressions of the form *p_X* representing a PP-attachment headed by the preposition *X*.

2.1 Noise reduction

The version of the BNC used in the WASPS project was the First Edition, which notoriously contains a large number of PoS-tagging errors. These errors are compounded at the parse stage, since the pattern-matching algorithm is not completely accurate. To minimise noise introduced by such errors, it was desirable to filter the data further. To this end, the words in each tuple of the processed data were checked against a computer-readable dictionary. Words in the data which did not appear with the correct PoS in the dictionary were identified and the corresponding tuples were removed from the data. This had the additional effect of removing many proper nouns from the data.

[2] from the project website www.itri.bton.ac.uk/projects/wasps/overview.html

3. The Probabilistic Model

The data collected in the previous section define a probability distribution on triples of the form $<N, V, R>$ where N is a noun, V is a verb, and R is a syntactic relation. The probability of the triple $<N, V, R>$ is given by the observed frequency of that triple in the data as a proportion of the total number of observations. Given this distribution we can formally define the notion of *strength of association* between a noun N and a verb V in a syntactic relation R by the probability $P(N, V, R)$.

We now attempt to construct a plausibility measure for a verb to act as the relation between two nouns in a noun-noun compound. A compound C is represented by the ordered pair $<N_1, N_2>$, N_1 being the modifier and N_2 being the head, (e.g. *<bread, knife>*). An interpretation for C is given by the triple $<V, R_1, R_2>$ where V stands for the verb relating the two nouns and R_1 and R_2 respectively represent legitimate syntactic relations by which the two nouns relate to the verb in this interpretation. For example, one very plausible interpretation for the compound *<bread, knife>* is *<slice, obj, p_with>*. This means that the modifier *bread* appears as the direct object of the verb *slice* and the head noun *knife* appears within a PP-attachment headed by *with*. This interpretation can be paraphrased in the context of the compound as *a knife with which one slices bread*.

Consider now the space of pairs of compounds and interpretations $<C, I>$. Under some derived probability distribution on this space, every pair $<C, I>$ has a probability associated with it, $P(C, I)$. A suitable measure for the plausibility of an interpretation for a given compound is $P(I \mid C)$. The interpretation I for which the expression $P(I \mid C)$ is greatest is judged to be the most plausible.

Given the internal form of compounds and interpretations, the space of pairs $<C, I>$ is equivalent to the space of 5-tuples $<N_1, N_2, V, R_1, R_2>$. The expression $P(I \mid C)$ is given by

$$P(V, R_1, R_2 \mid N_1, N_2) = \frac{P(N_1, N_2, V, R_1, R_2)}{P(N_1, N_2)}$$

For a given compound $<N_1, N_2>$, we wish to maximise this expression over all legitimate values of V, R_1 and R_2. Being independent of these values, $P(N_1, N_2)$

remains constant throughout, so it is sufficient to consider only the numerator. By the definition of conditional probability

$$P(N_1, N_2, V, R_1, R_2) = P(N_2, V, R_2).P(N_1, R_1 | N_2, V, R_2)$$

Under certain independence assumptions, the expression we wish to maximise is thus given by

$$= \frac{P(N_2, V, R_2).P(N_1, R_1 | V)}{P(V)}$$

Each of these three terms is derivable directly from the distribution obtained from the frequency data of the grammatical relations, under the prevailing assumption that nouns and verbs relate to each other within compounds with the same distribution as they do across the language at large.

3.1 The interpretation procedure

Given the formula above as a plausibility measure for interpretations, the procedure involved in finding a suitable interpretation for a compound $C = <N_1, N_2>$ is simply to evaluate the expression above for all legitimate values of V, R_1 and R_2 returning the interpretation $I = <V, R_1, R_2>$ for which the value of the expression is greatest. At this stage, it is necessary to add the requirement that R_1 and R_2 cannot stand for the same relation in a single interpretation. This is to prevent interpretations being produced where the verb has more than one subject or more than one direct object.

In practice, this procedure picks out a verb which is independently most strongly associated with the two nouns involved in the compound, together with syntactic relations establishing the roles played by each of the nouns in the event introduced by the verb. Associations between nouns and verbs — a form of linguistic knowledge — are the *only* source available to the system with which to make its decisions. The decisions made by the system can only reflect what is contained in the corpus data. The results which follow are thus as much a test of the content and suitability of the corpus as a test of the interpretation strategy itself.

4. Preliminary Results

The following are an illustrative selection of interpretations together with equivalent English paraphrases obtained from an initial test set compiled by hand. The test compounds are largely taken from examples of Barker and Szpakowicz (1998), used to demonstrate typical members of their compound classes.

bread knife
<slice, obj, p_with> *a knife with which one slices bread*
concert hall
<stage, obj, p_at> *a hall at which one stages a concert*
student protest
<stage, subj, obj> *a protest staged by students*
student discount
<entitle, obj, p_to> *a discount to which students are entitled*
oil pan
<heat, obj, p_in> *a pan in which one heats oil*

While the literature shows that out-of-context compound interpretation is widely subjective (Downing, 1977) making a quantitative evaluation of these results difficult, it is clear that the interpretations found for the compounds above are perfectly natural — in that a human could easily have made the same interpretation. Moreover, in the sense of being sufficiently informative, the interpretations produced are a lot more satisfactory than the interpretations for the same compounds of Barker and Szpakowicz (1998). Not all of the interpretations found by the system are as valid but even these few examples are encouraging enough to justify the approach and demonstrate the validity of the relational data as a resource for compound interpretation. Even where the output is clearly wrong

death penalty
<kick, p_to, obj> *a penalty which one kicks to death*

the 'thought-processing' involved is usually transparent and natural enough to see where the answers come from and to appreciate why the system outputs what it does.

Given that the system is clearly reasoning logically, it pays to examine the shortcomings in the data as well as the strategy which lead to incorrect results. Such shortcomings are due to a number of factors.

4.1 Noise

There are two types of noise that exist in the relational data gleaned from the BNC. There is, first, what might be described as *natural* noise, being observations in the form of frequency counts for pairs of words that genuinely lie in certain grammatical relations but between which there should be no semantic association. It is a fundamental assumption of the strategy that pairs of associated words should co-occur much more frequently than pairs of unassociated words. Nevertheless, it is inevitable that the observations for pairs of unassociated words will contribute to the noise in the data. Note that, since this noise is inherent in the corpus, it cannot easily be removed.

Secondly, there is the *artificial* noise, added to the genuine data as a result of errors in the processing. The issue of tagging errors in the unprocessed corpus has already been addressed (see Section 2.1). Even after attempts at noise minimisation by checking against a computer-readable dictionary, the noise from tagging errors is, in some cases, stronger than any genuine associations. The result

lemon juice
<wedge, subj, obj> *juice which is wedged by lemons*

is obtained because, on inspection of the corpus, the word *wedge* in the phrase *lemon wedges* is repeatedly, and incorrectly, tagged as a verb. The parser naturally analyses *lemon* as the subject of the verb *wedge* and an invalid association is formed. The filtering step outlined in Section 2.1 is useless in this case since the word *wedge* rightly appears in the dictionary as both noun and verb so the tagging error is not spotted.

In addition, since the pattern-matching technique is itself not entirely accurate even when the tagging is correct, yet more noise accumulates at the parse stage. While the example above demonstrates that with the current data, the level of noise is damagingly significant, it is conceivable, at least in principle, to improve the performance of the parser and apply it to a corpus with fewer tagging errors (the more recent World Edition of the BNC being one such corpus), and this kind of

noise may be dramatically reduced. In the meantime, it is essential that we find ways of minimising, or at least diluting, the existing noise.

4.2 Data sparseness

Despite the size of the BNC, it is inevitable from the sheer number of words in the lexicon that even very strongly related words are only going to appear in the same grammatical relation a small number of times. This is an example of the familiar problem of data sparseness where information contained in the corpus is distributed so thinly that it is difficult to observe the useful content over the noise. Given the significance of the amount of noise in the data outlined above, data sparseness is a serious issue in this case. What is required is a means of concentrating or focussing the genuine content in the corpus while controlling the level of noise. The next section describes an extension of the existing framework placing more emphasis on concepts rather than the lexicon in a direct attempt to alleviate the problems of data sparseness and noise.

4.3 Word sense issues

It is an inevitable result of working at the lexical level that associations are made between words rather than word senses. Polysemous words will form sets of associations relating to each of their senses. No explicit word sense disambiguation is carried out here as we lack the necessary machinery. It is hoped that in the noun-noun compound interpretation task, the associations which head and modifier hold in common will be enough to find a single, sensible interpretation even with polysemous words. After all, even out of context, the intended sense of *bank* in *bank loan* is clear to the human interpreter despite the ambiguity of the word *bank*. So there must be enough information in the semantic cues given by the head noun *loan* to disambiguate between word senses. Nevertheless, as the *death penalty* example above illustrates, the corpus can often throw up unexpected associations across word senses.

5. Incorporating Conceptual Hierarchies

The theory behind the mechanism built up so far is based on the assumption that we interpret noun-noun compounds by identifying a verb which we associate with both nouns and by fitting the two nouns into a predicate structure derived from that verb. Despite its semantic motivation, the measure of strength of association between nouns and verbs is ultimately a purely statistical measure governed by frequency of co-occurrence in the corpus linked to the semantic notion by the assumption that co-occurrence based on the underlying assumption that co-occurrence of words in a *syntactic* relation has a strong correlation with the words having a *semantic* association.

This approach is justified on the basis that we, as humans, do make mental associations between objects and certain activities — we *know* that knives are used for cutting and that bread needs slicing. To mimic the psychological model more precisely, however, we must consider that the mental associations which we develop exist at a conceptual level rather than between lexical items. It is not strictly between the words *knife* and *cut* that we form an association. It is rather with the set of knives and knife-like objects that we associate the generic action of cutting, slicing and sawing.[3] Moreover, the action of making, manufacturing and creating is associated with the entire class of *artifacts* — a class which subsumes many subclasses of concepts and is represented by a whole host of realisations in the lexicon.

A better approach would therefore be to attempt to learn associations between concepts and apply these to the compound interpretation task. The problem now is how to learn associations between concepts when all we have is lexical co-occurrence data. The assumption must now be that co-occurrence of two lexical items not only strengthens an association between the words at the lexical level but also reinforces the strength of association between all concepts for which these words are lexical realisations.

The model to be adopted is that of a conceptual hierarchy. Very general concepts, like the concept of a generic *object*, lie at the top of this hierarchy, while the more specific concepts (e.g. *knife)* lie at the bottom. Concepts are related in the hierarchy by the property of subsumption. One concept is subsumed by a (parent) concept if it inherits all of the parent's properties along with possessing some more

[3] Alternatively we may say that we do form links at the lexical level but *only* because of the more fundamental associations that exist between the underlying concepts.

specific, properties of its own. The concept *bread* is subsumed by the more general concept of *food*, which is, in turn, subsumed by the concept of *substance* or *matter*. The word *bread* is trivially a lexical realisation of the *bread* concept, but also a lexical realisation of all concepts which subsume the *bread* concept, all the way up the hierarchy. Hence, every lexical item is a lexical realisation of a number of concepts at different levels of the hierarchy. Conversely, every concept in the hierarchy commands a set of lexical realisations, being precisely that set of lexical items underneath it in the hierarchy.

Returning to the task of learning conceptual associations from lexical co-occurrence data, we construct two conceptual hierarchies – one for nominal concepts covering the nouns, and one for verbal concepts covering the verbs. The association between a nominal concept and a verbal concept can now be considered to be the accumulation of the lexical associations of all pairs of words underneath the two concepts in the nominal and verbal hierarchies. By accumulating the co-occurrence frequency values for every pair of such lexical items in the hierarchy, we may define a measure of *cumulative strength of association* between the two concepts.

5.1 A model for conceptual hierarchies

A conceptual hierarchy is modelled using the linguistic resource WordNet (Fellbaum, 1998). Concepts are represented in WordNet as sets of synonymous word senses (*synsets*). WordNet comprises a number of hierarchies and networks relating these synsets according to various semantic relations. One such relation us that of *hyponymy*, defined on both nouns and verbs. Hyponymy is closely connected to the notion of conceptual inheritance/subsumption and it is its structure in WordNet on which we model our conceptual hierarchies.

5.2 Compound interpretation within the conceptual framework

Given a noun-noun compound, the interpretation strategy at the lexical level outlined in Section 3, consisted of identifying a verb which was independently most strongly associated with both nouns. At the conceptual level, we look for the strongest cumulative association between a single verbal concept and nominal concepts subsuming the two nouns. As already observed, a noun can be

represented by nominal concepts at any level all the way up the hierarchy. To ensure that the notion of strength of association between a noun and a verbal concept continues to be well-defined, the association is said to exist at the conceptual level for which the cumulative strength of association between concepts is greatest. For example, given the compound *oak table*, the lexical framework returns

oak table
<dine, subj, obj> *a table dined by oak*

which is clearly nonsense. A sensible interpretation for this compound is *a table made of oak*. Conceptually, the *is_made_of* relation most naturally applies between man-made objects (WordNet's *artifact* class) and suitable materials. Hence, the associations between *oak* and *table* and the verbal concept *make* exist at the levels of *material* and *artifact* respectively. Because the conceptual associations involved in the interpretation of this compound are so general, and hence so far removed from the specific lexical associations, it is perhaps not surprising that the interpretation within the lexical framework is poor.

6. Issues for Further Work

This paper has outlined a framework for the automatic interpretation of noun-noun compounds using probabilistic reasoning (Section 3) and linguistic knowledge gleaned from the BNC (Section 2). Preliminary results (Section 4) are encouraging in principle but throw up a number of immediately apparent issues. The combined problem of data sparseness and the prevalence of noise in the data has been addressed in moving from a lexical to a conceptual framework (Section 5). It is not clear how this move will help to alleviate the problem of word sense ambiguity — indeed, it is anticipated that introducing a WordNet-based conceptual framework will in fact add to the problem — but addressing word sense issues in a conceptual framework is a matter of ongoing work.

It is worth at this stage addressing some of the basic assumptions on which this approach is based. The most fundamental of these, as justified in the introduction, is that an interpretation for a compound can be found be identifying a relation between head and modifier in the compound. It is generally accepted that, while counter-examples do exist, the majority of compounds are indeed *transparent* — in

that a relation genuinely does exist, justifying the effort in looking for one. Further assumptions concerning interpretations made within this framework are:

- that the interpretation is unaffected by context (equivalently, that there is a single context-neutral interpretation and that this is the only interpretation of interest);
- that it is at all necessary to find an interpretation (and hence identify the relation) in all cases.

Context

The issue of context is ignored by many authors in the field. Compounds can of course have multiple readings and context often plays a role in determining which is the most plausible. Indeed, it is usually possible to invent a context around any desired interpretation (consider *apple-juice seat* from Downing (1977)). It is a reasonable claim that, in general, a single reading is highly preferable in a context-neutral environment. In the case of the compound *bread knife*, for example, there seems (to this author) to be only one plausible context-neutral interpretation. Yet Downing (1977) shows experimentally that, given similar 'obvious' examples, subjects do *not* agree in their individual suggestions. Two issues arise from this. Firstly, if human subjects do not agree then there cannot be a single *correct* interpretation even in a context-free environment. In isolated tests then, any interpretation judged plausible cannot be deemed invalid. This has implications for the development of a quantitative evaluation of any interpretative system. Secondly, if context indeed plays a role in affecting plausibility judgements, then context analysis must be built into a comprehensive interpretation mechanism. Such a mechanism will need to be a lot more sophisticated than the present model to cope with contextual effects.

Lexicalisation

Owing to the productive nature of compounding in English, it is important to have the machinery available to interpret unseen cases. On the other hand, some compounds appear in the language so frequently that it is sensible for them to be *lexicalised* — treated as single lexical units (albeit with a certain amount of morpho-semantic structure) so that the meaning of the whole compound is 'hard-

wired' into the lexicon and can be obtained by a simple look-up process, thereby avoiding the need to work out an interpretation at all. The need for a separate treatment of lexicalised compounds is illustrated by those compounds which are *conventionalised.* In such compounds, the relation between head and modifier is often quite complex. The compound *home secretary*, meaning a secretary (in the *cabinet minister* sense) with responsibility for home affairs, is one such example. If this compound were not lexicalised then a lot of knowledge about society would be required to be able to choose this interpretation over any other.

Unfortunately, lexicons available to automated interpretation systems are not generally going to include lexicalised noun-noun compounds as individual entries. After all, the process of lexicalisation is a gradual linguistic process so it is impossible to prescribe which compounds are lexicalised at any one point in time. Without a vastly expanded lexicon, therefore, an automated system has no way of identifying which compounds are lexicalised and hence has no choice but to attempt to interpret all compounds it encounters. In cases such as the example above, without the necessary knowledge of society and culture, the system cannot expect to perform.

6.1 Further applications and extensions

Lexicalised compounds (and particularly the subset of conventionalised compounds) demonstrate that relations within compounds are more complex than can be represented by a single verb (or verbal concept). Indeed, to interpret such compounds, more knowledge than simple associations between nouns and verbs is needed — in the hardest cases, detailed cultural and social knowledge is required. What is outlined here is a first step towards genuine, feasible knowledge-based reasoning for the automatic resolution of linguistic constructions at the level of pragmatics. We say 'feasible' in the sense that the knowledge is automatically acquired, not hand-crafted and hence consistent and reproducible. Recent work in automatically learning domain-restricted knowledge (Pulman, 2000) illustrates the potential for applying encyclopaedic knowledge to the interpretation task where simple linguistic knowledge is not enough. While the mechanism outlined in this paper is particularly geared towards the interpretation of noun-noun compounds, there are some linguistic phenomena to which this method can easily be applied. Two of these are possessive constructions and the resolution of metonymy.

Possessives

The default interpretation for most possessive constructions is concerned with ownership or genuine possession. This is by no means the only interpretation in some contexts, however. The phrase *John's horse* of course may mean *the horse John owns*, but may also mean *the horse John is riding* or *the horse John trains* or indeed *the horse John has bet money on*. All of these interpretations pick out a verb particularly associated with horses (or some general concept subsuming *horse*) and with people, in the required grammatical relations. It is easy to see how the linguistic knowledge and reasoning mechanism as used for noun-noun compound interpretation may be applied directly to the interpretation of possessives.

Metonymy

It is not a huge leap either to see how similar knowledge about verbs and the semantic types of their arguments (i.e. which types are *most strongly associated* with which verbs) can be applied to the resolution of metonymy (Hobbs *et al.*, 1993). Given the sentence,

I am parked around the corner.

the knowledge required to establish that it is the speaker's *car* that is parked around the corner is that cars are things that are typically parked and that we can relate the speaker to the car (just as in the case of possessives) by the association with the verbs *drive* or *own* or indeed with *park*. Hence, while the relations involved in metonymy are more complex, there is nothing more required in terms of the type of knowledge and its application to produce perfectly satisfactory interpretations.

References

Ken Barker and Stan Szpakowicz. 1998. Semi-Automatic Recognition of Noun Modifier Relationships. In *Proceedings of the 36th Annual Meeting of the ACL*

and 17th International Conference on Computational Linguistics (COLING/ACL-98).

Ann Copestake and Alex Lascarides. 1997. Integrating Symbolic and Statistical Representations: The Lexicon Pragmatics Interface. In *Proceedings of the 35th Annual Meeting of the ACL.*

Pamela Downing. 1977. On the Creation and Use of English Compound Nouns. *Language* 53(4), pages 810–842.

Cécile Fabre and Pascale Sébillot. 1995. Calculability of the Semantics of English Nominal Compounds: Combining General Linguistic Rules and Corpus-based Semantic Information. Technical Report RR-2742.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database.* The MIT Press.

Jerry R. Hobbs, Mark Stickel, Douglas Appelt and Paul Martin. 1993. Interpretation as Abduction. *Artificial Intelligence* 63, pages 69–142.

Michael Johnston and Federica Busa. 1996. Qualia Structure and the Compositional Interpretation of Compounds. In *Proceedings of the ACL SIGLEX workshop on breadth and depth of semantic lexicons.*

Adam Kilgarriff and David Tugwell. 2001. WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. In *Workshop Proceedings COLLOCATION: Computational Extraction, Analysis and Exploitation, 39th ACL & 10th EACL*, pages 32–38.

Maria Lapata. 2002. The Disambiguation of Nominalisations. *Computational Linguistics* 28(3), pages 357–388.

Judith Levi. 1978. *The Syntax and Semantics of Complex Nominals.* Academic Press, New York.

Richard E. Neapolitan. 1990. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms.* Wiley.

Stephen G. Pulman. 2000. Statistical and logical reasoning in disambiguation. *Philosophical Transactions of the Royal Society Series A*, 358(1769) pages 1267-1280.