
The Ohio State University

Working Papers in Linguistics

No. 59

Edited by
Marivic Lesho
Bridget J. Smith
Kathryn Campbell-Kibler
Peter W. Culicover

The Ohio State University
Department of Linguistics

222 Oxley Hall
1712 Neil Avenue
Columbus, Ohio 43210-1298 USA

Winter 2010

© reserved by individual authors

The Ohio State University WORKING PAPERS IN LINGUISTICS

Working Papers in Linguistics is an occasional publication of the Department of Linguistics of Ohio State University containing articles by members (students and faculty) of the department. To download back issues of the OSU WPL series, please see our web page at:

<http://linguistics.osu.edu/research/publications/workingpapers/>

Information Concerning OSDL OHIO STATE DISSERTATIONS IN LINGUISTICS

Since October of 1994, dissertations that have been written by students in the OSU Linguistics Department since 1992 have been distributed by the graduate student-run organization OSDL. As of September 2006, we no longer provide hard copies of dissertations. Instead, we provide them in downloadable format at the following web page:

<http://linguistics.osu.edu/research/publications/dissertations/>

If you have any questions about OSDL or any of the dissertations it distributes, please email: osdl@ling.ohio-state.edu

INTRODUCTION

This volume of the Ohio State Working Papers in Linguistics continues to build on the revival of the Working Papers, which started with issue 58. This issue reflects the diversity of interests within the department, and wraps up the backlog caused by the hiatus in publishing. The issue is, as we used to name them, a *varia* issue, combining some older papers (Joseph & Lee, Riha,) with some newer papers (Klippenstein, Sampson, Baker & Brew), and representing multiple sub-disciplines in the field of linguistics. Multiple languages are also analyzed in this volume: Greek, Albanian, Early Modern English, Chinese, Japanese, and modern English. We have ordered the papers by a loose division of focus. The first two papers deal with phonology. The third and fourth papers focus on morphology. The final paper is computational in nature.

We have enjoyed resuming regular publication of the OSUWPL, and look forward to the next issue, which will be out in Fall 2010. For subscription information and back issues, please visit us on the web at:

<http://linguistics.osu.edu/research/publications/workingpapers>

Ohio State University Working Papers in Linguistics

No. 59

Table of Contents

Information concerning OSUWPL.....	iii
Information concerning OSDL.....	iii
Introduction.....	iv
Table of Contents.....	v
Greek ts/dz as internally complex segments: Phonological and phonetic evidence..... Brian D. Joseph & Gina M. Lee	1
Word-initial consonant clusters in Albanian..... Rachel Klippenstein	10
The Early Modern English genitive <i>its</i> and factors involved in genitive variation..... Selena Sampson	33
Lettered words in Chinese: Roman letters as morpheme-syllables..... Helena Riha	44
Multilingual animacy classification by sparse logistic regression..... Kirk Baker and Chris Brew	52

**GREEK TS/DZ AS INTERNALLY COMPLEX SEGMENTS:
PHONOLOGICAL AND PHONETIC EVIDENCE***

Brian D. Joseph & Gina M. Lee
The Ohio State University

Abstract

The “affricate dream” of Householder (1964), in which Modern Greek *ts/dz* are reduced to clusters of independently occurring segments (thus, *ts* is analyzed as /t + s/), is examined here in the light of two types of evidence not previously considered: instrumental measurements of the duration of the sounds in question

* This paper was written over 20 years ago, based on work that began in 1986, and it was presented at the Annual Meeting of the Linguistic Society of America in New Orleans in December 1988. It was originally intended for publication in a planned OSU WPL volume on Greek that never materialized, and the authors turned their attention to other projects. Since relatively little has been published in the intervening two decades on this particular issue in Greek phonology using the sort of evidence presented here (from instrumental phonetics and dialectal sound changes), it was thought appropriate to dust this off and present it in this form to the linguistic world. This decision is justified by the fact that the 1988 LSA presentation has been cited in the most definitive survey of research on Greek phonetics to date, Arvaniti (2007), where the author (pp. 114-117) summarizes the body of studies—four in all—that have dealt with the phonetics of the vexing problem of *ts* and *dz* in Greek: (her own) Arvaniti (1987), the LSA presentation Joseph & Lee (1988), Fourakis, Botinis, & Nigrianaki (2002), and Tserdanelis (2005). Also relevant are the as-yet unpublished Fourakis 2004 (based on Fourakis et al. 2002) and Joseph & Tserdanelis 2006. Work on the phonology of these sounds has been summarized recently by Malikouti-Drachman (2001). In part since the results of this paper have been cited in its 1988 (and largely unavailable) form, it seemed best to offer this version with little updating from a theoretical perspective, though with some bibliographic updating. In any case, moreover, it is our belief that the facts pointing to the analysis offered here should be of interest to phonologists of any theoretical persuasion and should be able to be fit into any theoretical framework. We owe a huge “thank you” to Marivic Lesho for her careful editing and for her work on making Figure 1 and to Adam Clark-Joseph for invaluable help with some of the statistics.

Brian D. Joseph and Gina M. Lee. Greek *ts/dz* as internally complex segments: Phonological and phonetic evidence. *OSUWPL* Vol. 59. pp. 1-10.

compared with related sounds, and the proper formulation of a dissimilatory dialectal sound change. This evidence shows that the best analysis recognizes these sounds as single segments but with internal complexity, as suggested, but not overtly argued for, in Joseph & Philippaki-Warbuton (1987).

1. Introduction

A long-standing problem in Modern Greek phonology concerns the analysis of the voiceless dental *ts* and its voiced counterpart *dz*.¹ Like similar sounds in many of the languages of the world, Modern Greek *ts* and *dz* show some characteristics that align them with other stop + sibilant clusters—in particular, *ps* and *ks*. At the same time, though, they present some traits that differentiate them from these clusters, thus suggesting status as single segments (i.e. [tʰ]/[dʰ]). Householder (1964) labeled attempts by linguists to reduce these sounds to clusters of independent segments, as illustrated in (1), the “affricate dream”:²

$$(1) \quad ts = /t/ + /s/ \qquad dz = /d/ + /z/$$

Some version of the affricate dream is generally preferred, for instance by Newton (1961, 1972), Setatos (1974), Arvaniti (1999), and Malikouti-Drachman (2001), though there are dissenters who accept the affricate analysis (e.g. Householder himself).

Among the indicators of cluster-like status are the following considerations. First, the range of clustering possibilities that voiceless stops enter into with fricatives shows a gap in the dentals. *p* + *s* and *k* + *s* both occur quite commonly, and even the combination of *t* + *θ* occurs marginally, as the examples in (2) indicate. A cluster analysis of Greek *ts* would thus fill this gap.³

$$(2) \quad \begin{array}{llll} \underline{k}s\acute{e}ro \text{ ‘know’} & \underline{p}s\acute{e}lno \text{ ‘chant’} & at\theta\acute{i}s \text{ ‘Attica’} & \underline{t}simb\acute{o} \text{ ‘pinch’} \\ \underline{a}ks\acute{i}a \text{ ‘value’} & \underline{t}aps\acute{i} \text{ ‘pan’} & (\text{underwear}) & \underline{k}u\acute{t}s\acute{o}s \text{ ‘lame’} \\ fl\acute{o}ks \text{ ‘fire’} & k\acute{o}nops \text{ ‘mosquito’} & \text{brand name)} & \underline{b}ats \text{ ‘slapping noise’} \end{array}$$

Second, a cluster analysis of *ts* as *t* + *s* explains an otherwise curious fact about *ts*. Greek tolerates a fairly wide range of clusters involving voiceless stops, including, in word-initial position, the sequences [str-, spr-, skr-, skl-, skn-, tm-, pn-, kn-, tr-, pr-, kr-], among others. However, *ts*, as well as *dz*, for that matter, does not participate in any clustering possibilities: for example, there are no words with *tsr- or *tsl-. In this way,

¹ Throughout we write these sounds in italics when referring to them in a nontechnical way, since the use of slashes or square brackets would imply that certain analytic decisions had been made, when in fact the point of this exercise is to explore some evidence relevant for those decisions.

² Some details of the claims regarding the voiced *dz* depend on other assumptions and claims that go well beyond the rather limited scope of this paper. Other possibilities exist for *dz*, depending on the resolution of Householder’s “voiced stop dream” (by which the voiced stops of Greek are reduced to sequences of nasal + voiceless stop), e.g. /nt/ + /z/ or even /d/ (or /nt/) + /s/.

³ As a result of the phonotactics of colloquial Greek, word-final examples of *ps* and *ks* do not occur; the examples given are from the “high-style”, generally literary, variety of Greek known as *katharevousa*. The example with word-final *ts* is an onomatopoe, though now some loanwords, e.g. *mats* ‘(football) match’, have this sequence also.

ts, and *dz* too, pattern with the clear voiceless stop + sibilant clusters, for there are no Greek words with *psr-, *psl-, *ksr-, *ksl-.

Running counter to these cluster-like indications for *ts* and *dz*, though, are a few facts that show these sounds to be different from *ps* and *ks*. From the standpoint of morphophonemics, it is noteworthy that sequences of the sounds that in a cluster analysis would constitute the *ts*, namely *t* + *s*, behave differently across a morpheme boundary from the sequences *p* + *s* or *k* + *s*. The relevant facts are given in (3), where it can be observed that the combination of morpheme-final *t* plus morpheme-initial *s* yields an *s*, whereas similar sequences with the labial or the velar voiceless stop yield clusters.

- (3) *fós* <= /fot + s/ ('light' + NtrNomSg; cf. NtrNomPl *fót-a*)
θésame <= /θét + s + ame/ ('put + Prfve + 1PIPst; cf. Pres *θét-ome*)
próvlepsa <= /provlep + s + a/ ('foresee' + Prfve + 1SgPst)
pléksame <= /plek + s + a/ ('knit + Prfve + 1PIPst)

Similarly, there are suffixes that begin with *ts* or *dz*, e.g. the hypocoristic *-tsos* and the occupational *-dzis*, as given in (4):⁴

- (4) *Mí-tsos* 'Jimmy' (from *Dimítris*)
Kó-tsos 'Connie' (from *Konstandínos*)
taksi-dzís 'taxi-driver'

There are, however, no suffixes that begin with *ps* or *ks*. It is significant, moreover, that here are suffixes with initial clusters, e.g. the feminine actor-noun suffix, as in (5):

- (5) *telefoní-tria* 'telephone operator' (cf. *telefoní-sa* 'I telephoned')

What this shows is that the absence of *ps* and *ks* from suffix-initial position is not a systematic fact about clusters in general in Greek but rather seems to be a matter relevant only to clusters with sibilant second members. That is, no Greek suffix begins with a stop + sibilant cluster; thus, since *-tsos* and *-dzis* occur, *ts* and *dz* by this criterion cannot be clusters.

Given these conflicting characteristics, it is not surprising that the rather considerable literature on this subject in Greek shows conflicting conclusions on the part of various analysts. In general, linguists have arbitrarily given more weight to one or the other type of behavior and have drawn their conclusions accordingly. For example, as noted above, Newton (1961, 1972), Setatos (1974), Arvaniti (1999), and Malikouti-Drachman (2001) all opt for a cluster analysis,⁵ while Householder (1964), on the basis of the morphophonemic evidence, opts ultimately for the single-segment analysis.

A solution to this dilemma was suggested by Joseph & Philippaki-Warburton (1987: 238), where it was proposed that, like affricates in many languages, Modern Greek

⁴ Actually, the occupational suffix, of Turkish origin, also has a *-ts*-initial allomorph after voiceless stops, e.g. *kaik-tsis* 'owner of a *kaiki* (a type of boat)'.

⁵ Either overtly stating they are so doing, or adopting it implicitly, via the absence of any mention of affricates in the phonemic inventory.

ts and *dz* constitute single segments but with a complex internal structure.⁶ Such an internally complex segment, as represented in segmental (“linear”) phonology, following Campbell (1974), is given in (6).

$$(6) \quad ts = [[t] [s]] \qquad dz = [[d] [z]]^7$$

One possible reinterpretation of this notion autosegmentally is given in (7) for *ts*, with a similar representation for *dz*.

$$(7) \quad \begin{array}{ccc} \text{CV-tier} & \text{C} & \text{(one element, i.e., unitary)} \\ & / \ \backslash & \\ \text{Segmental tier} & t \quad s & \text{(two elements, i.e., complex)} \end{array}$$

However, Joseph & Philippaki-Warburton (1987) merely asserted this possibility as a way out of the dilemma without giving any definitive argumentation to support this claim, beyond the observation that it allows these elements to have properties of both clusters and nonclusters. Accordingly, we present here one type of argument to support the Joseph & Philippaki-Warburton proposal—namely, the phonetic evidence concerning the duration of *ts* and *dz*. We present as well some diachronic evidence from a dialectal sound change that also is consistent with this proposal.

In presenting this evidence, we are attempting to address what has been a difficult problem internal to Greek linguistics—one that has generated considerable debate in the literature—without trying to draw general conclusions about the representation such sounds should or should not have in some particular theoretical framework or other. We do feel, however, that this evidence from one language may well be compatible with similar findings from other languages, and thus relevant for a general theory of complex (or contour) segments cross-linguistically.

2. Phonetic evidence

In undertaking this investigation, we are working under the assumption — one shared by many linguists, we believe, though not necessarily all — that wherever possible, phonological constructs should be closely tied to the phonetic reality of the elements they represent. Our approach, therefore, closely parallels such work as Hankamer & Lahiri (1986) or Miller (1987), as well as the work that now falls under the general rubric of “laboratory phonology”.⁸ To gain further insight into the status of Modern Greek *ts* and *dz*, we conducted an experiment involving five native speakers who were graduate students or junior faculty at The Ohio State University. Four spoke Athenian Greek and a fifth, who was fluent in Standard Modern Greek, natively spoke a northern Greek dialect; still, as the results show, dialect was not a factor.

⁶ Or more accurately perhaps in the terminology widely used since Sagey 1986, contour segments (with ordered multiple articulations). Malikouti-Drachman (2001) uses this terminology.

⁷ Assuming /d/ as underlying; [[d] [s]] is also possible.

⁸ See, for instance, the Cambridge University Press series, *Papers in Laboratory Phonology*, with several volumes based on the now biennial conference on work in this framework, Kingston & Beckman (1991) being the first such volume.

Each speaker read a corpus consisting of fifty-five sets of words, each set containing five words. The words were chosen to give examples of the primary sounds under investigation, *ts* and *dz*, as well as the (presumably) clear clusters *ps* and *ks*, and the single stops and fricatives /p, t, k, s, z/. The sound [d] was not considered because the medial occurrences of [d] was rare for our speakers, often being pronounced, by them as well as by other Greeks, with some degree of prenasalization or with a full preceding nasal.⁹ The participant in this experiment had an extremely small number of cases of “pure” [d] (i.e. not accompanied by a nasal in some form).

We recorded their utterances in an anechoic chamber and digitized the recordings at 10k Hz. Using a waveform editor, we measured the duration of these consonants in word-medial position. We considered only word-medial consonants for two reasons.¹⁰ First, it is easier to measure these sounds word-medially than word-initially. Second, there is a greater variety of words containing these sounds in medial position than in initial or final position (see footnote 3). Within each five-word set, we measured consonant duration in the second, third, and fourth words only, disregarding the first and last words because of possible effects of reading list intonation.

If the duration of *ts* turned out not to be particularly different from that of the stop + sibilant clusters, and if all differed from the single segments, then there would be reason to believe that *ts* and (by extension) *dz* are clusters. If, on the other hand, the duration of *ts* turned out to be quite smaller than that of the clusters, then there would be reason to believe that *ts* and *dz* are not clusters.

The results show that the duration of *ts* was, for all speakers, longer than that of the single segments and, importantly, shorter than that of the clusters /ps/ and /ks/. Figure 1 shows the results for all of the speakers taken together. On average, for all speakers, *ts* was 60.66 ms shorter than /ps/ and 53.04 ms shorter than /ks/. *ts* was 36.24 ms longer than (singleton) /t/ and 17.32 ms longer than (singleton) /s/. T-tests indicate that, for all speakers, the difference between the durations of *ts* and the stop + sibilant clusters was significant at the .01 level.

For all speakers, the difference between *ts* and /t/ was significant at the .05 level. For two speakers, the difference between *ts* and /s/ was significant at the .05 level. For all speakers, the duration of *dz* was on average 41.24 ms longer than /z/; the difference was significant at the .01 level for four speakers.¹¹

The experimental results therefore suggest that Greek *ts* and *dz* are not phonetically like clusters, nor are they phonetically like single segments, but rather are in

⁹ See Arvaniti & Joseph (2000, 2002, 2006) for some discussion of trends in the realization of voiced stops in Greek in the past thirty years.

¹⁰ Note that Arvaniti (1987), an instrumental study of clusters in Greek, looked at initial clusters only; see below for brief discussion of her results.

¹¹ One further comparison was made with [tr] clusters by way of gauging the duration of other combinations with /t/; we found that the [tr] duration for a given speaker was significantly longer than the *ts* duration ($p = 0.0236$; matched differences t-test, $df = 4$).

between clusters and segments. However, *ts* and *dz* appear phonetically to be more like segments than clusters.

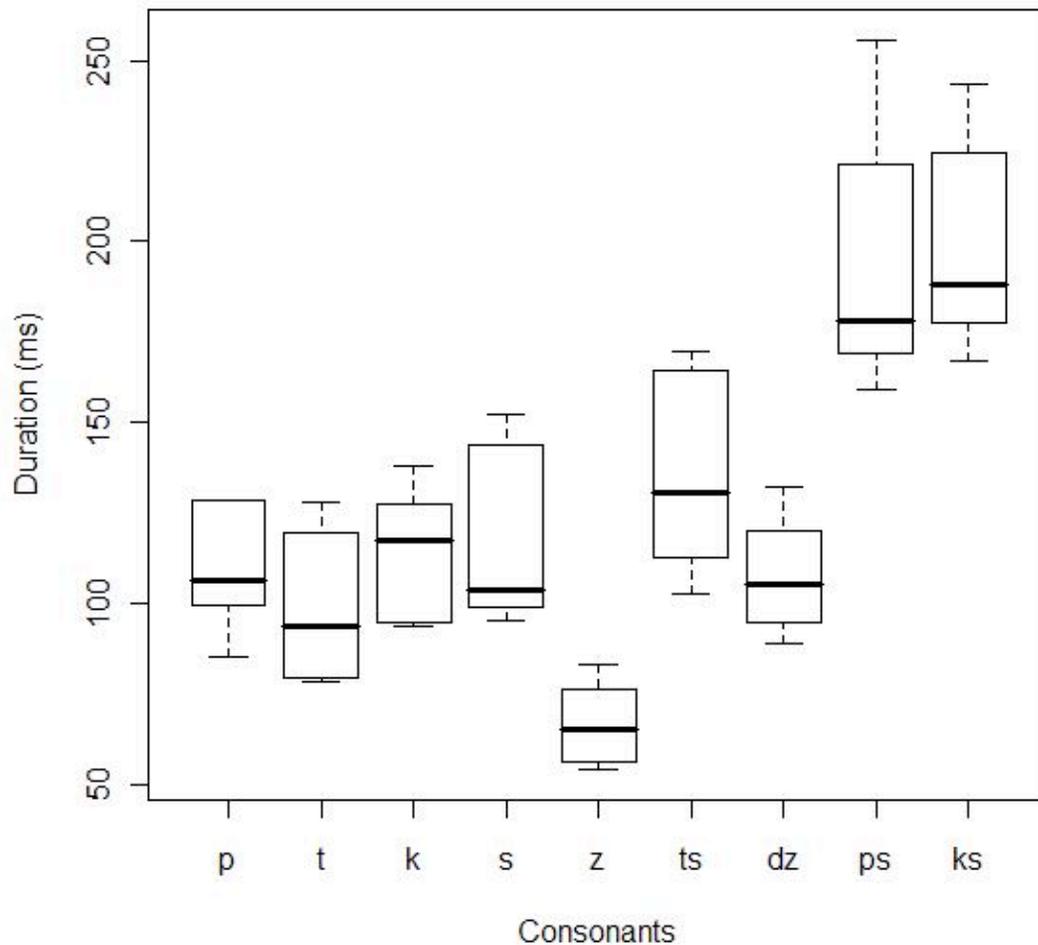


Figure 1. Duration results of all five speakers

As noted above (see footnote *), there are relatively few instrumental studies of Greek that have focused on the “affricate dream”, but to some extent they are consistent with what is reported here concerning a different status for *ts/dz* from that seen with *ps* and *ks*. Arvaniti (1987: 38), for instance, comments: “The present data seem to suggest (although this is rather speculative) that /ts/ is produced differently from /ps, ks/.” Among other things, she notes that “/ts/ as a cluster is significantly shorter than /ps/ and /ks/ for both subjects”. Admittedly, the evidence overall is not unproblematic,¹² and one could take a position that the phonetics are not an essential part of a phonological analysis, given that phonology can be taken to deal with abstract units and not the concrete physical realizations per se; nonetheless, the general outlines of the possible contribution of phonetics to the “affricate dream” should be clear.

¹² As Arvaniti (2002: 115) points out, however, “the shortening observed in [ts] also affects [st] when compared to [sp] and [sk]”, and this constitutes a problem for accounts of *ts* that draw on duration evidence.

3. Dialectal evidence from Cypriot

Further support for an analysis of *ts* and *dz* as internally complex (“contour”) units comes from some diachronic developments in Cypriot Greek, based on the description given in Pantelides (1923). In a few words, Cypriot [t] corresponds to Standard Greek *ts*, as seen in (8):

- (8) titsirízo ‘sizzle’ (Std. tsitsirízo)
 titsín ‘meat; breast’ (Std. tsitsí)

In the words given in (9), Cypriot [t] corresponds to Standard Greek [k], presumably from a prior stage of palatalization to *ts’* (the sound seen farther on in each word in (9), corresponding to Standard Greek [k]/[t] in palatalizing contexts):

- (9) tirts’éllin ‘ring’ (Std. krikéli)
 terats’ja ‘carob tree’ (Std. keratjá)

The diachronic sound change that led to these correspondences involved a dissimilation of [t + s(’)] to [t], triggered by a following [t + s(’)], and it can be formalized as in (10):

- (10) t + s => t / __... t + s

This formulation of the change in a cluster analysis of *ts*, however, is rather ad hoc, or at least more complicated than it might be otherwise, in that [t] needs to be stated both in the input and in the conditioning environment. Moreover, /ps/ and /ks/ do not undergo or condition this change. Based on these facts, *ts* and (by extension) *ts’* cannot be clusters.

Furthermore, as (11) indicates, there is a Cypriot word that shows *ts’* (corresponding, again, to Standard Greek [k] via palatalization) dissimilating to [t] in the context of a following [s]:

- (11) teparíssin ‘cypress tree’ (Std. kiparísi)

As formalized in (12), under a unitary segment analysis of *ts* as $\widehat{[ts]}$ (though the Americanist and Slavist [c] is given below as typographically more congenial here), the change requires an unnatural statement since there is no clear relation between the triggering segment (an [s]) and the change that occurs (c => t):

- (12) c => t / __ ... s

As seen in (13), however, a more natural rule can be formalized by taking *ts/ts’* as internally complex segments, whereby there would be an overtly represented sibilant portion of the complex unit that could be lost via dissimilation in the context of a following sibilant.

- (13) [[t] [s(’)]] => [[t] [Ø]] / __ ... [s]

4. Conclusion

The two pieces of evidence presented here — the results of instrumental measurements and the diachronic dialectal evidence of a dissimilatory Cypriot sound change — do not by themselves prove the superiority of one analysis of Greek *ts/dz* over another, and as some of the discussion above indicates (see especially footnote 12 and some of the references in footnote *), there are problematic aspects to the instrumental analysis. But when considered along with other facts, even in the face of conflicting evidence, each argument is consistent with an analysis of Greek *ts* and *dz* as units that are internally complex, and thus each constitutes a piece in the on-going debate concerning the status of these sounds.

References

- Arvaniti, Amalia. 1987. *The Timing of Consonant Clusters: A Study of Initial Consonant Clusters in Modern Greek*. Cambridge: University of Cambridge M.Phil. dissertation.
- Arvaniti, Amalia. 1999. Illustrations of the IPA: Modern Greek. *Journal of the International Phonetic Association* 19: 167-172.
- Arvaniti, Amalia. 2007. Greek Phonetics: The State of the Art. *Journal of Greek Linguistics* 8: 97-208.
- Arvaniti, Amalia & Brian D. Joseph. 2000. Variation in voiced stop prenasalisation in Greek. *Glossologia* 11-12. 131-166. [Preliminary version (1999) in *Ohio State University Working Papers in Linguistics* 52: 203-232.]
- Arvaniti, Amalia & Brian D. Joseph. 2002. Modern Greek [b d g] in the early 20c.: Evidence from folk and rebetika songs. *Recherches en linguistique grecque*, vol. I: 67-70. Paris: L' Harmattan.
- Arvaniti, Amalia & Brian D. Joseph. 2006. Early Modern Greek /b d g/: Evidence from Rembetika and Folk Songs. *Journal of Modern Greek Studies* 22. 73-94.
- Campbell, Lyle. 1974. Phonological features: Problems and proposals. *Language* 50: 52-65.
- Fourakis, Marios. 2004. Temporal characteristics of [ps], [ts], and [ks] in Modern Greek. Paper presented at The First (Ever) Martin Luther King Day Symposium on Modern Greek Linguistics, 19 January 2004, Columbus, The Ohio State University.
- Fourakis, Marios, Antonis Botinis, & Elena Nigrianaki. 2002. Xronika xarakteristika ton simfonikon akoluθion [ps], [ts], ke [ks] stin eliniki. *Proceedings of the 6th International Conference on Greek Linguistics* [<http://www.philology.uoc.gr/conferences6thICGL/>].
- Hankamer, Jorge & Aditi Lahiri. 1986. The timing of underlying and derived geminates. Paper presented at the Annual Meeting of the Linguistic Society of America, New York, December 1986.
- Householder, Fred W. 1964. Three dreams of Modern Greek phonology. In Robert Austerlitz (ed.), *Papers in Memory of George C Papageotes* (issued as Supplement to *Word* 20.3), 17-27.
- Joseph, Brian D. & Irene Philippaki-Warburton. 1987. *Modern Greek* (Croom Helm Descriptive Grammar Series). London: Croom Helm.
- Joseph, Brian D. & Gina M. Lee. 1988. Greek *ts/dz* as internally complex segments: Phonological and phonetic evidence. Paper presented at the Annual Meeting of the Linguistic Society of America, New Orleans, December 1988.

- Joseph, Brian D. & Georgios Tserdanelis. 2006. On the phonetic description and IPA notation of affricates. Poster presented at the Annual Meeting of the Linguistic Society of America, Albuquerque, January 2006.
- Kingston, John & Mary Beckman (eds.) 1991. *Papers in Laboratory Phonology: Between the Grammar and Physics of Speech*. Cambridge: Cambridge University Press.
- Malikouti-Drachman, Angeliki. 2001. Greek phonology: A contemporary perspective. *Journal of Greek Linguistics* 2: 187-243.
- Miller, Ann. 1987. Phonetic characteristics of Levantine Arabic geminates with differing morpheme and syllable structures. In Mary Beckman & Gina Lee (eds.) *Papers from the Linguistics Laboratory 1985-1987 (Ohio State University Working Papers in Linguistics 36)*, 120-140.
- Newton, Brian. 1961. The rephonemicization of Modern Greek. *Lingua* 10: 275-284.
- Newton, Brian. 1972. *The generative interpretation of dialect. A study of Modern Greek phonology*. (Cambridge Studies in Linguistics 8). Cambridge: Cambridge University Press.
- Pantelides, X. 1923. Etimolojika ke fonitika. *Leksikografikon Arxion tis Mesis ke Neas Elinikis* 6: 116-124.
- Sagey, Elizabeth. 1986. *The representation of features and relations in nonlinear phonology*. Cambridge, MA: MIT Ph.D. dissertation.
- Setatos, Michalis. 1974. *Fonolojia tis kinis neolinikis*. Athens: Papazisis.
- Tserdanelis, Georgios. 2005. *The role of segmental sandhi in the parsing of speech: Evidence from Greek*. Columbus, OH: Ohio State University Ph.D. dissertation.

WORD-INITIAL CONSONANT CLUSTERS IN ALBANIAN

Rachel Klippenstein
The Ohio State University

Abstract

Albanian has a wide but not unrestricted range of initial consonant clusters. This paper lays out some constraints on such clusters; e.g., there are no clusters of two voiced stops, nor of voiced obstruent + voiceless obstruent. Dictionary data is supplemented by phonetic data from a native Albanian speaker, which helps determine how well orthographic evidence reflects pronunciation. I find that vowel epenthesis in obstruent-obstruent clusters is rare; schwa is sometimes elided to form clusters that are not orthographically evident, but less often than expected; and clusters written with voiceless obstruent + voiced obstruent are pronounced as such, at least sometimes.

1. Introduction

Albanian has a wide range of word-initial consonant clusters. Some of the more remarkable ones include *çd* [tʃd] in *çdo* ‘every’, *tk* in *tkurrje* ‘contraction’, *zhvl* [ʒvl] in *zhvleresoj* ‘devalue’; the longest contain four consonants, such as *zmb*r [zmbɾ] in *zmbrops* ‘repel’, and *çmpl* [tʃmpl] in *çmpleks* ‘untwine’. With such a rich variety of clusters, Albanian offers many interesting issues about syllabicity, the relationship between the phonology of morphologically simple and morphologically complex words, and the relationship between spelling and pronunciation. However, few people have yet addressed these challenges; work on the topic is essentially limited to lists of clusters (e.g., Buchholz & Fiedler 1987:46–50), and does not provide any substantial account. Before the interesting issues raised by Albanian clusters can be investigated in depth, a basic description and analysis of the facts is needed. In this paper I provide a step towards this goal by providing a description and basic analysis of two-consonant clusters (as well

as the first two consonants of longer clusters). I use dictionary-based evidence to provide an overview of constraints on consonant clusters, and supplement this with instrumental phonetic data to check how closely the spellings given in the dictionaries reflect the actual pronunciation of a native speaker. In addition to checking whether initial combinations of consonant letters are in fact pronounced as clusters, I also investigate the possibility that the elision of schwa in initial syllables results in consonant clusters that are not spelt as such. Finally, I investigate whether there is in fact a voicing mismatch in obstruent clusters spelt as *çb* and *çd*.

1.1 Background on Albanian

Albanian is an Indo-European language in its own branch of the Indo-European language family. It has two major dialect groupings: Geg, the northern variety, spoken primarily in northern Albania and Kosovo, and Tosk, the southern variety, spoken primarily in southern Albania, as well as in various diaspora communities in Italy, Greece, and elsewhere; there are many smaller dialect divisions within these two main groupings (Newmark et al. 1982:6–7, Friedman 2006:14). Standard Albanian is based primarily on Tosk, but has some Geg features (Newmark et al. 1982:8).

		labial	dental/alveolar ¹		palato-alveolar	palatal	velar	glottal
stop	voiceless	p	t			q /c/	k	
	voiced	b	d			gj /ʒ/	g	
fricative	voiceless	f	th /θ/	s	sh /ʃ/			h
	voiced	v	dh /ð/	z	zh /ʒ/			
affricate	voiceless		c /t͡s/		ç /t͡ʃ/			
	voiced		x /d͡z/		xh /d͡ʒ/			
nasal		m	n			nj /ɲ/		
tap			r /r̥/					
central approximant			rr /r̥/			j ²		
lateral approximant			l /l̥/	ll /l̥/				

Table 1. Obstruents and nasals (orthography = IPA unless otherwise indicated)

As background for discussing consonant clusters, an overview of the consonant inventory of Standard Albanian may be helpful. The obstruents, nasals, and sonorants are shown in Table 1 above; when orthographic representation differs from IPA, the IPA equivalent is indicated.³ The values given here are based on Buchholz & Fiedler (1987:37–42).

¹ Buchholz & Fiedler (1987) classify *th*, *dh* as apico-dental, *t*, *d*, *ll* as alveolar-dental (dental in Geg, alveolar in Tosk), and *s*, *z*, *c*, *x*, *n*, *l*, *r*, *rr* as alveolar. Newmark et al. (1982:9–10) describe *t*, *d*, *th*, *dh*, *n* as apico-dental, and *s*, *z*, *c* as apico-alveolar; *x* /d͡z/ appears to have a typo in its description and presumably is meant as apico-alveolar along with *z*. According to Friedman (2006:1), *t*, *d*, *c*, *x*, *ll*, *n*, *r*, *rr* are ‘alveolar, NOT dental, (except in some Geg)’.

² Buchholz & Fiedler (1987) consider [j] an allophone of /i/, not a consonant phoneme. I follow Newmark et al. (1982:10, 13–14) in considering it a consonant.

³ The vowel system of Standard Albanian consists of the vowels *i*, *y*, *e*, *a*, *o*, *u*, *ë*. With the exception of *ë*, the orthographic representation of vowels aligns reasonably closely with IPA; *ë* is a central unrounded vowel whose precise quality varies; it may be stressed (Newmark et al. 1982:11–12). I will refer to it as schwa and use the IPA symbol /ə/ when necessary. Albanian also has diphthongs *ie*, *ua*, *ye*, *ue* (Newmark et al. 1982:12).

1.2 Writing and speech

The relationship between writing and speech is a complicated one. Speech exists before writing both historically and in the life of a speaker. For good reason, therefore, linguists typically view written language as dependent upon spoken language, a reflection which is considerably distorted by constraints that the written medium imposes. In addition, once a written system is established, it tends to change slower than the associated spoken language, giving rise to additional mismatches between writing and speech. However, the relationship between writing and speech is not entirely unidirectional, with influence from speech flowing towards writing, and never the other way round. Writing also influences spoken language, as in the case of spelling pronunciations, where speakers come to pronounce a word based on its spelling rather than on its traditional pronunciation. The interaction between writing and speech means that there is a non-arbitrary relationship between them, and written language can help in understanding spoken language, so long as the limitations of the relationship are kept in mind and writing is not taken as a simple substitute for speech.

The present written standard for Albanian was developed fairly recently, in several stages over the course of the 20th century. At the Congress of Manastir in 1908, two ways of writing Albanian were established as acceptable: the present phonetically based Latin alphabet, and a system based on the Turkish/Arabic alphabet. The Literary Commission of Shkodër in 1916/1917 and the Educational Congress of Lushnjë in 1920 determined that the southern Geg dialect of Elbasan should be the basis for the standard, and this was taught at teachers' training school. However, this decision did not take hold, and people continued writing in both Geg and Tosk. Gradually, a shift occurred towards usage of Tosk with an admixture of Geg, encouraged in part by the use of Tosk in official documents. In the early 1950s, the Albanian Writers' Union and the National Conference on Orthography decided that the literary standard should be Tosk alone. The culmination of standardization efforts was the 1972 Congress of Albanian Orthography, which laid out the rules of orthography for a Tosk-based Standard Albanian, and led to the publication of official orthographic works. This Tosk-based standard continued (and continues) to have some Geg features, especially in lexicon and morphology (Newmark et al. 1982:6–9, Moosmüller and Granser 2006:122–123).

With this basis, it is reasonable to hypothesize that written standard Albanian reflects carefully spoken standard Albanian to a considerable degree, since it was intentionally designed to be phonologically grounded and the time since standardization is quite short, so that drastic phonological changes are unlikely to have caused the spoken language to change significantly while the writing system remained constant. Nevertheless, this is no guarantee that the written language is an accurate guide to the phonetics of the spoken language, and even where it reflects careful speech, casual or fast speech is likely to differ. Thus, it is necessary to confirm orthographically based hypotheses with phonetic data.

2. Standard Albanian consonant clusters from written sources

As mentioned above, Albanian has a wide range of clusters allowed word-initially, but possibilities are not unconstrained. In this section I work towards a full analysis of

Albanian word-initial clusters by examining the constraints governing which clusters do and do not occur, based on evidence from written sources. I limit my investigation to the initial two consonants of a word; primarily this includes two-consonant clusters, but I also include the initial two consonants of three-consonant clusters, where these do not seem to occur without a third consonant. I assume that in the occasional cases where two consonants are found as the initial two consonants of a three-consonant cluster, but do not occur as an initial two-consonant cluster, the lack of a two-consonant cluster is probably an accidental (i.e., nonsystematic) gap. I impose this limitation for two reasons: first, to keep the investigation to a manageable size, and second, because two-consonant sequences are the most basic level of complexity, which form the necessary background for any fuller account.

As far as I am aware, very little work has been done to investigate the constraints on Albanian consonant clusters. Buchholz & Fiedler (1987:47ff) provide a list of word-initial clusters, but give no analysis of the principles constraining them.

In this project, I assemble and analyze a list of consonant clusters based on Newmark (1998), supplemented by Stefanllari (1996). Newmark (1998) is a very thorough dictionary including many rare, obsolete, and dialectal words; non-standard words are easy to identify since they are marked with an asterisk. I restrict my analysis to clusters found in words not marked as non-standard, in order to avoid combining data from dialects that may have differing phonologies. Since Newmark (1998) contains many words that are probably not familiar to most native speakers,⁴ I compare the list of clusters derived from Newmark with those found in Stefanllari (1996), a much smaller dictionary, almost all of whose words are probably familiar to most Albanian speakers. Clusters found in Newmark (1998) but not in Stefanllari (1996) are italicized; clusters found in only one word in Newmark (1998) and none in Stefanllari (1996) I consider marginal, and put in italics and parentheses. It is conceivable that the competence of some Albanian speakers might not include these clusters; however, the fact that they occur in some words, even if only rare ones, suggests that they are not systematically excluded by Albanian grammar.

Newmark (1998) indicates by means of italics that certain instances of schwa (written as <ë> in Albanian orthography) may be elided. In many cases, Newmark indicates that a schwa between a word-initial consonant and another consonant may be elided, creating a word-initial cluster. Some of the clusters that would be formed by elision are also found in cases without elision (e.g., [ps] in *pse* ‘why’, *pësoj* ‘undergo, suffer’), while others are not (e.g., [kʃ] in *këshill* ‘council’); in addition, there are some clusters that are possible without elision, but appear to be impossible as results of elision (e.g., [ʃt] in *shtet* ‘state’ but not in *shëtit* ‘stroll around’⁵). I discuss clusters without elision and with elision separately.

2.1 Clusters without elision

⁴ I encountered many that were not familiar to the speaker I worked with.

⁵ Newmark (1998) indicates that the stress in *shëtit* is on the second syllable, so the failure of *ë* to elide is not attributable to it being stressed.

Clusters without elision from Newmark (1998) are shown in tables 2 and 3. The major constraints on consonant clusters on this chart are indicated by shading, as noted in the key to the table; these constraints are discussed in more detail below.

	p	t	q /c/	k	c /ts/	ç /tʃ/	f	th /θ/	s	sh /ʃ/	b	d	gj /j/	g	x /dz/	xh /dʒ/	v	dh /ð/	z	zh /ʒ/			
p	█								sp	shp													
t	<i>(pt)</i>	█					ft		st	sht													
q			█			çq	fq		sq	shq													
k		<i>tk</i>		█		çk			sk	shk													
c					█																		
ç						█																	
f							çf		sf	shf													
th								█		shth													
s	ps								█														
sh	psh									█													
b											█									zb	zhb		
d												█					vd			zd	zhd		
gj													█				vgj			zgj	zhgj		
g														█						zg	zhg		
x															█								
xh																█							
v										<i>(shv)</i>		<i>(dv)</i>									zv	zhv	
dh																						gdh	
z																							
zh																							
j	pj	tj		kj	çj	çj	fj	thj	sj		bj	dj									vj	dhj	zj
r	pr	tr		kr	çr	çr	fr	thr		shr?	br	dr		gr							vr	dhr	
rr	<i>pr</i>			<i>kr</i>									<i>gr</i>								<i>(vrr)</i>		
l	pl			kl		çl	fl		sl	shl	bl		gl								vl		
ll	pll			kl			fl		sl	<i>(shll)</i>	bl		gl								vl		<i>(zll)</i>
m		tm		km		çm			sm	shm													zm
n	<i>(pn)</i>					çn		<i>(thn)</i>	<i>(sn)</i>	shn				<i>gn</i>									
nj						çnj																	
h				<i>(kh)</i>		<i>(çh)</i>																	

Table 2. Clusters where C₁ is an obstruent

Notes to Table 2:

- Underlying clusters from Newmark 1998 not marked as nonstandard.
- *Italicized* clusters not in Stefanllari 1996.
- (*Italicized and parenthesized*) clusters in only one word in Newmark 1998 and no words in Stefanllari 1996, counted as marginal; see Appendix 1 for list of marginal words.
- Only with a morpheme boundary between the consonants: *çb, çf, çh, çn, çnj, çq, çr, çrr, shth, zhb, zhd, mv*.

Key to shading:

No obstruent clusters with voicing mismatch (exceptions: <i>çb, çd</i> , marginal <i>shv</i>); No clusters of two voiced stops	C1 not <i>x, xh, q, gj</i> ; C1 not oral sonorant (exception: <i>rrj</i>); no clusters of <i>th, dh</i> + obstruent	
No geminates	C2 not <i>c, ç, xh</i> ; no clusters of obstruent + <i>x</i>	

	j	r [r]	rr [r]	l	ll [ɫ]	m	n [n,n̥]	h
p						mp		
t								
q								
k							nk	
c								
ç								
f								
th								
s								
sh								
b						mb		
d							nd	
gj							ngj	
g							ng	
x							nx	
xh								
v						mv		
dh								
z								
zh								
j	■		rrj			mj		
r		■				mr		(hr)
rr			■			mrr		
l				■		ml		
ll					■	ml		
m						■		
n								
nj								
h								■

Table 3. Clusters where C_1 is a sonorant

2.2 Clusters without elision: overall patterns

Geminate consonants do not occur word-initially in Albanian. (In fact, Albanian does not have geminates word-internally either; orthographic *ll* and *rr* represent /ʎ/ and /r̥/ respectively.)

There are several consonants that do not begin clusters in Standard Albanian as represented in Newmark (1998): the voiced affricates *x*, *xh* /d͡z/, d͡ʒ/ and the palatal stops *q*, *gj* /c, ɟ/. There are fairly plain historical explanations for at least some of these. The palatal stops have several sources. They developed by palatalization of pre-Albanian **k*, **g* before **j* and front vowels, and from Proto-Albanian **kl*, **gj*.⁶ Additionally, *gj* developed from Proto-Indo-European **s* before a stressed vowel, and from pre-Albanian **j* (Demiraj 1996:196–200; Beekes 1995:261–263). Since *q/gj* always developed before a vowel (or before a glide which was then lost), it was never in a position to be followed by another consonant.

⁶ The development of *q*, *gj* from **kl*, *gj*, is relatively recent, since some outlying dialects still have clusters: Standard Albanian *gjuhe* ‘language’ = *gluhë* in an Arvanitika dialect in Greece (Demiraj 1996:198).

The palato-alveolar affricate *xh* [d͡ʒ] has a special status in Albanian. It occurs primarily in loanwords from Turkish, as well as in some loanwords from other languages (including English), and some sound-symbolic forms (Curtis 2008). This sound would not begin a cluster in loanwords from Turkish, as the source sound in Turkish could not be the first member of an initial cluster, since clustering in Turkish is very limited. This does not fully explain why it does not begin clusters in sound-symbolic words. However, if *xh* at one time existed only in loanwords from Turkish, and was afterwards employed for sound-symbolic use, then it would first have become established in the language as a segment that occurred word-initially before vowels but not before other consonants; this may have become a phonotactic constraint that sound-symbolic forms adhered to when they were introduced.

The historical reasons that these consonants cannot begin clusters do not mean that there are no synchronic phonological reasons. The historical developments gave rise to a state of the language in which initial clusters beginning with these consonants do not appear. When a language learner is presented with the data of Albanian, there is nothing to encourage the learner to posit the possibility of such clusters. They are not sporadic gaps in an otherwise full range of possible clusters, but a consistently absent category. Given this consistent absence, it is plausible that speakers would exclude it from their grammars. The hypothesis in the preceding paragraph about the reason for the absence of *xh*-initial clusters in sound-symbolic forms is an example of how this could apply.

In addition to consonants that cannot begin clusters, there are also consonants that cannot end clusters. Specifically, there are no clusters ending with three of the four affricates: *ç*, *c*, *xh* [t͡ʃ], [ts], [d͡ʒ] (the fourth affricate, *x* [dz] is found after *n*, as mentioned later).

2.3 Obstruent-obstruent clusters

In clusters of two obstruents, voicing mismatches are avoided. There are no cases of a voiced obstruent followed by a voiceless obstruent, and clusters of a voiceless obstruent followed by a voiced obstruent are very restricted. Newmark (1998) gives the clusters *çb* [t͡ʃb] (e.g., *çbind* ‘dissuade’), *çd* [t͡ʃd] (e.g., *çdo* ‘every’), and marginally, *shv* [ʃv], which occurs in the word *shvenk* ‘flash-pan’ (a cinematographic term), which appears to be a borrowing from German *Schwenk* ‘pan’ (also in a film context; Messinger et al. 1993). This is the first case where *ç* [t͡ʃ] seems to combine more freely than other consonants. Newmark et al. (1982:19) indicate that this may be a place where orthography is misleading; they state that the negative prefix *ç* becomes *zh* before a voiced consonant (e.g., *ç-* + *duk* = *zhduk* ‘cause to disappear’). However, in the examples they give, the assimilated form is spelt with a *zh-*, leaving a mystery of whether the few forms spelt with *çd-* and *çb-* are also assimilated or not.⁷ This question will be investigated further in the phonetic study below.

Clusters of two stops do occur, but they are very restricted. The only such clusters consist of two voiceless stops; the cluster *tk* occurs in *tkurrje* ‘contraction’ and a few

⁷ While *çboj* and *çbind* have the negative prefix in question, *çdo* does not, but rather a morpheme meaning ‘what’; this may not follow the same morphophonemic rules as the negative prefix.

other words from the same root, and *pt* occurs marginally in the interjection *ptu* ‘stylized spitting to represent spite/contempt for someone’. In addition to the consonants discussed earlier, obstruent-obstruent clusters cannot begin with the dental fricatives *th*, *dh* /θ, ð/, and cannot end with *x* /d͡z/.

Some of the clusters described in this section, particularly the clusters with voicing mismatch and those with two stops, raise questions about underlying and surface forms, and the relationship of orthography to these. Assuming that these spellings represent some level of phonological reality, which level do they represent? Do they represent an underlying level, or something close to it, with phonological processes affecting the cluster so that it surfaces in a different form, such as [d͡ʒb] or [kət]? Or do they represent a surface level, aligning well with the phonetic realization of these clusters? Does this have any relation to broader tendencies in the relationship between writing and speech? The question of the phonetic realization of these clusters will be taken up below in section 3.2.

2.4 Obstruent-sonorant clusters

There are fewer clearly definable restrictions on obstruent-sonorant clusters than on obstruent-obstruent or sonorant-sonorant clusters: most sonorants (especially *j*, *r* /r/, *l*, *ll* /l/, *m*) cluster quite freely after a wide range of obstruents. Three of the four liquids, *rr* /r/, *l*, *ll* /l/,⁸ do not follow most coronals, but there are exceptions: the clusters *çrr*, *çl*, *sll*, *shl*, *shll*, *zll* [t͡ʃr, t͡ʃl, sɫ, ʃɫ, zɫ] do occur.

The palatal nasal *nj* [ɲ] occurs only after *ç* [t͡ʃ] (e.g., *çnjerëzor* ‘inhuman’) — once again, *ç* patterns more freely than other consonants. In all cases of *çnj*, the *ç* is a separate morpheme, a negative prefix. The ability of *ç* to cluster more freely than most consonants, (and especially than other affricates), combined with the fact that in these cases the *ç* is a prefix, raises questions about how the phonology of morphologically complex words relates to that of morphologically simple words.

2.5 *h* in clusters

Generally, *h* does not occur in clusters either as the first or second member. However, Newmark et al. list a few marginal cases. As a first member it occurs marginally in the cluster *hr*, in the clearly borrowed word *hrushovian* ‘Khrushchevian’.⁹ As a second member it occurs marginally in the cluster *kh* the onomatopoeic word *khu-khu* ‘coughing sound of someone choking’¹⁰ and in the linguistic term *çhundorëzim* ‘denasalization’ (in which the *ç* is a negative prefix.)

⁸ The fourth liquid is *r* [r].

⁹ The speaker I worked with did not know this word; it would be interesting to find out how it is pronounced by speakers who do use this word—is it [hr], [xr], [ɣ], or something else?

¹⁰ The speaker I worked with knew this word but said the initial consonant was simply a *k*, and would spell it accordingly. More investigation would be needed to find out whether other Albanian speakers pronounce it with an onset other than [k], whether a cluster or fricative [x] or something else.

2.6 Sonorant-initial clusters

The only cluster beginning with an oral sonorant in Standard Albanian¹¹ is *rrj* [rj] (e.g., *rrjedhim* ‘result’).¹² There are clusters of *m* followed by all of the oral sonorants. The other nasals, *n* and *nj* [ɲ], however, are not followed by oral sonorants. There are no clusters of two nasals in Standard Albanian¹³. There are also clusters of nasal + obstruent, including voiceless obstruents. These include *mv* (the only cluster of nasal+fricative) in e.g. *mvehtësi* ‘independence, individuality’, and (what appear to be) clusters of a nasal followed by a homorganic stop or affricate, including some cases with voiceless stops, such as *mp* in *mposht* ‘defeat’. These nasal-stop clusters can occur as part of larger clusters, such as *zmb* [zmbɾ] in *zmbrops* ‘repel’, and *çmpl* [tʃmpl] in *çmpleks* ‘untwine’. Such clusters raise questions about syllabicity and the role of sonority in syllabification: are clusters such as *mp* and *mv* onsets, or do they involve a syllabic nasal? If such sonorant-obstruent clusters are onsets, what about obstruent-sonorant-obstruent-sonorant clusters such as *çmpl*; are they also onsets? If so, what does this say about the role of sonority in syllabification? In addition to questions about the syllabification of clusters such as *mb* and *ng*, there is also a question whether they are clusters at all, or whether they are in fact prenasalized stops. In the next section, I discuss some facts related to these possibilities, though I come to no firm conclusion.

2.7 #NC

There are three main possibilities for the phonological identity of what appears to be a word-initial nasal-stop cluster (NC). These possibilities are phonetically quite similar, though there may be subtle differences, but phonologically they are distinct, and are tied to differences in whether the NC is tautosyllabic or heterosyllabic, and whether it is monosegmental or bisegmental. These possibilities are outlined in (1).

(1)	Syllabic nasal+stop	bisegmental	heterosyllabic /#n.d/ ¹⁴
	Onset cluster	bisegmental	tautosyllabic /#nd/
	Prenasalized stop	monosegmental	tautosyllabic /#nd/

There are two ways to try to determine which of the possibilities in (1) occurs in a given case: phonetic and phonological. While the different possibilities are phonetically similar, there may be subtle differences which make it possible to distinguish between them phonetically; such a method for distinguishing clusters from prenasalized stops will be discussed below. A phonological method, instead, attempts to determine how an NC fits in to the phonological system of the language. For instance, a language might have suffix

¹¹ Other clusters beginning with oral sonorants occur in other dialects which are outside the scope of this paper; Geg, for instance, has clusters of *rrn-*, as in the title of the book *Rrno vetëm për me tregue* by Zef Pllumi (1995).

¹² It is worth noting that all instances of *rrj* were followed by the vowel *e*.

¹³ Tosk dialects spoken in Greece, known as Arvanitika, show a development of *mj-* into *mnj-* in e.g. *mnjekrë* ‘beard, chin’, under the influence of Modern Greek (Brian Joseph, p.c.)

¹⁴ A variant on this possibility is that the two consonants are phonetically preceded by or broken up by a vowel: [#ən.d] or [#nə.d]. Phonetic data would reveal if there is a vowel present; my phonetic study found no such evidence.

with a disyllabic allomorph that attaches monosyllabic roots and a monosyllabic allomorph that attaches to words of two or more syllables; in such a language, if roots like *ndal-* take the monosyllabic allomorph, they must be disyllabic, and the *n* must therefore be syllabic.

Riehl (2008) both details a phonological method for distinguishing the different types of NCs (not only in initial position) and finds a phonetic distinction between prenasalized stops and clusters. In Riehl's phonological method, NCs which are heterosyllabic are necessarily clusters, while NCs that are not 'separable' — whose stop component does not occur as an independent segment outside NCs — are necessarily prenasalized stops. Tautosyllabic separable NCs are considered clusters by default, and are identified as prenasalized stops only if other phonological evidence points that way — that is, 'if NC sequences appear to be treated as single segments by the phonology, in contrast to clear consonant clusters' (Riehl 2008:24). Riehl (2008:52–62) also argues that there are no cases of prenasalized voiceless stops, and that the occasional reports of them have been confounded by other factors, e.g., failing to distinguish between tautosyllabicity and monosegmentality, or analyzing phonetically voiced prenasals as voiceless to economize on the language's feature inventory.

Phonetically, Riehl (2008) finds that prenasalized voiced stops are roughly equivalent in duration to plain nasals, while NC clusters (even when tautosyllabic) are substantially longer. (This applies in particular to NCs where the C is a voiced stop; NCs where the C is an affricate or voiceless stop are substantially longer than plain nasals even when they are monosegmental (Riehl 2008:272–275).)

Albanian NCs are clearly separable: their stop/affricate components occur as independent segments; thus they are not obvious cases of prenasalized stops. According to Buchholz & Fiedler (1987:43–44), Albanian NCs are tautosyllabic, both word-initially and word-internally. For the speaker I was working with, this was not intuitively clear. The speaker found it hard to decide whether words like *mposht* 'defeat' (with initial [mp-]) and *ngjall* 'revive' (with initial [ŋʝ-]) were one syllable or two. When asked how many notes she would sing them on, she replied with no hesitation that she would sing each word on one note.¹⁵ However, intuition may be a misleading guide to syllabification (Riehl 2008:21). Ideally there would be not only intuitions, but more directly phonological evidence indicating whether these clusters are tautosyllabic or heterosyllabic, and whether the nasal forms a syllable of its own or not. I will tentatively follow Buchholz & Fiedler and my speaker's singing syllabification and assume that Albanian NCs are tautosyllabic. This leaves them in the 'inseparable tautosyllabic' category, where by default they are to be considered clusters, though they may be considered prenasalized stops if further evidence warrants. Further investigation would be needed to determine whether phonological evidence exists. Phonetic evidence about

¹⁵ The speaker's syllabification judgments were sometimes problematic in other places; e.g., when first asked how many syllables there were in *mbiemri* 'adjective, surname', she said that there were two: *mbi* and *emri*. In these cases, asking her how many notes she would sing them on gave a more expected syllabification; however, it is possible that this reflects conventional singing patterns rather than directly representing phonology. (On a later occasion, asked again how many syllables there were in *mbiemri*, she responded that there were 3: *mbi*, *em*, and *ri*.)

the duration of NCs relative to nasals would also be informative, but is not incorporated into the present study.

2.8 Clusters with elision

	p	t	q	k	c	ç	f	th	s	sh	b	d	gj	g	x	xh	v	dh	z	zh
p	█		█					█												
t		█		kət			fət	█												
q			█																	
k				█																
c	█		█	█	█															
ç	█		█	█	█	█														
f							█													
th								█						gəth						
s	pəs			kəs					█											
sh	pəsh			kəsh						█										
b											█									
d												█								
gj													█							
g														█						
x	█			█											█					
xh	█			█												█				
v																	█			
dh																		█		
z																			█	
zh														gəzh						█
j																			věj	
r																			vër	
rr	përr			kërr							bërr	dërr	gërr					vërr		
l		təl		kəl																
ll	pəll			kəll		fəll		shəll			dəl								vəll	
m				kəm		fəm														
n				kən																
nj																				
h																				

Table 4. Clusters where C_1 is an obstruent

- Clusters formed by elision from Newmark 1998 not marked as nonstandard

Key to shading (same as table 2, except clusters starting with oral sonorants not shaded):

No obstruent clusters with voicing mismatch (exceptions: <i>çb</i> , <i>çd</i> , marginal <i>shv</i>); No clusters of two voiced stops	█	C_1 not <i>x</i> , <i>xh</i> , <i>q</i> , <i>gj</i> ; no clusters of <i>th</i> , <i>dh</i> + obstruent	█
No geminates	█	C_2 not <i>c</i> , <i>ç</i> , <i>xh</i> ; no clusters of obstruent + <i>x</i>	█

	j	r	rr	l	ll	m	n	nj	h
p						mëp			
t						mët			
q						mëq			
k				lëk		mëk			
c									
ç									
f									
th									
s						mës			
sh						mësh			
b			rrëb	lëb					
d						mëd			
gj						mëgj			
g						mëg			
x									
xh									
v						mëv			
dh				lëv		mëdh			
z						mëz			
zh						mëzh			
j									
r						mër			
rr						mërr			
l						mël			
ll						mëll			
m				lëm					
n				lën		mën			
nj						mënj			
h						mëh			

Table 5. Clusters where C_1 is a sonorant

In addition to words beginning with orthographic consonant clusters, there are words where consonant clusters may be formed if orthographic schwa (*ë*) is not pronounced. Newmark (1998) italicizes *ë* in some words to indicate that it may be unpronounced. In some words, a schwa is italicized between a word-initial consonant and another consonant. In these cases, if the vowel is omitted, a consonant cluster would result. For example, Newmark (1998) gives *mësim* ‘education’; if the *ë* is omitted, the result would be an initial [ms] cluster. In the following discussion, I parenthesize cases of *ë* that Newmark italicizes. Table 4 shows all the initial clusters formed by elision according to Newmark (1998). This differs from Table 2 in the following significant ways. First, there are a wider range of clusters beginning with oral sonorants—mostly *l*, e.g. in *l(ë)vere* ‘rag’, *l(ë)mosh(ë)* ‘alms’, *l(ë)bardh* ‘give a white appearance’, but also *rr* in *rr(ë)byth* ‘force backwards’. (Consonant clusters still do not begin with *j*, *r*, and *ll*.) There are no clusters beginning with any of the nasals except *m*, but *m* can begin a cluster followed by almost any consonant, with the exceptions of the affricates (*c*, *ç*, *x*, *xh* /*ts*, /*tʃ*, /*dʒ*, /*dʒ*/), which do not participate in any clusters formed by elision, and *f*, *th*, *b*, *j*. The consonants *s*, *z* and *zh* do not begin clusters with elision in Newmark, and there are almost no clusters with two voiced obstruents (the only exception is *g(ë)zh-* in e.g. *g(ë)zhøj(ë)* ‘shell’).

Schwa elision raises questions about the relevance for phonological description and analysis of formal or careful speech pronunciations vs. informal or fast speech pronunciations; in some sense, careful speech forms seem more basic and fundamental, with fast speech forms being secondary and derived from them. Yet, it seems probable that people hear far more fast and informal speech than careful and formal speech, and so people must often learn words from informal speech rather than from formal speech, making informal speech in a sense more basic at least for certain types of things. What role does each of these types of speech play in a speaker's knowledge of their language, and thus how should it be taken into account in phonological analysis?

3. Phonetic study

3.1 Speaker

The study involved a single speaker who was female, in her early 30s, and came from Prishtina in Kosovo. She spoke both Geg and Geg-influenced Standard Albanian; she reported that her parents spoke to her in Standard Albanian and Standard Albanian was the language of her schooling. This study investigated only her Standard Albanian. Geg influence was evident in the fact that she did not distinguish *r* and *rr* (standardly [r] and [r̥]), and merged the palatal stops *q* and *gj* (standardly [c] and [ɟ]) with the palato-alveolar affricates *ç* and *xh* ([tʃ] and [dʒ]).

3.2 Stimuli and recording process

Based on dictionary work and elicitation sessions, 84 words were selected to be recorded. The words discussed here included 17 words with orthographic obstruent-obstruent clusters, and 20 words with *ë* between the first two consonants (which might be elided); 15 words with initial NCs (e.g., *ndal*), 15 with initial stops (e.g., *dal*), and 11 with initial nasals (e.g., *nam*) were recorded in order to investigate NCs, but results from these are not discussed here, as I decided after recording that the stimuli and recording conditions were not sufficiently controlled to reliably measure timing. 6 words with other clusters were recorded but not used in analysis. The obstruent-obstruent clusters tested in this study were selected as follows. For a complete list of words, see Appendix 2; for a list of words recorded but not analyzed, see Appendix 3.

- (2) *çb, çd, çk, çf* (x2) ([tʃb, tʃd, tʃk, tʃf])
 These clusters represent all the available clusters of an affricate followed by a consonant. Newmark has examples with *çf* [tʃf] and *çq*, [tʃc] but the speaker in this study was not familiar with these words.
- (3) *shp, sht, shk, shf* ([ʃp, ʃt, ʃk, ʃf])
 These clusters are equivalents of those in (2) with a voiceless fricative in place of the affricate, and voiceless consonants where those in (2) had voiced consonants.
- (4) *zhb, zhd, zhg, zhv* ([ʒb, ʒd, ʒg, ʒv])
 These clusters are the voiced equivalents of those in (3); thus, in relation to the clusters in (2) they have voiced fricatives in place of the affricates, and voiced second members where those in (2) had voiceless ones.

- (5) *vd, tk, kf, gdh* ([vd, tk, kf, gð])
 These clusters were chosen for a variety of reasons: *tk* is the only stop-stop cluster familiar to my speaker, and I reasoned that this was one of the most likely contexts for epenthesis;¹⁶ *vd* and *gdh* were the only all-voiced obstruent clusters familiar to my speaker that did not begin with *z* or *zh*;¹⁷ *kf* does not occur in Newmark (1998), but was found by elicitation in *KFOR-i* ‘Kosovo Forces’.

Obstruent-obstruent clusters that were not tested include clusters beginning with *s* (*sp, st, sq, sk, sf*), *z* (*zb, zd, zgj, zg, zv*), *f* (*ft, fq, fsh*), *k* (*kth, ks*) and *p* (*ps, psh*).

The 20 words chosen to test the possibility of elision included some in which Newmark indicated that *ë* could be elided, and some in which he did not; they also included some in which pre-recording elicitation with the speaker indicated that she thought she would pronounce it (at least sometimes) with an initial consonant cluster, and some in which she did not. The two classifications did not line up: there were words where Newmark indicated elision and the speaker did not, and vice versa. For a full list of words and details about whether Newmark and the speaker indicated elision for each word, see Table 6 below.

The speaker was asked to say these words in frames designed to elicit various rates of speech, as given in (6) and (7), where _____ represents the slot into which the speaker was to insert the word. (Note the presence of an isolated instance of the target word before the instances in sentence contexts.)

- (6) ‘them’ frame
 Albanian

 Tani do të them _____ përsëri.
 Tani do të them _____ shpejt.
 Tani do të them _____ ngadalë.
 Translation
 _____ (‘isolated repetition’)
 Now I’ll say _____ again. (‘again repetition’)
 Now I’ll say _____ quickly. (‘fast repetition’)
 Now I’ll say _____ slowly. (‘slow repetition’)
- (7) ‘përsërisë’ frame¹⁸
 Albanian

 Tani do të përsërisë _____ përsëri.
 Tani do të përsërisë _____ shpejt.
 Tani do të përsërisë _____ ngadalë.

¹⁶ Newmark (1998) gives an interjection with *pt*; to my speaker this cluster did not seem like a possible word beginning.

¹⁷ Newmark (1998) gives words with *dv* and *vgj*, but these clusters were not familiar to my speaker.

¹⁸ The final *ë* in *përsërisë* was rarely pronounced.

Translation		
_____.		(‘isolated repetition’)
Now I’ll repeat _____ again.		(‘again repetition’)
Now I’ll repeat _____ quickly.		(‘fast repetition’)
Now I’ll repeat _____ slowly.		(‘slow repetition’)

The frames and the target words were presented on separate pieces of paper. The target words were organized in 8 groups of 10 or 11 words each, with two groups on each piece of paper. Each group of words was recorded into a separate file, using the computer program Praat (Boersma 2008). Before recording, the speaker was instructed to do what she said she would do when saying the words in the frames. This was effective in eliciting different speech rates.

3.3 Results: Do obstruent-obstruent sequences have epenthetic vowels?

The first question investigated by this study was whether epenthetic vowels break up sequences of two orthographic obstruents. I found that this occurs very rarely, and did not find any contexts where it happened consistently.

As discussed above, 17 words with orthographic obstruent-obstruent sequences were recorded. In investigating this question, I have only analyzed those in the ‘them’ frame, due to the difficulty in distinguishing between the final *s* in *përsërisë* and an initial fricative in the target word; thus, 4 repetitions of each word are analyzed here, giving a total of 68 tokens. Among these were two tokens with clear epenthetic vowels, listed in (8).

- (8) Tokens with clear epenthetic vowels
 One token of [tʃəb] in *çboj* ‘I undo’¹⁹ in isolation
 One token of [zəv] in *zhvillim* ‘development’ in isolation

Phonetically, there is no clear-cut division between a long release and a brief vowel. A short release may be clearly a release, and a non-brief vowel may be clearly a vowel, but the area in between is gradient and not categorical. Among the words recorded, I found two tokens with a release that comes close to being a vowel, listed in (9).

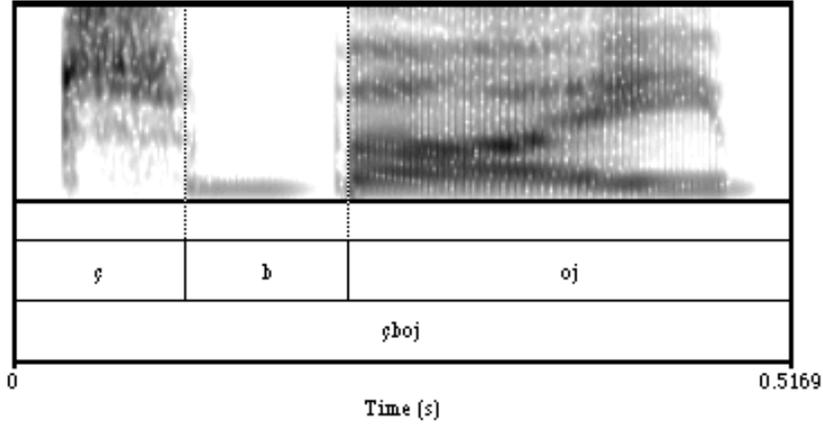
- (9) Tokens with near-vowel releases
çboj in *Tani do të them ‘çboj’ përsëri* ‘now I’ll say *çboj* again’
vdekja ‘the death’ in *Tani do të them ‘vdekja’ ngadalë* ‘now I’ll say *vdekja* slowly’²⁰

¹⁹ For this word, Newmark (1998) gives standard *zhbën* (3sg; 1sg = *zhbëj*) and nonstandard *çbën* (3sg; 1sg = *çbëj*); the speaker’s *çboj* (1sg; 2sg *çbon*) is not listed, but I have used it as an example of the cluster *çb* which does occur in other standard words in Newmark, particularly *çbojatis* ‘discolor’ and *çbind* ‘dissuade’.

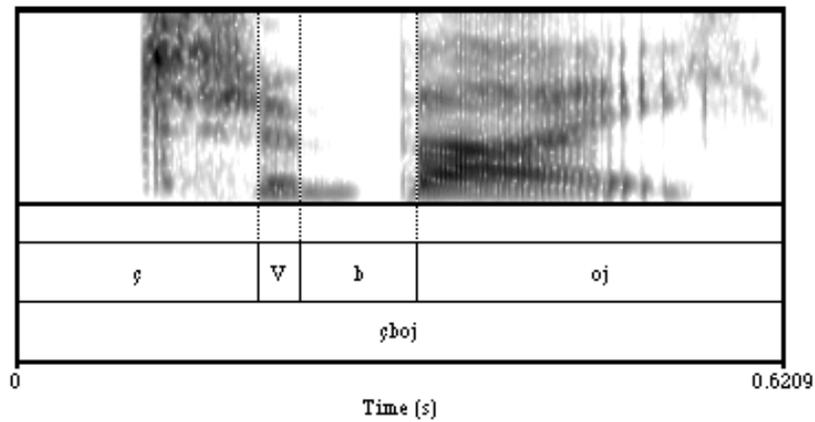
²⁰ Kelly Maynard has called to my attention that the near-vowel release in *vdekja* may be due to the fact that the speaker is from Kosovo, since in Kosovo dialects, this word begins with plain *d*, so *vd* may be an unfamiliar cluster.

Spectrograms of *çboj* said without epenthesis, with an epenthetic vowel, and with a release that is almost a vowel are given in 10–12.

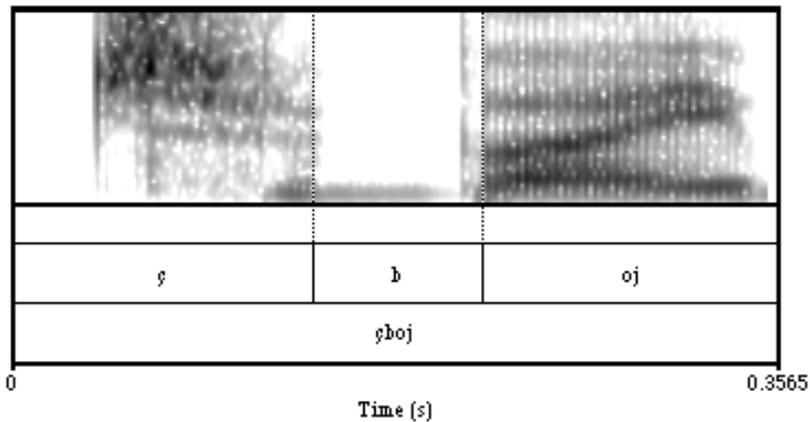
- (10) *çboj* with voiceless *ç* [t̪], voiced *b*, and no epenthetic vowel:



- (11) *çboj* with an epenthetic vowel:



- (12) *çboj* with a release that is almost a vowel:



Although there were a few cases of epenthesis in words said in isolation, it is absent in the vast majority of cases, even in isolation. Although not all types of obstruent-

obstruent clusters were tested, the results from those that were tested suggest that epenthesis in obstruent-obstruent clusters is uncommon, and that the occasional instances are more likely to happen in slow or careful speech than in casual or fast speech.

3.4 Results: Can the existence of clusters formed by elision be confirmed acoustically?

The second question investigated in this study is whether it is possible to confirm acoustically that elision of schwa produces clusters which are not indicated orthographically.²¹ The data showed that this does take place, but less than anticipated.

Pre-recording elicitation had determined a number of words with *ë* in the first syllable, some of which the speaker said she would say with an initial cluster (without the *ë*) in spoken Albanian, and some of which she indicated that she would not omit the vowel in.²² These data did not align very well with what was found in recordings. Clusters were not found in any cases where the speaker said she would not have them, and were also only occasionally found in cases where she said she would have them in spoken Albanian. All cases of elision occurred in fast repetitions. 20 words were investigated for elision. Of the 40 tokens in fast repetitions, one was disfluent, leaving 39 fluent tokens, 22 of which belonged to words that the speaker indicated elision in, and a further 11 of which Newmark indicated elision in, while 6 tokens belonged to words that neither Newmark nor the speaker indicated elision in.

Elision occurred once in the frame *tani do të them ___ shpejt* (with the word *dëllirë* ‘pure’) and 3 times in the frame *tani do të përsërisë ___ shpejt* (with the words *mësuesja* ‘the teacher’, *bërryl* ‘elbow’, and *mësim* ‘education’). A spectrogram of *mësim* with elision is given below in (13). There was also one case where not only the vowel, but also the initial consonant was omitted in the word *kështjella* ‘the castle’. Thus, out of the 22 fast speech tokens of words where the speaker indicated that she would elide, less than one-fourth showed elision. In addition, there were 3 tokens where *ë* was very reduced but not entirely gone. A full chart of cases with and without elision is given in Table 6; this table also indicates whether the speaker said that she would elide the vowel, and whether Newmark (1998) indicates that elision is possible.

²¹ Strictly speaking, this study does not prove that these cases are elision, since it does not prove that schwa is underlyingly present phonologically. However, the fact (discussed below) that it is usually present strongly suggests that it is phonologically present.

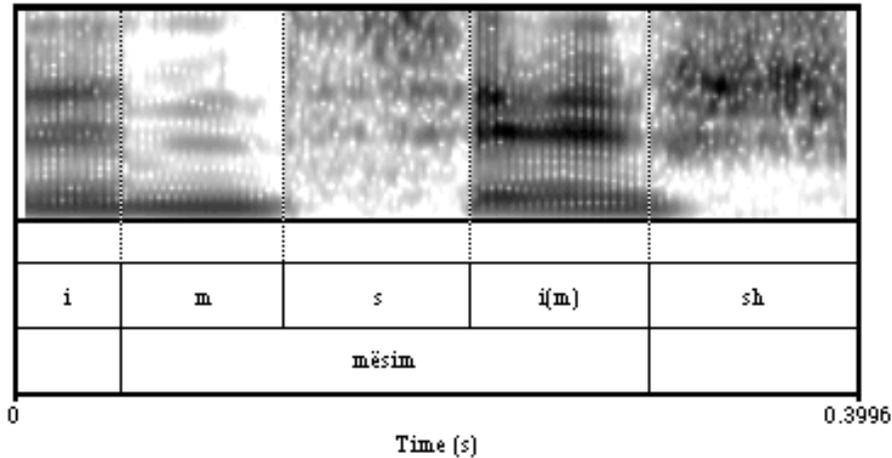
²² The speaker was asked questions like ‘do you know any words that start with *kt*?’, and often volunteered words that she said had the particular cluster in question in spoken Albanian, but were written with an intervening *ë* in written Albanian. On occasion she was asked particularly how she would pronounce a given word with *ë* in the first (orthographic) syllable, and would say it and indicate whether she would omit the vowel or not.

Albanian	gloss	Tani do të them ____ shpejt	Tani do të përsërisë ____ shpejt	Speaker said would elide?	Newmark shows elision?
kësaj	this (infl.)	x	x	(not asked)	yes
këmishë	shirt	x	x	would	yes
mëshova	I did put pressure on	x	x	would	yes
kënetë	marsh	x	x	wouldn't	yes
këtu	here	x	x	would	yes
mëtejshëm	later	x	x	(not asked)	yes
gëzim	joy	x	x	would	no
dënim	punishment	x	x	wouldn't	no
gjëzim	thundering sound	x	x	wouldn't	no
lëkundje	shaking	x	x	?	no
gëzuar	joyous	x	x	would	no
kështjella	the castle	x	ELISION; k also elided	would	no
kësulkuqe	Little Red Riding Hood	x	x	would	no
lëmoshë	alms	x	x	?	yes
mësuesja	the teacher	x	ELISION	would	yes
gëzhojë	shell	x	(very reduced)	would	yes
bërryl	elbow	x	ELISION	(not asked)	yes
mësim	education	(very reduced)	ELISION	would	yes
mëditje	daily wages	(very reduced)	(disfluency; no data)	wouldn't	yes
dëllirë	pure	ELISION	x	would	yes

Table 6. Elision data: Elision in fast speech; words listed in the order they were recorded. Note increasing elision towards the end of the recording session.

The fact that there were more elisions in the *përsërisë* frame than in the *them* frame is probably due to several factors that likely increased speech rate. First, since *përsërisë* is longer than *them*, the whole frame was longer, which could contribute to its components being pronounced faster. Second, the *përsërisë* repetitions followed the *shpejt* repetitions, and the speaker's rate of speech increased over the course of the task.

While the phonetic data provide evidence that *ë* may sometimes be elided, they do not provide evidence that it is usually elided in some words. This could be due to one of two things. First, it may be that (at least for this speaker), *ë* is not in fact elided in these words, despite speaker intuitions to the contrary. Second, it may be that the task did not elicit the speech registers in which *ë* is elided. For example, since the task depended on writing, it may be that even at relatively fast speech rates, an orthographically-based pronunciation was elicited, rather than a pronunciation that would be used in everyday speech.

(13) *mësim* with elision of schwa

It is worth drawing attention to the fact that while epenthesis sometimes occurred in orthographic clusters, and a vowel was sometimes elided in words with *ë* between the first two consonants, the cases with orthographic clusters usually did not show epenthesis, and the cases where *ë* was in a position to be elided usually did not show epenthesis. This gives evidence that the orthographic presence or absence of *ë* between the first two consonants does in fact reflect a phonological reality.

3.5 Results: Are *çb*, *çd* phonetically voiceless followed by voiced [tʃb, tʃd] as spelling suggests?

The third question investigated in this study is whether Albanian does in fact have clusters of a voiceless obstruent followed by a voiced obstruent—specifically, are *çb* and *çd* phonetically [tʃb, tʃd] as the spelling suggests, or not? The phonetic evidence showed that they can be phonetically voiceless + voiced, and are in several tokens, but that they are not always.

The initial *ç* is generally pronounced voicelessly, as [tʃ]; however, in the fast repetition of both *çboj* and *çdo* after *përsërisë*, some assimilation occurs: the final *s* of *përsërisë* is pronounced as voiced [z], and the initial *ç* of the following word is pronounced as [ʒ], producing a cluster across the word boundary of [z#ʒ].²³

The second consonant in the cluster, *b* or *d* was sometimes voiced, but was voiceless in many tokens. It is worth noting that this was true also of the voiced stops in the ‘all-voiced’ clusters *zhb*, *zhd*, and *zhg*. A spectrogram of *çboj* with the initial cluster pronounced [tʃb] has already been given in 10.

4. Conclusion

The rich, largely unexplored clustering possibilities of Albanian consonants provide many interesting questions and avenues for investigation. In this paper I have done

²³ The final *ë* in *përsërisë* was elided in these cases, as in most cases at all speech rates.

necessary groundwork for the investigation of Albanian word-initial consonant clusters. I have not only provided a list of clusters, but have determined some of the constraints on clusters not formed by elision, which I summarize as follows:

- Clusters of voiced obstruent followed by voiceless obstruent are not permitted.
- Clusters may not begin with palatal stops (*q, gj*) or voiced affricates (*x, xh*).
- Affricates (*c, ç, x, xh*) may not be second members, except that *nx* is allowed.
- Obstruent-obstruent clusters may not begin with dental fricatives (*th, dh*).
- The only voiceless obstruent that may precede a voiced obstruent is *ç*.
- The only consonant that may precede *nj* is *ç*.
- *rr, l, ll* do not follow most coronals, but there are exceptions.
- *h* does not normally participate in clusters.

Based on phonetic data, I have determined several things. First, obstruent-obstruent clusters are not normally broken up by an epenthetic vowel, though they occasionally are. Second, elision may form consonant-consonant clusters that are not orthographically indicated; however this was observed less than anticipated, possibly due to the effects of reading pronunciation even in fast speech. Third, the clusters *çb* and *çd* may be pronounced with a voicing mismatch, with a voiceless affricate followed by a voiced consonant, but sometimes assimilation takes place and they are either all voiced (in certain fast speech conditions) or all voiceless.

Further research is needed in several areas. Phonetic and phonological motivations for the clustering constraints that I have described should be investigated. Determining when and under what conditions schwa-elision occurs requires work with more speakers and probably a more natural elicitation task — or a range of types of elicitation in order to determine what types of language use promote and inhibit elision. In addition, research needs to be extended beyond the initial two consonants into clusters of three and four consonants; for example, are the constraints on C_1C_2 the same in three and four consonant clusters as in two-consonant clusters? Are the constraints on C_2C_3 at all related to the constraints on C_1C_2 ? A conclusive determination of whether *mp, mb* etc. represent clusters or prenasalized stops also requires further work; Riehl's (2008) phonetic criteria provide a promising method of investigation; combining phonetic and phonological investigation could help both to determine the status of NCs in Albanian and to test whether Riehl's findings are supported in languages besides the ones she studied.

Appendix 1: Newmark's (1998) entries for words counted as marginal in 0

pn:	pneumoni'a <i>nf</i> [<i>Med</i>] pneumonia
pt:	ptu <i>interj</i> stylized spitting to represent spite/contempt for someone
kh:	khu-khu <i>onomat</i> coughing sound of someone choking
çh:	çhund or ëz i'm <i>nm</i> [<i>Ling</i>] denasalization
shv:	shvenk <i>nm</i> [<i>Cine</i>] Swiss-pan, flash-pan
dv:	dvier• <i>vt</i> = degjenero'•n
vrr:	vrrro'më <i>nf</i> slut, slattern = mola're
zll:	zlloti <i>nm</i> zloty (Polish money)
zhhd:	zh dhjam o's• <i>vt</i> to take the (excessive) fat off of []
hr:	hrushovia'n <i>nm</i> Khrushchevian
thn:	thnegël <i>nf</i> [<i>Entom</i>] ant

Appendix 2: Stimulus words: obstruent-obstruent clusters

(Abbreviated definitions based on Newmark 1998)

Obstruent-obstruent clusters: No elision

çboj	'undo' (nonstandard; standard <i>zhbëj</i>)
çdo	'every'
çka	'what'
çfarë	'what'
çfejoj	'make X break off X's engagement' (nonstandard; standard <i>shfejoj</i>)
zhbllokoj	'unblock'
zhdukje	'disappearance'
zhgunë ²⁴	a kind of white wool fabric; a cloak made of this fabric
zhvillim	'development'
shpoj	'pierce'
shtoj	'increase'
shkoj	'go'
shfajësoj	'exonerate'
vdekja	'the death'
tkurrje	'contraction'
KFOR-i	'Kosovo Forces'
gdhiu	'it dawned'

²⁴ Newmark (1998) spells this word *zhgun*; the spelling above was offered by the speaker in my study.

Appendix 3: Stimulus words not analyzed in this paper

<i>Sonorant-obstruent clusters</i>		<i>Stop/affricate</i>		<i>Nasal</i>	
mposht	‘defeat’	posht	‘low’	mos	‘don’t’
mposhti	‘defeat’ (infl.)	posti	‘the post’	mosha	‘the age’
mplaket	‘he gets old’	plaket	‘he gets old’	mjetet	‘the tools’
mbiemri	‘the surname’	bileta	‘the ticket’	minuta	‘the minute’
mbështetje	‘support’	bërthama	‘the kernel’	mërzitem	‘get bored’
mbledhje	‘meeting’				
ndal	‘stop’	dal	‘go out’	nam	‘reputation’
ndoshta	‘maybe’	dosja	‘the file’	nofka	‘the nickname’
nxënës	‘pupil’	xixa	‘the spark’	nëna	‘the mother’
nxehje	‘heating’	xixat	‘the sparks’	nënat	‘the mothers’
				nisje	‘departure’
ngadalë	‘slowly’	gazetë	‘newspaper’	natyrë	‘nature’
ngushëllime	‘condolences’	gazolina	‘the gasoline’		
		gabimisht	‘mistakenly’		
ngjall	‘resurrect’	gjallë	‘alive’		
ngjak	‘in (your) blood’	gjak	‘blood’		
ngjarje	‘event’	gjendje	‘state’		

Miscellaneous

tmerr ‘terror’

çmim ‘price’

mllef ‘anger’

Xrxa (Village name) The speaker seemed to say this with a syllabic *r*; when asked, she said it was probably Xërxa in written Albanian.tlynë According to my speaker, this was something similar to butter but not the same. This appears to be the same word that Newmark (1998) gives as tëlyen ‘butter’, with an elidable *ë*.msheftas ‘hide-and-peek’ (Nonstandard; see *mshef-* in Newmark (1998))

References

- Beekes, Robert S. P. 1995. *Comparative Indo-European linguistics*. Amsterdam: John Benjamins.
- Boersma, Paul & David Weenink. 2008. Praat: Doing phonetics by computer (Version 5.0.21) [Computer program]. Retrieved April 28, 2008, from <http://www.praat.org/>
- Buchholz, Oda & Wilfried Fiedler. 1987. *Albanische Grammatik*. Leipzig: VEB.
- Curtis, Matthew. 2008. *Xhorxh, xhuxhmaxhuxh, and the xhaxhalar*: The allolinguistic status of Albanian /xh/. Paper presented at the 16th Balkan and South Slavic Conference, May 1–4, Banff, Alberta, Canada.
- Demiraj, Shaban. 1996. *Fonologjia historike e gjuhës shqipe*. Tirana: Toena.
- Friedman, Victor A. *Albanian grammar*. 2006. Online publication of The Slavic and East European Language Research Centre. Available at http://www.seelrc.org:8080/grammar/pdf/albanian_bookmarked.pdf . Last accessed Sept. 20, 2008.
- Messinger, Heinz, Gisela Türck & Helmut Willmann. 1993. *Langenscheidt compact German dictionary*. Berlin: Langenscheidt.
- Moosmüller, Sylvia & Theodor Granser. 2006. The spread of Standard Albanian: An illustration based on an analysis of vowels. *Language Variation and Change* 18(2): 121–140.
- Newmark, Leonard. 1998. *Albanian-English Dictionary*. Oxford: Oxford University Press.
- Newmark, Leonard, Philip Hubbard & Peter Prifti. 1982. *Standard Albanian: A reference grammar for students*. Stanford: Stanford University Press.
- Pllumi, Zef. 1995. *Rrno vetëm për me tregue*.
- Riehl, Anastasia Kay. 2008. *The phonology and phonetics of nasal obstruent sequences*. Dissertation: Cornell University.
- Stefanllari, Ilo. 1996. *Albanian-English English-Albanian Dictionary* (Hippocrene Practical Dictionary). New York: Hippocrene Books.

THE EARLY MODERN GENITIVE *ITS* AND FACTORS INVOLVED IN GENITIVE VARIATION¹

Salena Sampson
Ohio State University

Abstract

This article explores the variation between the emergent genitive *its* and the periphrastic form *of it* in Early Modern English, situating this case in the larger picture of English genitive variation. As previous studies have often focused on non-pronominal possessors (given that Present Day English pronominal possessors often appear pronominally, with limited variation), this early pronominal genitive variation provides unique insight as it illustrates some of the same factors significant in pronominal genitive variation as in other cases. Additionally, as neuter pronouns commonly correlate with inanimate referents, this variation provides new evidence on the independence of weight and animacy in genitive variation. The importance of another factor, pressure from the pronoun paradigm, is also illustrated.

1. Introduction

The variation between genitives (e.g. *the book's cover*) and *of*-constructions (e.g. *the cover of the book*) has been a topic of investigation in studies of both historical and Present Day English (e.g. Rosenbach 2002, 2005; Rosenbach and Vezzosi 2000; Leech, Francis and Xu 1994; Altenberg 1982). Previous studies, however, have tended to focus on constructions involving non-pronominal possessors, as most pronominal possessors,

¹ Earlier versions of this paper were presented at the Dictionary Society of North America conference at the University of Chicago and at LSA in Chicago.

or possessive pronouns, in Present Day English strongly prefer a prenominal position, thereby limiting variation. However, in spite of apparent differences in distribution, at least in Present Day English, a more unified analysis may be possible. This study shows how the emergence of the Early Modern neuter genitive *its*, and the resulting variation between this emergent form and the periphrastic form *of it*, provide special insight into the relationship between pronominal genitive variation and other cases of genitive variation in that they demonstrate the same factors to be significant in this early case of variation as in other non-pronominal cases. In particular, this study highlights the importance of weight as a factor in determining genitive variation. As the use of neuter pronouns commonly correlates with inanimate referents, these results are also significant in that they provide new evidence on the independence of weight and animacy in genitive variation.

Prior to the Early Modern period, the form *his* served as both the masculine and the neuter third person singular genitive possessive pronoun form, as seen in (1) below.

- (1) The wide sea with all his billows raves. (Pope 1725: XI. 195)

As grammatical gender was lost in English, it became increasingly awkward to use this form, more and more associated with masculine gender, in neuter contexts, as can be seen by the increased avoidance of this form in neuter contexts. By the time of the earliest attestations of the new analogical form *its* in the middle of the sixteenth century, according to corpus data, the neuter genitive *his* was already dramatically in decline, making up only around 26% of the total third person singular neuter genitive constructions. Instead, speakers used a number of alternate constructions, such as *thereof*, *of the same*, and most notably the periphrastic form *of it* (Nevalainen and Raumolin-Brunberg 1994).

2. Previous Literature

There has been very little research devoted to the emergence of this new genitive pronoun, though it is remarkable as it is one of the major grammatical developments of the period and also constitutes an addition to a rather conservative closed class, the system of personal pronouns (Baugh and Cable 2002). Other than standard textbook accounts, noting *its* as a new analogical form with the basis of analogy being other 's genitives (e.g. *John's*, *the book's*), there is just one prior study providing a more detailed look at the emergence of the genitive *its* and making use of corpus data, as cited above (Nevalainen and Raumolin-Brunberg 1994). This study considers a number of factors in the selection between *its* and comparable periphrastic forms, with a focus on the relationship between the possessor and the possessum. As weight has been found a major factor in the choice between 's genitives and *of* forms generally, both historically (Altenberg 1982) and in contemporary English (Rosenbach 2002), Nevalainen and Raumolin-Brunberg (1994) consider weight as a potential factor, but ultimately rejects it as an unimportant factor.

There is a fair amount of research on the selection between 's genitive forms and *of* constructions (taking the form 'the N of NP'), as mentioned above; and an assortment of features has been considered in the selection between the two forms – the relationship

between possessor and possessum, animacy, weight, and phonological factors. While all of these factors have been found to be significant in the selection between these forms in contemporary English, a more recent study poses the question of whether or not animacy and weight are two distinct factors, as they have been shown to be highly statistically correlated – with animate nouns, which are generally lighter, more frequently occurring as prenominal possessors (Rosenbach 2005). Rosenbach ultimately argues that these factors are distinct, using both experimental and corpus data.

3. Historical Insight into Contemporary English

The emergence of the new form *its* and this period of variation and are naturally of interest in their own right as this form constitutes an addition to a closed class system, the personal pronouns. In addition, however, the patterns of usage associated with the new form *its* may illuminate a broader spectrum of English genitive constructions. As the innovative form *its* is generally understood as an analogical form, based on analogy with other *'s* genitives, perhaps the early competition between *its* and *of it* can shed light on the larger question of the selection between *'s* genitives and *of* constructions. Specifically, since *its* is a neuter form, it provides a unique opportunity for exploring the distinctness of two previously explored factors – animacy and weight.

4.1. Corpus and Tools

Whereas previous work on the emergence of the genitive *its* (Nevalainen and Raumolin-Brunberg 1994), has made use of only the Early Modern English sections of the Helsinki corpus, resulting in a small sample (only 107 instances of the genitive *its*), the current study makes use of the Lampeter Corpus, which is a larger corpus specifically devoted to Early Modern English. This corpus of approximately 1.1 million words of running text is comprised of Early Modern English tracts, with a balanced selection of tracts pertaining to subject matter, divided into six categories: religion, science, law, economics, politics, and miscellaneous.

Processing of the data consisted of a combination of the use of a basic concordancing program and hand editing of the resultant concordance data. This combination allowed for a more detailed analysis, ensuring that only cases where variation could at least in principle be considered possible would be included.

4.2. Selectional Criteria

With regard to this consideration, only constructions where the possessor-possessum relationship was subjective, as illustrated in (2a), objective, as illustrated in (2b), or possessive, as illustrated in (2c), were included, as these are the contexts which have previously been identified as choice contexts, where there may be variation between the two constructions (Rosenbach 2002).

- (2) a. The sun was observed before *its* setting to appear of a pale and dead color
- b. They and their instruments were the first kindlers *of it*
- c. It continued acting *its* illegal cruelties, upon all occasions

In example sentence (2a), the possessor *its* acts as the subject of the gerund *setting*; in example sentence (2b), *it* acts as the object of the possessum *kindlers*; and in example (2c), *its* stands in a general possessive relationship with *illegal cruelties*. While there certainly are statistical tendencies for the preference of one form versus the other in these cases, with *its* being preferred in subjective and possessive relationships, and *of it* being preferred in objective relationships, it has been argued that a choice between forms is at least theoretically possible (Rosenbach 2002) for each of these. (Compare such relationships with partitive relationships, for example, which categorically require *of* constructions: “one of the geese” versus **“the geese’s one”*.)

In addition, there is a further restriction related to definiteness. Since, the possessive pronoun *its* acts as a definite determiner, in that it cannot be used in addition to another definite determiner such as *the*, a definite determiner is required to head the noun phrase in the cases of periphrastic constructions with *of it* in order to establish real equivalence. Therefore, all cases of the periphrastic *of it* attaching to nouns with no determiner have been excluded. In other words, syntactically, (3) and (4) have been treated as equivalent.

- (3) *its* N
- (4) the N *of it*

Also, with regard to syntax, cases involving postmodification (e.g. “its appearance *in print*”) have been controlled for, following Rosenbach (2005). Since postmodifiers range in syntactic complexity, including, for example, relative clauses, prepositional phrases, and other postmodifiers; premodifiers serve as a better measure of weight as they represent less variation in syntactic complexity, often being simply adjectives. Though, as Rosenbach (2005) points out, weight and syntactic complexity correlate, selecting a specific, concrete measure of weight, such as premodification, may help parse out the effects of one versus the other.

Finally, fixed phrases, most notably *its own* and *its self*, must be excluded from a variationist analysis given their high degree of collocation which more importantly reflects the impossibility of an alternative comparable *of* construction (e.g. **“the self of it”*).

Though the idea of grammatical variation is controversial, as it is difficult to argue that different constructions truly mean “the same thing” (for a recent affirmation of these difficulties, see Guy 2007), these measures have been taken in an effort towards a variationist analysis of these genitive constructions.

5.1. Variation Related to Time

Analysis of data from the Lampeter Corpus confirms previous accounts on the time line of the emergence of the new form *its*, as displayed in Figure 1 below. The new form *its* first becomes common in print in the early to middle seventeenth century, as can be seen in the relatively equal proportions of *its* and *of it* in the first two decades represented in the corpus. Prior to these earliest decades represented in the Lampeter Corpus,

attestations of the new form are comparatively less frequent. The relatively equal counts for *its* and *of it* constructions during the first two decades represent the growing frequency of *its*, but also the lingering usage of other periphrastic forms such as the previously mentioned *thereof* and *of the same*, the lower total number of neuter genitives represented in the table for these decades being a product of the use of these other forms. Subsequent to the 1660's, the innovative form *its* is relatively more common. Though the new form is not commonly found in writing until the mid seventeenth century, it is important to note the often conservative nature of written texts. The new form, then, may have been in circulation, perhaps in spoken discourse, for some time before that period.

Decade	Form	Text Genres						Total
		Econ.	Pol.	Law	Rel.	Sci.	Msc.	
1640	<i>its</i>	5	1	6	12	10	0	34
	<i>of it</i>	2	1	6	21	4	1	35
1650	<i>its</i>	3	5	11	13	3	1	36
	<i>of it</i>	10	3	1	0	10	9	33
1660	<i>its</i>	17	12	0	27	62	57	175
	<i>of it</i>	16	7	0	12	12	3	50
1670	<i>its</i>	9	3	0	20	40	10	82
	<i>of it</i>	10	3	3	10	3	4	33
1680	<i>its</i>	9	17	5	27	52	16	126
	<i>of it</i>	15	17	3	2	17	5	59
1690	<i>its</i>	12	18	3	6	16	0	55
	<i>of it</i>	1	14	18	10	6	6	64
1700	<i>its</i>	39	7	2	24	39	10	121
	<i>of it</i>	10	4	8	15	6	2	45
1710	<i>its</i>	9	16	9	14	19	22	89
	<i>of it</i>	1	1	9	8	12	4	35
1720	<i>its</i>	0	3	5	26	19	39	92
	<i>of it</i>	4	0	5	19	20	5	53
1730	<i>its</i>	2	5	18	3	17	25	70
	<i>of it</i>	1	7	21	2	7	2	40
Total:	<i>its</i>	105	87	59	172	277	180	880
	<i>of it</i>	80	57	74	119	97	41	447

Figure 1. Relative Distribution of the Genitive *its* and *of it* in the Lampeter Corpus

In an effort to faithfully represent the emergence of the new form *its* with regard to the dimension of time, all instances of *its* and *of it* have been included in this table, including fixed collocations, such as *its own* and *its self* and cases with postmodification, since these comprise a considerable portion of the early usages, with 62 individual instances of the collocation *its own*, for example. These figures then are more useful in depicting trends over time of the emergence of this new form, without regard to whether or not the forms are completely interchangeable in each circumstance. A variationist analysis with statistical comparison taking into consideration all of the selectional criteria identified above appears in subsequent sections.

5.2. Variation Related to Subject Matter

There is variation in the choice of neuter genitive in relation to the subject matter of the text, as well. Scientific texts have one of the highest proportions of the innovative form, which accords with previous accounts (Nevalainen and Raumolin-Brunberg 1994). Interestingly, however, religious tracts show a similarly high count of the form *its*, though they have traditionally been thought to be a more formal register, less conducive to use of innovative forms (Altenberg 1982). Altenberg does however note that religious texts constitute “one of the most heterogeneous genres” (p. 256), with texts which are expected to be read (as opposed to heard), having higher proportions of the *'s* genitive in general. The discrepancy in results may point to the heterogeneity in this text type, as well as perhaps a need for a wider range of linguistic features to be considered in the labeling of this text type as generally more “conservative” or “innovative”.

5.3. Variation Related to Weight

Weight as discussed in terms of grammatical variation has been characterized and measured in a number of different ways. In the case of genitive variation, while some studies have considered the direction of the syntactic branching of the possessor and possessum (e.g. Jucker 1993), other studies have attempted to characterize syntactic complexity of the two noun phrases in terms of number and type of constituents (e.g. Altenberg 1982), and yet others have simply counted relative number of words (e.g. Biber et al 1999, Altenberg 1982). While all of these measures have been previously used to discuss weight and the relative weight of possessors and possessums in studies of genitive variation, Rosenbach (2005, p. 617) argues that if we are concerned chiefly with “weight”, it is best to control for syntactic complexity. Rosenbach then argues that this may be accomplished by counting premodifiers on the noun phrases: premodifiers are almost always adjectives and are more constrained in variety than postmodifiers, which may be prepositional phrases with varying lengths and complexity or varying types of dependent clauses, among others (Rosenbach 2005). By this measurement, the possessor in (5) would be “heavier” than the possessum as it is modified; and the possessum in (6) would be “heavier” than the possessor by the same reasoning.

- (5) The *red* book’s cover
- (6) The book’s *leather* cover

In the case of variation with regard to the neuter genitive pronoun, the possessor will always be a comparatively light element, being only a single word, *it*. Therefore, if the possessum is modified, it will be heavier than the possessor.

The previously observed trend in genitive variation with full noun phrases is that heavier elements generally appear later in the construction. So, if the possessor is heavier, this will generally make an *of* construction more likely; whereas, if the possessum is heavier, this will generally make an 's genitive more likely. In the case of genitive variation with the neuter genitive pronoun, we would then predict that if the possessum is modified, these conditions would prefer the new form *its*, as the possessum would then be heavier than the possessor and would be expected to appear after the comparatively lighter possessor.

Given an increased sample set from the Lampeter Corpus, relative weight, as measured by premodification of the head noun, does prove to be a significant factor in the selection of a genitive form (χ^2 , $p < .01$), the relative counts being displayed in Figure 2 below. Specifically, premodified heads prefer the new form *its*, as seen in the contrast between (7) and (8). Here, the modified head *love* in (7) takes the pronominal form *its*, and the non-modified head *Inhabitants* in (8) takes the periphrastic form *of it*:

- (7) God forbid that ever this Parliment should lose any of *its* first love to Religion.
- (8) It is named thus originally from the Lappi or Lappones, the Inhabitants of *it*.

Modification	Genitive Form	
	<i>Its</i>	<i>Of it</i>
Premodification	186	68
No Premodification	482	307
Total:	668	375

Figure 2. Variation and Premodification

This is in accordance with previous predictions that items with more weight appear later (Rosenbach 2002, 2005; Altenberg 1982), though contrasting with previous conclusions (Nevalainen and Raumolin-Brunberg 1994), derived from an analysis of the relatively smaller sections of the Helsinki Corpus. Given the fact that these new results regarding the importance of weight in this case of genitive variation match up with predictions with regard to weight made by related non-pronominal constructions, one interpretation that suggests itself is that the larger corpus has allowed for a previously unavailable statistical comparison of relatively low frequency occurrences, modified noun phrases with neuter pronoun possessors from the period when *its* was first in usage and already a relatively low frequency occurrence.

5.4. Avoidance of Repetition of the Same Form

Another syntactic pattern that is significant in the selection between forms relates to the avoidance of repetition of structure. Specifically, the new form *its* is more likely to occur

in the object of a preposition in a prepositional phrase headed by *of* than a noun phrase modified by *of it* (χ^2 , $p < .01$), as seen by the relative counts displayed below in Figure 3:

Position	Genitive Form	
	<i>Its</i>	<i>Of it</i>
In OP headed by <i>of</i>	112	27
Other	556	348
Total	668	375

Figure 3. Variation with Regard to Position: Whether in Object of Preposition

This variation is illustrated by the following examples in (9) and (10), with examples such as (10) being more common.

- (9) ...notice being given to the generality of the Trustees of the meeting, and
of the end of it
- (10) ...have been the greatest obstructor's of its relief heretofore

Though the use of the neuter genitive or *of* construction in the object of a prepositional phrase headed by *of* is a relatively low frequency occurrence, there are clear patterns in the choice between the genitive and the *of* construction in this context. Specifically, the use of the genitive is more frequent in these contexts, and may be attributed to considerations involving prosody or the avoidance of repeating the same form. This pattern, too, fits predictions made by patterns of use with other genitive constructions, where *combinations* of 's genitives and *of* constructions are most common in the case of nesting genitives, at least as early as the Early Modern period (Altenberg 1982).

6. Discussion

As can be seen from the previous syntactic evidence, in the period when the innovative form *its* first emerged, patterns of variation between *its* and *of it* correspond with larger patterns of variation between 's genitives and *of* constructions both in Early Modern and in Present Day English. As the new form *its* was formed by analogy with other 's genitives, these similarities in patterns of usage, though previously unobserved, perhaps are of little surprise.

However, in addition to being influenced by larger trends in genitive variation with full noun phrases, the new form seems eventually to show influence from the rest of the pronoun paradigm as well. Jucker (1993), in a corpus analysis of Present Day English, found that 98.5% of pronominal possessors take the form of a personal pronoun, as opposed to an *of* construction, leading Rosenbach (2002) to treat pronouns as a categorical environment with regard to genitive variation. Similarly, Nevalainen and Raumolin-Brunberg (1994) found only 50 instances of the construction *of it* in the Lancaster-Oslo/Bergen Corpus of Present Day English. These findings suggest that periphrastic possessive constructions involving pronouns, including *of it*, are rather uncommon in Present Day English.

While the periphrastic form *of it* may have been relatively common in English prior to the point when the innovative form *its* became established in the Early Modern period, periphrastic possessive constructions involving most of the other personal pronouns even in the Early Modern period were nearly categorically absent, with one clear exception – periphrastic possessives involving the pronoun *them*.

Construction	Occurrences with <i>of</i>	Total Occurrences of Pronoun
The N of <i>me</i>	3	1,226
The N of <i>you</i>	5	3,342
The N of <i>him</i>	24	2,910
The N of <i>her</i>	3	1,572
The N of <i>us</i>	13	2,155
The N of <i>them</i>	174	5,162
The N of <i>it</i>	447	12,887

Figure 4. Frequencies of *of* Constructions with Other Pronouns²

We can surmise that this exception with the form *them* is likely not related to number, as the frequency of periphrastic possessive constructions involving *us* is not similarly elevated. Instead, the distribution of possessive constructions involving *them* seems to pattern with the distribution of *it* possessives. The pronominal possessive *their* still far outnumbers *of them* constructions (7,356 instances of *their* as opposed to only 174 instances of the *of them* construction) as compared to *its* and *of it* (with 880 instances of *its*, and 447 instances of *of it*). The comparatively high frequency of periphrastic *of it* constructions in the Early Modern period, especially in light of their gradual decline in frequency since then seems to reflect a period of instability in which the new form *its* was still in the process of being established. Remarkably, what these two periphrastic possessive constructions, *of it* and *of them*, have in common seems to relate to the previously mentioned factor of animacy.

While all of the other personal pronouns – *me*, *him*, *you*, *her*, *us* – are used to refer almost exclusively to animate referents, *it* is most commonly used to refer to inanimate referents, and *them* may be used to refer to inanimate referents. It seems then that the factor of animacy may be playing an important role in the slight elevation in frequency of these two periphrastic pronominal possessive constructions. Yet, as these forms constitute a clear minority of the total inanimate pronominal possessive constructions (when compared with *its* and *their*), it appears as if these two periphrastic forms may be experiencing some pressure from the rest of the pronominal system, which has chiefly animate referents, and which clearly favors pronominal possession, as other animate nouns do.

7. Conclusions

Given the clear patterns of usage associated with weight, the variation between *its* and the *of it* constructions provides more evidence in favor of the distinctness of animacy and

² Note that the counts for *her*, *you*, and *it* may be somewhat inflated as these are raw counts, therefore including the determiner *her* and nominative *you* and *it*, whereas the counts for *me*, *him*, *us*, and *them* only reflect counts for objective case pronouns.

weight. It also provides evidence that the same constraints that operate synchronically on established constructions may come into play with the emergence of a new analogical form, in this case suggesting that there is no reason to treat this instance of variation differently than other cases of genitive variation, in spite of previous claims. When Nevalainen and Raumolin-Brunberg (1994) previously conclude that “pronouns do not behave in a similar fashion with genitive nouns or *of* phrases” (p. 194), they base this claim on an apparent lack of correlation between patterns of modification and weight in general NP genitive variation and those patterns of variation related to the innovative form *its*. Given the larger corpus used for this study, it does, however, appear that the initial variation between *its* and *of it* conforms to previously observed syntactic patterns of use with full NP genitives. Still, there may be some truth in the original claim, if for somewhat unexpected reasons. The innovative form *its* stands in a unique position in the history of the English language: as an analogical form, it is initially subject to the same patterns of variation as generally observed in the choice between *'s* genitives and *of* genitives. However, as this new form settles into the pronoun paradigm, it appears to be subject to competing pressures related to the patterns of usage associated with other pronominal genitives, which generally appear pronominally. And, yet, the periphrastic form *of it* has not disappeared from the language entirely and appears to represent a case of stable variation at this point in the language – a lingering testament to the strength of competition between different language internal factors, and the persistence of variation.

References

- Altenberg, Bengt. 1982. *The Genitive V. The Of-Construction: A Study of Syntactic Variation in 17th Century English*. Eds. Claes Schaar and Jan Svartvik, Malmö: CWK Gleerup.
- Baugh, Albert and Thomas Cable. 2002. *A History of the English Language*, 5th ed. New Jersey: Prentice Hall.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London/New York: Longman.
- Guy, Gregory. 2007. Grammar and Usage: the Discussion Continues. *Language* 83(1): 2-4.
- Jucker, Andreas. 1993. The Genitive Versus the Of-Construction in Newspaper Language. In Andreas Jucker (ed.), *The Noun Phrase in English. Its Structure and Variability. anglistik + englischunterricht* (49). 121-136.
- Leech, Geoffrey, Brian Francis, and Xunfeng Xu. 1994. The Use of Computer Corpora in the Textual Demonstrability of Gradience in Linguistic Categories. In Catherine Fuchs and Bernard Victorri (eds.) *Continuity in Linguistic Semantics*. Amsterdam: J. Benjamins. 57-76.
- Nevalainen, Terttu and Helena Raumolin-Brunberg. 1994. The Standardization of the Third Person Neuter Possessive. In Dieter Stein and Tiekens-Boon van Ostade (eds.) *Towards a Standard English*. New York: Mouton de Gruyter. 171-216.
- Rissanen, Matti and Ossi Ihalainen. 1984. *The Helsinki Corpus of English Texts. Diachronic and Dialectal*. Helsinki: University of Helsinki.
- Rosenbach, Anette. 2002. *Genitive Variation in English*, Eds. Elizabeth Traugott and Bernd Kortmann. New York: Mouton de Gruyter .

- Rosenbach, Anette. 2005. Animacy Versus Weight as Determinants of Grammatical Variation in English. *Language* 81(3): 613-644.
- Rosenbach, A. and Letizia Vezzosi. 2000. Genitive Constructions in Early Modern English: New Evidence from Corpus Analysis, *Stability, Variation and Change of Word-Order Patterns over Time*, Eds. R. Sornicola et al. (2000): 285-307.
- Schmied, Josef and Eva Hertel. 1998. *The Lampeter Corpus of Early Modern English Tracts*, Chemnitz.

LETTERED WORDS IN CHINESE: ROMAN LETTERS AS MORPHEME-SYLLABLES

Helena Riha
Oakland University

Abstract

In English individual letters are used to represent syllables, morphemes, and words in abbreviations. These uses of letters have been borrowed readily into Chinese, while the use of letters to represent phonemes in spelled words is less common. I discuss why the use of letters to represent units larger than the phoneme is more common in Chinese than their use in spelled words and what this reflects about Chinese morphology. I also argue that since abbreviations and letter-symbol words use letters as components of their structure, they show an interaction between orthography and morphology that should be recognized in morphological studies.

1. Introduction

In English individual letters normally represent phonemes, and sequences of letters indicate the pronunciation of syllables, morphemes, and words. There are cases, however, when individual letters represent units larger than phonemes, usually syllables, morphemes, or words in examples such as *PJs* (< pajamas), *e-mail* (*e* < e(lectronic)), *DJ* (< disc jockey), and ‘*n* ‘and’. Individual letters also have this function when they are used as *letter symbols* with contextual meanings, such as *A* in *grade A*, *A* in *A-line* (*dress, skirt*, and other women’s garments), *B* in *vitamin B*, *X* in *X-ray*, and so on. These various functions of letters are marginal in English, but interestingly, they are borrowed easily into Chinese, as are words formed with letters used in these ways.

2. Describing lettered words

There are many words in contemporary Chinese that are written fully or partly with roman letters. These so-called *lettered words* (*zìmǔcí*) are words written fully or partly in roman letters rather than Chinese characters. Many are borrowings, primarily from English, and most are initialisms (e.g. *NBA*, *WTO*), acronyms (e.g. *NASA*, *NAFTA*), or *hybrid words* composed of a roman letter component and a Chinese character component (e.g. ATM机 ‘ATM machine’). Many hybrid words are letter-symbol words in which one or more letters are used as symbols with particular contextual meanings that are usually the same as those in English. In X光 ‘X-ray’, for example, *X* means ‘unknown identity’; in A区 ‘area A’, *A* means ‘first’. A growing number of lettered words are also natively created initialisms, such as *KTV* ‘karaoke TV’, and hybrid words, such as PC机 ‘personal computer’, lit. ‘PC-machine’. Acronyms and ordinary spelled words are much less common. There are no native spelling-to-sound rules for their pronunciation, making them difficult to pronounce if one does not know English.

Lettered words have gained in popularity in Chinese in the last few decades and now form an established category of new words in the language. An appendix of lettered words is included in the *Xiàndài Hànyǔ Cídiǎn* ‘Dictionary of Modern Chinese’, the authoritative dictionary of Standard Mandarin commonly used by the public in the People’s Republic of China. Each new edition includes an ever greater number of lettered word entries in the appendix, indicating the growing influence of lettered words on the language.

3. Corpus study of lettered words in Chinese

I investigated lettered words in the *Chinese Gigaword Third Edition*, a corpus of Chinese newswires (Graff 2007), examining newswires from Xinhua, China’s state news agency, during the sixteen-year period from 1991 to 2006 (Riha 2008a). My aim was to describe the characteristics of lettered words and to understand trends in their use during that time. The corpus study revealed that the use of lettered words in Xinhua newswires increased continuously from 1991 to 2006. Based on this finding, I expect that the use of lettered words will most likely continue to grow in popularity in both newswires and in other written contexts.

The most common types of lettered words in the Xinhua corpus are English initialisms and acronyms, such as *WTO* and *APEC*, and hybrid words with both roman letter and Chinese character components, such as K他命 ‘ketamine’, pronounced [kei²ta⁵⁵min⁵¹]¹ in Standard Mandarin. Ordinary spelled words are much less common.

The use of initialisms, acronyms, and hybrid words is perhaps best appreciated by examining a sample of a Chinese text that contains lettered words. The text in (1) is an excerpt from an article in China’s *People’s Daily Newspaper*, quoted in Zhang (2005), which includes numerous lettered words.

¹ As noted in §4 on the use of lettered words in speech, the tones for roman letter names have yet to be standardized in Mandarin. I therefore use question marks in place of numerals to indicate the tones of letter names in phonetic transcriptions.

(1)

APEC记者招待会后，我约了STV的记者和一群MBA、MPA研究生朋友，讨论中国加入WTO后IT业对GDP的影响。读MBA的张小姐本来想去.COM当CEO，但觉得IT业风险大...随后大家相约关掉BP机，也不上Internet的QQ和BBS聊天，而是去了KTV唱卡拉OK。

“After the APEC press conference, I invited a reporter from STV and a group of MBA and MPA graduate student friends to discuss the impact of the IT industry on the GDP after China’s entry into the WTO. Miss Zhang, an MBA student, originally wanted to go to a .COM to be a CEO, but she thinks the IT industry is too risky ... Afterwards, we all decided to turn off our beepers, and we didn’t chat on QQ or BBSs on the Internet. Rather, we went to a KTV to sing karaoke.”²

The excerpt contains fifteen unique lettered words, only two of which are spelled words (*.com* and *Internet*). The thirteen other lettered words consist of one acronym (*APEC*), nine initialisms (*STV*, *MBA*, *MPA*, *WTO*, *GDP*, *CEO*, *QQ*, *BBS*, *KTV*), and three hybrid words that contain an initialism and a Chinese character component (IT业 ‘IT industry’, lit. ‘IT-industry’, BP机 ‘beeper’, lit. ‘BP-machine’ (BP < beeper), 卡拉OK ‘karaoke’ [ka²¹⁴la⁵⁵ou³kei³]). This distribution of word types is clearly different from that of most English texts, which contain primarily spelled words and relatively few initialisms, acronyms, or letter-symbol words.

The large number of initialisms, acronyms, and hybrid words in the excerpt is representative of the corpus findings as well. The list in (2) shows the ten most common roman letter words found in the Xinhua corpus (Riha & Baker 2010).

(2) NBA, GDP, DNA, WTO, APEC, OK, H5N1, IT, IBM, CBA

Nine of the ten words are initialisms and one is an acronym (*APEC*). None of the most frequent words in the corpus are spelled words. I interpret this as further evidence of a preference in Chinese for initialisms and acronyms over spelled words.

4. Use of lettered words in speech

In contrast to most ordinary words in Chinese, many lettered words are used primarily in writing rather than in speech. Lettered words appear frequently in news writing, technical writing, and computer-mediated communication (Zhang 2005, Gao 2007). Speakers differ, however, in how much they use lettered words in speech. The use of lettered words in the speech of Chinese speakers varies based on numerous factors, such as the speaker’s level of education, English fluency, age, profession, personal interests, and

² Translation provided by the author.

³ The word 卡拉OK ‘karaoke’ is a borrowing of the Japanese word カラオケ, lit. ‘empty-orchestra’, pronounced [karaoke] in Japanese ([oke] < orchestra). (Thanks to Kuniko Nielsen for the phonetic transcription of the Japanese word.) The Chinese characters 卡 [ka²¹⁴] and 拉 [la⁵⁵] have no meaning in this word. They are used as phonetic characters to represent the pronunciation of the first two syllables. The letters *O* and *K* are pronounced with their customary letter name pronunciation in Mandarin, [ou³kei³].

whether the speaker lives in an urban area. These factors can be summed up as an individual's level of participation in China's modernization (Riha 2006).

A comparison with English will help to illustrate how lettered words are primarily a product of the written language. English has a variety of words and expressions associated with a formal register that educated speakers use in writing but have few occasions to say in daily life. One category of this type is words and phrases borrowed from Latin. Speakers may vary in their pronunciation of certain words and phrases because they rarely need to say them aloud even though they recognize them in writing, and because they may rarely hear them spoken. Similarly, educated Chinese speakers may recognize a variety of lettered words in print but may not necessarily need to say them aloud and may not know their pronunciation because they rarely hear the words pronounced.

Another complicating factor regarding the pronunciation of lettered words is that their pronunciations have not been standardized and are not given in dictionaries. The appendix of lettered words in the *Xiàndài Hànyǔ Cídiǎn* provides glosses for lettered words but not their pronunciation. Presumably, speakers are expected to learn the pronunciations by listening to others rather than by consulting the dictionary.

5. Pronunciation of lettered words

In both initialisms and letter-symbol words, roman letters are usually pronounced individually with their letter names. Roman letters have official letter name pronunciations in China that are provided in dictionaries, but these pronunciations are generally not used to pronounce roman letters in lettered words. An alternate set of letter name pronunciations modeled on English letter names has emerged for this purpose (Riha 2006, Riha 2008a).

As with the pronunciation of individual letter names and letters in initialisms, the pronunciation of many acronyms borrowed from English is based on their English pronunciation. Thus, *APEC* is pronounced approximately as [eipek] and *IBM* as [aibiem]. In cases where speakers do not know that a particular acronym is conventionally pronounced as a word in Chinese, they pronounce it as an initialism.⁴ Natively created lettered words are normally also pronounced as initialisms rather than acronyms. A full discussion of the pronunciation of lettered words is provided in Riha (2008a). The main point here is that the pronunciations of lettered words have not been standardized and continue to evolve as lettered words make inroads into contemporary Chinese.

6. Popularity of lettered words

My corpus study shows that initialisms, acronyms, and letter-symbol words are considerably more frequent in Chinese newswires than spelled words, which I interpret to mean that these types of words are generally preferred to spelled words in Chinese. These

⁴ The same is true in English. For example, individuals who are not familiar with the *FTSE 100* (*Financial Times-Stock Exchange 100 Share Index*) would not know that it is commonly called the *Footsie* (Scott 2003) rather than being pronounced letter by letter.

types of words are preferred because they can be interpreted in terms of the most salient unit of organization in Chinese speakers' mental grammar: the Chinese *zi* (*zi*), or morpheme-syllable-character. Individual *zi* are the basic unit of metalinguistic knowledge for Chinese speakers (Riha 2008b). This view is reflected in Chao's (1968: 136-8) assertion that *zi* are the *sociolinguistic word* in Chinese, that is, "that type of unit, intermediate in size between a phoneme and a sentence, which the general public is conscious of, talks about, has an everyday term for, and is practically concerned with in various ways." Chao notes that while the sociological word in English is the familiar notion of *word*, in Chinese it is the *zi*. Larger units such as words and phrases are created from combinations of *zi*, as described by Chao (1968), Li & Thompson (1981), and Packard (2000), among others. The popular view of *zi* also appears to show that this unit is most salient for Chinese speakers. Packard (2000: 15) explains that in popular usage, the spoken morpheme-syllable and the character with which it is written are "one and the same thing" due to assumption that the spoken *zi*, or morpheme-syllable, can always be visually rendered as a written *zi*, or character.

I suggest that there is a structural correspondence between *zi* and individual letters when used in initialisms and acronyms and when used as letter symbols that makes them particularly suitable for incorporation into Chinese. Just as Chinese words and phrases are composed of sequences of morpheme-syllable-characters, these types of lettered units are composed of one or more *morpheme-syllable-letters* and possibly also one or more morpheme-syllable-characters. Chinese speakers essentially select linguistic units in English that are comparable to individual *zi* and sequences of *zi* in Chinese and integrate them directly into the spoken and written language to use alongside *zi*. Letter symbols and lettered words are borrowed, and new lettered items are subsequently created by analogy with borrowed items. For example, the letter *A* was borrowed with many of its meanings in English, one of which is 'first in order or in a series'. The letter *A* and this meaning of *A* are now used in expressions such as *A餐* 'set meal A' (from Hansell 1989), where *A* is used to indicate the first of several set meals on a menu. Similarly, a variety of English initialisms and acronyms have been borrowed, including *TOEFL*. The initialisms *HSK* and *PSC*⁵ are most likely formed by analogy with *TOEFL* and use the same process of abbreviation, albeit for romanized Mandarin words rather than English words.

This is not to say that only letter symbols, initialisms, and acronyms can be borrowed into Chinese and that "regular" spelled words cannot because they lack a correspondence between the individual letters and *zi*. As pointed out by Thomason (2001: 63), a language can borrow "anything" from another language, and my study certainly does show instances of spelled words in the Xinhua corpus. Chinese speakers simply appear to have a greater affinity for the use of letters as morpheme-syllable-letters than the ordinary use of letters as symbols for phonemes. Letters used in this alternate manner can be interpreted by analogy with similar units in Chinese speakers' native language system, whereas letters used to represent phonemes cannot.

⁵ *HSK* stands for *Hànyǔ Shuǐpíng Kǎoshì*, the romanized name of 汉语水平考试, the proficiency test of Standard Mandarin for foreign learners. *PSC* stands for *Pǔtōnghuà Shuǐpíng Cèshì*, the romanized name of 普通话水平测试, the proficiency test of Standard Mandarin for native speakers of Chinese varieties and other languages in China.

7. Roman letters as a new type of *zi*

Roman letters have the following functions in Chinese, among others: 1. to write foreign spelled words (*Microsoft*, *Olympics*), 2. to write romanized Chinese morphemes and words (*Běijīng*, *Shànghǎi*), 3. to write foreign initialisms and acronyms (*IBM*, *NAFTA*), 4. to create initialisms based on romanized Chinese words (*HSK*, *PSC*), 5. to write Chinese syllables that do not have a corresponding written character (*K* in *K他命* ‘ketamine’), and 6. to have independent meanings as morphemes that in many cases are the same as those in English (*A* means ‘first’, ‘top’, ‘best’, etc.; *X* means ‘unknown identity’). The manner in which letters are used in all of these functions except the first two, writing spelled words, give them the characteristics of *zi* and allow them to be interpreted in the same way as *zi*, as described below.

Chinese *zi* have the following features: 1. one or more pronunciations as a syllable, 2. one or more associated meanings or grammatical functions, and 3. a self-contained written form that fits into the imaginary ‘equidimensional square’ of Chinese writing (Boodberg 1957). This imaginary square requires each *zi*, no matter how complex its internal structure, to fit into a square of the same size. In addition to ordinary *zi* with these three features, there are also a small number of *zi* used as phonetic characters with no associated meaning or grammatical function. These *zi* are normally used to write loanwords (T’sou 2001). Examples include 咖 [*ga*⁵⁵] and 喱 [*li*³⁵], used in writing 咖喱 ‘curry’.

Example (3) shows the two *zi* in the compound meaning ‘gingko’. Each is written in an equidimensional square. The second *zi* is more complex internally than the first, but both fit into a square of the same size.

(3)

白	果
---	---

 ‘gingko’; 白 [*pai*³⁵] ‘white’, 果 [*kuo*²¹⁴] ‘fruit’

When individual letters are used to represent words (e.g. *PC机* ‘PC’, lit. ‘PC-machine’), syllables (*BB* ‘baby’), letter symbols (*维生素 A* ‘vitamin A’), or phonetic symbols (*K他命* ‘ketamine’), they also have the three features of *zi*. They have a letter name pronunciation that in many cases is just one syllable in length, an independent written form that can fit into an equidimensional square, and an independent meaning, grammatical function, or phonetic function, as shown in (4).

(4)

e	书
---	---

 ‘e-book’; *e* [*ji*[?]] ‘e(lectronic)’, 书 [*shu*⁵⁵] ‘book’

Example (4) shows that just like the Chinese *zi* 书 ‘book’, which is pronounced as a syllable and has an associated meaning and an independent written form, the letter morpheme-character *e* also has these same three features: it is pronounced as a syllable, has an associated meaning, and has an independent written form. Letters that can be interpreted as morpheme-syllable-characters in this manner have become a new set of *zi* in Chinese. These *roman letter zi* can stand alone as morphemes and words; they are pronounced as individual units, usually one syllable in length; and they can be combined in ways that are similar to those of *zi* to create new words and phrases (Riha 2008a).

8. Conclusion

In this article I have shown that Chinese speakers have a preference for using letters to represent morpheme-syllables, the basic linguistic unit in Chinese, rather than phonemes, as is normally done in English. English words in which letters represent morpheme-syllables are uniquely suited to the *zi*-based (morpheme-syllable-based) grammar of Chinese and are commonly used in the contemporary language. This use of letters is exploited in Chinese by interpreting letters in a Chinese manner, that is, as individual *zi*; sequences of letters are similarly interpreted as sequences of *zi*.

In the use of letters as roman letter *zi*, there is an interaction between orthography and morphology that has yet to be recognized in morphological studies. Roman letters used as morpheme-syllable-characters fit well into Chinese in part because each letter fits into the equidimensional square of Chinese writing in the same way as individual Chinese characters. The written form of roman letter *zi* is an integral part of their functionality in Chinese.

As familiarity with English continues to increase in Chinese society, there may be less reliance on the use of roman letters by analogy with Chinese *zi*, and more ordinary spelled words may be used. Nonetheless, lettered words will most likely continue to be primarily initialisms, acronyms, and letter-symbol words since they have the characteristics of Chinese *zi*.

References

- Boodberg, Peter A. 1957. The Chinese script: An essay in nomenclature (the first hecaton). *Bulletin of the Institute of History and Philology Academia Sinica* (Taipei) 39:113-120.
- Chao, Yuen Ren. 1968. *A grammar of spoken Chinese*. Berkeley: University of California Press.
- Chinese Academy of Social Sciences, Dictionary Editing Group. 2005. 现代汉语词典 *Xiàndài Hànyǔ Cídiǎn* [Dictionary of Modern Chinese]. Beijing: Commercial Press.
- Gao, Liwei. 2007. *Chinese Internet language: A study of identity constructions*. Munich: LINCOM Europa.
- Graff, Dave. 2007. *Chinese gigaword third edition*. Linguistic Data Consortium, Philadelphia.
- Hansell, Mark D. 1989. *Lexical borrowing in Taiwan*. University of California, Berkeley dissertation.
- Li, Charles N., and Sandra A. Thompson. 1981. *Mandarin Chinese: A functional reference grammar*. Berkeley: University of California Press.
- Packard, Jerome. 2000. *The morphology of Chinese*. Cambridge: Cambridge University Press.
- Riha, Helena. 2006. The pronunciation of lettered words in Mandarin Chinese. Paper presented at NWAV 35 (New Ways of Analyzing Variation 35), Columbus, Ohio.
- Riha, Helena. 2008a. *Lettered words and roman letter characters in Chinese writing: A study of alphabetic writing in Chinese newswires*. The Ohio State University dissertation.

- Riha, Helena. 2008b. Aronoff and sociological words. (Letter to *Language*.) *Language* 84, 1-2.
- Riha, Helena, and Kirk Baker. 2010. Lettered words: Using roman letters to create words in Chinese. In *Proceedings of the 13th International Morphology Meeting (IMM 13)*. Amsterdam: Benjamins.
- Scott, David L. 2003. *Wall Street words: An A to Z guide to investment terms*. Boston: Houghton Mifflin.
- Thomason, Sarah G. 2001. *Language contact: An introduction*. Washington, DC: Georgetown University Press.
- T'sou, Benjamin K. 2001. Language contact and lexical innovation. *New terms for new ideas: Western knowledge and lexical change in late imperial China* (= *Sinica Leidensia* 52), ed. by Michael Lackner, Iwo Amelung, and Joachim Kurtz, 35-56. Leiden: Brill.
- Zhang, Yihua. 2005. Lexical concerns about neologisms in Chinese lexicography: A cognitive approach to the motivated structure of new words and expressions. *Words in Asian Cultural Contexts: Proceedings of the 4th Asialex Conference (Asialex 2005)*, ed. by Vincent BY Ooi, Anne Pakir, Ismail Talib, Lynn Tan, Peter KW Tan, and Ying Ying Tan, 380-92. Singapore: National University of Singapore.

MULTILINGUAL ANIMACY CLASSIFICATION BY SPARSE LOGISTIC REGRESSION

Kirk Baker and Chris Brew

Abstract

This paper presents results from three experiments on automatic animacy classification in Japanese and English. We present experiments that focus on solutions to the problem of reliably classifying a large set of infrequent items using a small number of automatically extracted features. We labeled a set of Japanese nouns as \pm animate on the basis of reliable, surface-obvious morphological features, producing an accurately but sparsely labeled data set. To classify these nouns, and to achieve good generalization to other nouns for which we do not have labels, we used feature vectors based on frequency counts of verb-argument relations that abstract away from item identity and into class-wide distributional tendencies of the feature set. Grouping items into suffix-based equivalence classes prior to classification increased data coverage and improved classification accuracy. For the items that occur at least once with our feature set, we obtained 95% classification accuracy. We used loanwords to transfer automatically acquired labels from English to classify items that are zero-frequency in the Japanese data set, giving increased precision on inanimate items and increased recall on animate items.

1. Introduction

Distinguishing animate from inanimate noun phrases is important for a number of morphological and other linguistic processes. For example, most languages exhibit some type of syntactic or morphological alternations that reflect perceived distinctions in animacy (e.g., pronoun categories like s/he vs. it, or marking animate direct objects differently from inanimate ones). Animacy is associated with other potentially useful properties, including sentience, autonomy, and intentionality. The ability to make accurate inferences about the animacy of potential noun phrase referents is important for many natural language processing tasks such as pronoun resolution (Oräsan and Evans 2001), machine translation (Ilarraza et al. 2002), and language generation (Zaenen et al. 2004).

The core sense in which we are using the term “animacy” is straightforward. Animals (including humans) are animate; vegetables, minerals and manufactured objects are not. But it is far from straightforward to make detailed claims about the nature of the concepts that underly the linguistic distinctions that are typically labeled with the terms “animate” or “inanimate”. There is no fully convincing way of establishing that two languages are operating with the same concept of animacy. Worse, there are no generally agreed criteria for deciding that a particular linguistic pattern is really a reflection of any concept of animacy. The relevant distinction could very well turn out to be agency, sentience or the ability to move in an apparently autonomous fashion. We have no intention to resolve either of these difficulties here. We will provide examples of animacy distinctions in action, and assume without detailed argument that the ontologies partly encoded in the linguistic distinctions of different languages are sufficiently similar for the enterprise of cross-linguistic information transfer to make sense.

Most of the literature on animacy distinctions and its relevance for human language makes reference to an animacy scale, which arranges objects along a sort of discretum from animate to inanimate. The number of categories is not fixed, but depends to some extent on the degree of resolution deemed appropriate for the description of the phenomenon at hand. Sometimes a three-way distinction between humans, other animates and inanimates is made (e.g., Zaenen et al. 2004), but often humans and animals are both treated as animate (e.g., van Nice and Dietrich 2003, Bresnan and Hay 2006) and contrasted with everything else. According to Bresnan and Hay (2006), Bresnan et al. (2005) used a four way animacy distinction that consisted of the categories human, organization, animal/intelligent machine and inanimate. Yamamoto (1999) provides an elaborate radial hierarchy view of animacy with 1st person pronouns at the center, 2nd person pronouns and other people connected to that and extending outward to supernatural being, animals, machines, plants, etc.

As we will see in the examples below, language users can and will make flexible use of their language's conventions about animacy. A simple example from English is that pet animals can be referred to using either the masculine and feminine pronouns “he” and “she” or the neutral pronoun “it”. People who like pets and think of them as family members are more likely to choose the gender marked pronouns. If you do not like them, you will be more likely to use the neuter pronoun. In extremis the same thing can be done with babies, although the risk of opprobrium is correspondingly greater.

Patterns related to animacy show up in various ways in different languages. One relatively common way is for animate objects to be marked with some kind of case marker while inanimate objects are not. In Spanish, for example, patient noun phrases are marked with the preposition *a* if they are animate (1b), but not if they are inanimate (1a)(Yamamoto 1999:47, 13):

- (1a) Ha comprado un nuevo libro.
 have:3SG bought a new book
 'He has bought a new book.'
- (1b) Ha comprado *a* un nuevo caballo
 have:3SG bought *to* a new horse
 'He has bought a new horse.'

Similar facts are reported for Bantu languages like Swahili (Woolford 1999) and Native American languages like Blackfoot (Bliss 2005). Animacy has been claimed to play a role in languages with nominal classifier systems. For example, Japanese has an extensive system of numeral classifiers that selectively attach to subsets of the language's nouns. These are subsets are clearly related by properties such as animacy, shape, size and function (Iida 1999). A few examples are shown below. In the column labeled "Animacy" there are examples of four different classes of animate entities, each with its own numeral classifiers. Roughly, the *-hiki* class is for smallish animals, while the *-tou* class is for larger animals.

Animacy		Shape		Function	
<i>-nin</i>	<i>hito-ga san-nin</i> 'three people'	<i>-hon</i>	<i>ohashi-ga san-bon</i> 'three chopsticks'	<i>-dai</i>	<i>kuruma-ga san-dai</i> 'three cars'
<i>-hiki</i>	<i>hebi-ga san-biki</i> 'three snakes'	<i>-mai</i>	<i>shatsu-ga san-mai</i> 'three shirts'	<i>-hatsu</i>	<i>juusei-ga san-batsu</i> 'three gunshots'
<i>-tou</i>	<i>uma-ga san-tou</i> 'three horses'				
<i>-wa</i>	<i>tori-ga san-wa</i> 'three birds'				

Table 1. Some Japanese numeral classifiers.

Even in languages where animate and inanimate nouns have the same morphological properties, there are distributional differences in the occurrences of animate and inanimate nouns. For example, a corpus study of the syntactic distribution of animate and inanimate noun phrases in Swedish (Dahl and Fraurud 1996) found significant differences in the proportion of animate NPs that occurred as direct objects (13.0%) versus indirect objects (83.1%) and as transitive subjects (56.5%) versus intransitive subjects (26.0%). English evidences animacy preferences in double object word order (e.g., *give me money* versus *give money to me* and possessives (e.g., *'s* used more with

animates; *of* used more with inanimates) (Zaenen et al. 2004).

Taking animacy as a category that exists independently of linguistic categorizations of it necessitates a distinction between grammatical animacy and non-linguistic animacy.¹ Grammatical animacy is a type of lexical specification like person or number agreement that shows up in verb inflection and case marking. Like much conventionalized linguistic structure, the exact nature of the relationship between grammatical animacy and the correlates of non-linguistic animacy that supposedly underlie its use can be unclear. For example, Polish has been described as a language that employs grammatical animacy (Gawronska et al. 2002). Polish has four animacy classes that are distinguishable in terms of their accusative case marking and verb agreement:

- superanimate - grammatically masculine, human. Accusative form equals the genitive.
- animate - masculine or feminine living things. Singular accusative is the genitive; plural accusative is the same as the nominative.
- inanimate - masculine, feminine, or neuter non-living things. Accusative form equals the nominative.
- semi-animate - grammatically masculine, not living. Accusative form patterns like the animates.

The semi-animate nouns do not fall into any discernible semantic category, comprising things like names of dances, games or some actions; food, money, and negative mental states (Yamamoto 1999).

Similar phenomena exist for other languages noted for their sensitivity to animacy. For example, in Algonquian languages the distinction between animate and inanimate nouns overlaps in part with pronominal distinctions made in languages like English, but there are also many cases where correspondences are difficult to discern. For example, in Plains Cree, items like *opswākan* 'pipe', *mīhkwan* 'spoon', and *āpoy* 'paddle' are grammatically animate but do not correspond in a biological or grammatical sense to English nouns that are considered living (Joseph 1979). In Blackfoot as well, grammatical animacy cross-cuts sentience: nouns like *ato'ahsim* 'sock', *isttoan* 'knife', *pokon* 'ball', *po'taa'tsis* and 'stove' are grammatically animate but not alive (Bliss 2005). Closely related Bantu languages encode animate object agreement differently from each other depending on additional factors like definiteness, agentivity, person, and number (Woolford 1999). In spite of this sometimes unintuitive disconnect between grammatical and non-linguistic animacy, a great deal of research has gone into exploring the ways in which animacy distinctions surface in language.

A variety of natural language phenomena are said to be sensitive to distinctions in animacy. For example, in English the choice of the genitive is influenced in part by the animacy of the possessor. Jäger and Rosenbach (2006) conducted an experimental study

¹ Some authors make this distinction in terms of 'syntactic animacy' versus 'semantic animacy' (e.g., Yamamoto 1999:50 and references therein).

in which subjects were asked to read a short text such as the example below and choose which of the two underlined possessive constructions was more natural:

A helicopter waited on the nearby grass like a sleeping insect, its pilot standing outside with Marino. Whit, a perfect specimen of male fitness in a black suit, opened [the helicopter's doors/the doors of the helicopter] to help us board.

Subjects were more likely to choose the 's construction when the possessor was animate (e.g., *the boy's eyes*), than when the possessor was inanimate (e.g., *the fumes of the car*).

Animacy also plays a role in English in the syntax of verbs of giving (e.g., *give, take, bring*). Such verbs can be realized in two ways: the double object construction (2a), or the dative alternation (2b).

- (2a) Who gave you that watch?
- (2b) Who gave that watch to you?

Bresnan and Hay (2006) analyzed instances of the verb *give* that occurred in a corpus of New Zealand and American English talkers, and found that non-animate recipients were more than 11 times more likely to be realized in the prepositional dative than animate recipients were. (They also found that non-animates were more likely to appear in the double object construction in New Zealand versus American English.)

Dowty (1991) presents an analysis of a class of English verbs that are similar to verbs like *give* in that they permit a syntactic argument alternation that does not correspond to a difference in interpretation. Dowty (1991) refers to this as the *with/against* alternation, and an example is shown below (Dowty 1991:594, 62):

- (3a) John hit the fence with the stick.
- (3b) John hit the stick against the fence.

In contrast, a semantically similar class of verbs is claimed to not permit this alternation (Dowty 1991:596, 65):

- (4a) swat the boy with a stick
- (4b) *swat the stick at/against the boy

This class, which includes verbs like *smack, wallop, swat, clobber* is said to be distinguished by the fact that its verbs restrict their objects to human or other animate beings, and entail a significant change of state in their direct object arguments.

Animacy has also been claimed to play a role in Japanese grammar as well. As illustrated in Table 1, numeral classifiers are sensitive to animacy distinctions. According to Iida (1999) animacy is the most important of four basic semantic features that play important roles in determining which classifier is selected for a given noun:

ANIMACY > FUNCTION > SHAPE > CONCRETENESS.

Thus, in assigning a classifier to an object that could be counted with either one, (e.g., a snake which is both long and animate), animacy takes precedence. Robot dogs can be counted with animate counters (*-hiki* or *-tou*); according to Iida (1999), when “one feels a high degree of animacy in a certain object, one can count it with animate classifiers; on the other hand, when one does not find animacy in what one is counting, there is no way to use animate classifiers”.

Japanese has a number of other lexical items which are correlated with the animacy of their arguments. Although nouns are typically not marked to indicate number, there is a plural marker *-tachi* which, when it is used, is largely restricted to animate nouns (e.g., *watashi-tachi* 'us', *hito-tachi* 'people'). Japanese has two distinct verbs meaning 'to exist', *iru* and *aru* that show tendencies for selecting animate and inanimate subjects, respectively:

- (5a) *dansa-ga iru* = ANIMATE
 dancer NOM *exist*
 'There's a dancer.'
- (5b) *toosutaa-ga aru* = INANIMATE
 toaster NOM *exist*
 'There's a toaster.'

We are interested in the problem of learning a statistical classifier that can reliably distinguish animate from inanimate noun phrase referents on the basis of their contextual distribution in a large text corpus. As the examples above illustrate, we cannot assume that there is a single sense of animacy that applies across all languages. However, we are comfortable with the assumption that at some cross-linguistic level, perceptual properties commonly associated with animacy such as sentience, agency, and intentionality overlap one another and with the presumed biological basis of an animate/inanimate distinction.

In the empirical work reported here, we correlate animacy information across two historically unrelated languages, Japanese and English. Our hypothesis is that combining lexical information from multiple languages allows for more reliable automatic semantic classification than is possible using cues from only one language. We make no claim that the notions of animacy that are involved are exactly parallel. If pressed, we would argue that two independently drawn languages may make similar choices about the gross distinction between clearly animate entities (for example: human beings and large wild animals) and clearly inanimate entities (for example: rocks and pieces of wood). We would not necessarily expect usable consensus on the trickier cases (statues, robots, thermostats or the personified form of The North Wind).

2. Previous Work on Automatic Animacy Classification

Orăsan and Evans (2001) describes a method for animacy classification of English nouns that applies an instance based learning method to WordNet sense annotated corpus data. In a two step process, they first categorize WordNet senses by animacy, and use that information to classify nouns whose sense is unknown. The assumption motivating their methodology is that a noun with lots of animate senses is more likely to refer to an

animate entity in a discourse; conversely, a noun with a majority of inanimate senses is more likely to refer to an inanimate entity.

Oräsan and Evans (2001) defines animacy as the property of a NP whose referent, when singular, can be referred to pronominally with an element from the set $\{he, him, his, himself, she, her, hers, herself\}$. They explicitly reject classifying animals as animate, considering references to pets as *he* or *she* to be an “unusual usage of senses”.

The first part of their method involved manually annotating 20026 NPs in a 52 file subset of SEMCOR (Palmer et al. 2005) for animacy. In order to determine the animacy of a given sense, they work bottom up to calculate the proportion of animate hyponyms of that sense. If all of the hyponyms of a given sense are animate, that sense is classified as animate. If not, they use a chi-square test to determine whether the proportion of animate to inanimate hyponyms is reliably skewed enough for that sense to be labeled animate.

A similar procedure is used to label verbs senses as \pm animate. In this case, they use the proportion of animate or inanimate nouns that appear as subjects of a verb sense to decide the animacy label.

The number of animate and inanimate senses are used to make a vector representation of each noun. The features in this vector are:

- noun lemma
- number of animate noun senses
- number of inanimate noun senses
- number of animate verb senses
- number of inanimate verb senses
- ratio of animate to inanimate pronouns in the document

The last feature, the ratio of animate to inanimate pronouns, is calculated as the count of $\{he, she\}$ in the document divided by the count of $\{it\}$.

The second part of their methodology involved using TiMBL (Daelemans et al. 2003) to classify nouns on the basis of these feature vectors. They settled on a k-nearest neighbors classifier (k=3) using gain ratio as a weighting feature. The most important feature was the number of animate noun senses followed by the number of inanimate noun senses.

Oräsan and Evans (2001) offers two evaluations of their method: 5 fold cross validation on SEMCOR, and training on SEMCOR and testing on a set of texts from Amnesty International. Their results are shown in Table 2. Overall accuracy on both corpora is around 98%, and precision and recall for both classes ranges from about 90% to 98%.

	Animate		Inanimate	
Accuracy (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)

97.5	88.9	91.0	98.7	98.4	SEMCOR
	n = 2512		n = 17514		
97.7	94.3	92.2	98.4	98.8	Amnesty International
	n = 537		n = 2585		

Table 2. Classification results from Oräsán and Evans (2001), Table 3

Oräsán and Evans (2001) note that the most problematic words for their system were named entities, as these tended not to occur in WordNet. Other problematic cases were animate nouns that did not have an animate sense in WordNet (e.g., “Bob”). Because named entities were “constantly misclassified”, Oräsán and Evans (2001) removed them from their dataset. Oräsán and Evans (2001) also report that their system learned that most unknown words were inanimate, and in most cases correctly classified them as such. Given that 87% of the NPs in SEMCOR were inanimate, and that 83% of the Amnesty International data were also inanimate, it is not clear how much is actually being generalized by their system. Removing only problematic items from consideration makes interpreting their results difficult and illustrates some of the problems associated with a system based mainly on static lexical lookup.

Ilarraza et al. (2002) presents a dictionary based method for assigning animacy labels and a reliability score to Basque nouns that relies on manually labeling a small set of items and using synonymy and hypernymy relations to extend those labels to a larger set. The insight motivating their proposal is that conventions for defining words in a dictionary can be exploited to iteratively classify nouns on the basis of selective annotation.

According to Ilarraza et al. (2002), dictionary definitions come in three forms that can be used to extrapolate semantic information about lexical entries. The classical definition relates the entry to a hypernym plus some differentiating information, i.e.,

airplane: a **vehicle** (*hypernym*) that **can fly** (*differentia*)

Another method consists of specific relators that determine the semantic relationship between the entry and the definition:

horsefly: **name** (*relator*) given to a kind of **insect** (*related term*)

The third method is to list synonyms of the entry:

finish: **stop** (*synonym*), **terminate** (*synonym*),

Ilarraza et al. (2002) labeled the 100 most frequent hypernyms and relators that occurred in definitions in a Basque monolingual dictionary, and gave these a reliability rating of 1.0. Assuming that the animacy of a hypernym applies to all of its hyponyms, they search the dictionary and for each definition of an entry apply the animacy label of its relator or hypernym. That entry's reliability score is equal to the number of labeled keywords over the number of senses. For example, *armadura* 'armor' is labeled -animate with a

reliability of 0.66 because it occurred with three definitions, two of which contained a labeled hypernym:

Noun	# def	# hype	Labeling process			Anim.	Rel.
<i>armadura</i>	3	2	multzo [-]1	babesgarri [-]1	soineko []	[-]	0.66
<i>armor</i>			<i>collection</i>	<i>protector</i>	<i>garment</i>		

Table 3. Example animacy labeling process from Ilarraza et al. (2002).

This labeling process iterates such that if *armadura* 'armor' appears as a hypernym or relator of another noun in the dictionary, that entry can be labeled using the information assigned to *armadura* 'armor' in the previous iteration. After an iteration, synonyms are classified according to the labels just assigned. Ilarraza et al. (2002) states that automatic labeling asymptotes after about 8 iterations, with 75% of the nouns in the dictionary covered.

Ilarraza et al. (2002) offer two evaluations of their system, one in terms of items in the dictionary and another on corpus data. For the dictionary evaluation, they selected 1% of the nouns (123 items) for hand checking, and report 99.2% accuracy (75.1% recall). They do not report the breakdown of animate to inanimate items, do not state whether evaluation was in terms of senses of the checked items, nor explain the criteria for selecting the verification set. Ilarraza et al. (2002) report an overall recall of 47.6% of noun types, but do not report accuracy. Of the 3434 nouns labeled, 356 were classified as animate, and 3078 were classified as inanimate.

Øvrelid (2006) presents a machine learning approach to classifying nouns in Norwegian as \pm animate that is based on decision trees trained on relative frequency measures of morphosyntactic features extracted from an automatically annotated corpus. Øvrelid (2006) uses a set of linguistically motivated features that potentially correlate with animacy. Nouns are represented as vectors containing the relative frequency of each of the following features:

- transitive subject/direct object: the prototypical transitive relation involves an animate subject and an inanimate direct object.
- demoted agent in passive: a correlation is assumed between animacy and agentivity
- reference by personal pronoun: Norwegian pronouns distinguish antecedents by animacy (e.g., *han/hun* 'he/she' vs. *den/det* 'it-MASC/NEUT').
- reference by reflexive pronoun: assumes that the agentive semantics of the reflexive might favor animate nouns
- genitive -s: assumes that possession is a property of animate entities

Øvrelid (2006) reports several experiments looking at the effect of frequency on classification accuracy. The first evaluation is based on 40 high frequency nouns, evenly split for animacy. She reports overall accuracy of 87.5% when all features are used, but does not report precision and recall for each class separately. Leave one out training and testing shows that SUBJECT and REFLEXIVE are the best individual features (85% and 82.5% accuracy). Her second experiment looks at the classification accuracy of nouns

that occurred around 100, 50, and 10 times in the corpus (40 items in each bin). Classification accuracy drops to 70% using the classifier trained for the first evaluation. Although she does not report precision and recall, these values can be calculated from the confusion matrix for the nouns with frequency 100 in Øvrelid (2006:52, Table 5), for some indication of classifier performance.

Accuracy (%)	Animate		Inanimate	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
70.0	83.3	50.0	64.3	90.0

Table 4. Precision and recall values calculated from Øvrelid (2006).

Øvrelid (2006) also investigated the effect of backing off to the most frequent features, and found that only using SUBJECT, OBJECT and GENITIVE performs slightly better for the nouns with frequency 100 and 50, but shows no improvement for the less frequent group. A backoff classifier, trained on nouns of similar frequency, showed some improvement when using all the features, but was unchanged for individual features.

3. Experimental Setup

3.1 Data Set

Japanese has a number of lexical features related to animacy that make it relatively easy to automatically label a subset of nouns without recourse to an external lexical resource. We used three of these to initially label a set of training data as animate or inanimate:

- the plural suffix *-tachi*, which typically attaches to animate nouns.
- Japanese has two verbs meaning 'to exist'. *iru* is typically used with animate subjects, while *aru* typically occurs with inanimate ones.

We used CaboCha v0.53 (Kudo and Matsumoto 2002), a deterministic dependency analyzer for Japanese, to extract verb-argument pairs from the 229 MB (30M word) Japanese Business News Text corpus (Graff and Wu 1995). From its output, we extracted nouns marked by *-tachi* and subjects of *iru* and labeled them animate; subjects of *aru* were labeled inanimate. We defined subject as a nominative case-marked dependent of a verb, and object as an accusative case-marked dependent of a verb. We eliminated items that occurred as subjects of both verbs, and verified the remaining items against a native Japanese speaker's judgments.

We also identified and extracted English loanwords that appeared in the corpus and back-transliterated these into the original English source word (e.g., *herikoputaa* → helicopter). We used WordNet2.0 (Fellbaum 1998) to classify the English words as ±animate, and retained those nouns which had only animate or only inanimate senses. We assume that the animacy of the English transliteration is the same as that of its Japanese counterpart².

² It is certainly possible to find cases where the WordNet animacy label of an English source word

We obtained animacy labels for 14389 words (4433 animate, 9956 inanimate); 63% were English loanwords. Table 5 contains a categorization of items within the two animacy classes in terms of a set of fairly coarse-grained WordNet distinctions.

Semantic Category	Percent	Examples
Animate		
Occupation	59%	<i>chiji</i> ‘governor’, <i>gakusei</i> ‘student’
Proper Name	18%	<i>kiyouko</i> ‘Kiyoko’, <i>heminguwei</i> ‘Hemingway’
Human Being	15%	<i>obaachan</i> ‘grandma’, <i>watashi</i> ‘I’
Animal	8%	<i>saru</i> ‘monkey’, <i>iruka</i> ‘dolphin’
Inanimate		
Concrete	40%	<i>jushi</i> ‘resin’, <i>opaaru</i> ‘opal’
Activity	27%	<i>ragubii</i> ‘rugby’, <i>tsutae</i> ‘conveying’
Abstract	20%	<i>gihou</i> ‘technique’, <i>konomi</i> ‘taste’
Location	7%	<i>itarii</i> ‘Italy’, <i>sabaku</i> ‘desert’
Organization	6%	<i>renmei</i> ‘association’, <i>hamasu</i> ‘Hamas’

Table 5. Characteristics of the data set.

While these labeling methods are heuristic in nature, we have checked a sample of the results, and believe the labels to be of a quality comparable to what would be obtained by more costly methods. Of course, in focusing on the cases where simple indicators give reliable results, we sacrifice coverage relative to approaches that make heavier use of informed human judgment.

3.2 Feature Set

We considered three criteria in choosing a feature set: accuracy, coverage, and generality. One commonly used source of features is the pooled adjacent context approach (Redington et al. 1998), which uses counts of words that occur in a sliding window surrounding the target word. Variants of this approach are often used for similar lexical acquisition tasks such as word sense disambiguation, (e.g., Kaji and Morimoto 2005), and computing them requires little prior knowledge of the language at hand. However, contextually associated words are not always good cues to animacy: *soldier* and *tank* may often co-occur within the same word window, but *soldier* is animate while *tank* is inanimate. On the other hand, *soldier* and *tank* are not equally likely as subjects of a verb like *believe*. These facts suggest that some consideration of the structural relations between words will be useful in discriminating animate from inanimate items.

differs from the animacy of that loan in Japanese. For example, WordNet2.0 lists one sense for the noun *super*, which is animate (superintendent, super), whereas in Japanese *supaa* usually refers to a supermarket. In general, the transferred labels were reliable.

Many verbs are known to impose semantic restrictions on one or both arguments (e.g., the subject and object of *murder* are typically both people), indicating that verbs may work well for animacy classification. Relative to a set of specific syntactic or morphological features, verbs should provide reasonable coverage of the data set. Using verbs as features affords some measure of cross-linguistic consistency, as verb-argument relations are in principle language-independent (with language-specific instantiations). Therefore, we restricted our feature set to verbs governing as subject or object the nouns we wanted to classify. For each noun in our study we created three feature vectors based on the number of its subject occurrences, and three feature vectors based on the number of its object occurrences.

Subject (Object) Frequency: frequency with which a noun occurs as the subject (object) of a verb. Features are individual verbs, and values are counts. Of the three representations, we expect this one to do the best job of preserving distinctions between items. However, most verbs will occur with a small subset of nouns, making generalization relatively difficult.

Verb Animacy Ratio: for each verb in the training set, we calculated its subject animacy ratio as the number of animate subjects divided by the total number of subjects (likewise for the object animacy ratio) and substituted this new value for the original frequency of that verb when it occurred as a feature of a particular noun. These ratios were calculated over the training data, and applied to the features in the test set. Test features that did not occur in the training set were discarded.

The individual verbs are still used as features, but values are animacy ratios rather than counts. We are now paying attention to the general predictive power of the verb, rather than the frequency with which it occurs with a particular noun. We expect that with this representation, some distinctions between items will be lost because any time two items share a feature, they will have the same value for it. but expect generalization to improve, because the feature values are now representative of class-wide tendencies across the training data, and are based on a larger quantity of data.

Average Verb Animacy Ratio: for each noun, we created a vector with a single feature whose value was the average animacy ratios of the verbs that occurred with that noun at least once. This representation essentially eliminates item-level distinctions, but should generalize across classes well even with sparse data. We expect that a single estimate of class-wide tendencies will be robust to some of the variation associated with a large number of infrequent features.

3.3 Classifier

Any of a number of machine learning techniques are suitable to the task of automatic animacy classification given the proposed feature set. We used Bayesian logistic regression (Genkin et al. 2004), a sparse logistic regression classifier that efficiently deals with a large number of features. This model has been shown to work well for other natural language classification tasks including lexical semantic verb classification (Li and

Brew 2008; Li, Baker and Brew 2008), text categorization (Genkin et al. 2004) and author identification (Madigan et al. 2005). Bayesian regression is based on a sparse logistic regression model that uses a prior distribution favoring feature weights of zero, simultaneously selecting features and providing shrinkage. Detailed description of the model is available in Baker (2008). We used a publicly available implementation of the model³ and specified a Laplace prior with mode zero.

4. Experiments

We ran three experiments on animacy classification in English and Japanese. The first experiment establishes baseline classification accuracy using feature vectors based on frequency counts of verb-subject and verb-object relations. The second experiment examines the impact that grouping items into equivalence classes prior to classification has on data coverage and classification accuracy. The third experiment focuses on classifying zero-frequency items by training an English classifier on translations of the Japanese items and transferring those labels back onto the Japanese data set.

4.1 Experiment One

The purpose of this experiment is to classify Japanese nouns as \pm animate, comparing the coverage and classification accuracy of feature vectors containing subject (object) counts, verb animacy ratios, and average verb animacy ratio. The results of the classification are shown in Table 6.

In terms of coverage, object counts accommodate slightly more of our data set (36%) than subject counts (33%). When combined, subject and object counts cover 40% of the data set (meaning that the remaining 60% were not parsed as subject or object of any of the verbs in the feature set⁴). The combined feature set tends to have the highest precision and recall within each feature type, and the best performing combination of feature set and feature type is average verb animacy ratio with combined subject and object.

Feat	Cvg (%)	Acc (%)	Inanimate		Animate	
			Prec (%)	Rec (%)	Prec (%)	Rec (%)
Subject (Object) Frequency						
Subject	33	83.6	84.7	95.3	78.3	49.8
Object	36	85.9	86.2	97.5	83.4	44.8
Subj+Obj	40	85.7	86.5	95.6	82.0	57.1
Verb Animacy Ratio						
Subject	33	83.1	85.9	92.6	71.9	55.6
Object	36	85.9	89.2	93.3	71.2	59.8
Subj+Obj	40	84.9	87.3	93.2	75.7	60.6
Average Verb Animacy Ratio						

³ <http://www.bayesianregression.org/>

⁴ We excluded *iru* and *aru* from the feature set because we used these two verbs to select the data set.

Subject	33	86.8	88.7	94.3	79.7	65.0
Object	36	88.1	89.1	96.6	82.6	57.9
Subj+Obj	40	88.0	89.2	95.4	83.5	66.7

Table 6. Classification results for Japanese.

The most frequent baseline for the covered portion of the data set is about 50%. All of the feature types outperform the baseline by 30-38%. Object counts perform as well as subject counts across the feature types. Overall, precision and recall for the animate nouns ($p=79\%$, $r=58\%$) tends to be considerably lower than for the inanimate nouns ($p=88\%$, $r=95\%$). Precision of the animate class is lowest when using verb animacy ratios as feature values (74% vs. 80%).

4.2 Experiment Two

The purpose of the second experiment is to group nouns into equivalence classes prior to classification and examine the corresponding effect on data coverage and classification accuracy. As mentioned in Experiment 1, the most comprehensive feature set (combined subject and object counts) only covers 40% of our data points. Therefore, we were interested in a way of forming noun classes that does not depend on feature counts.

We realized that many of the items in our data set are morphologically similar to compound nouns. Most compound nouns are subtypes of the head noun (e.g., *sports car* is a type of *car*), and compound nouns with a common head often share properties of the head, including its animacy. For example, in English, compounds such as *postman*, *fireman*, *salesman* are all types of *man*, and for the purposes of gross animacy classification further distinctions are not necessary. Many Japanese nouns are morphologically similar to the compounds in the English example above. In particular, it is common for words of Chinese origin to have a compound-like morphology, and Japanese orthography often makes this structure explicit. For example, a number of words end in the suffix *-jin* 'person' (orthographically the single character 人): *kajin* 'poet', *kyojin* 'giant', *shuujin* 'prisoner', *tatsujin* 'expert', etc. Another class of words ends in the suffix *-hin* 'manufactured good' (orthographically the single character 品): *shinsouhin* 'bedding', *buhin* 'parts', *youhin* 'supplies', *shouhin* 'prize', etc. In both cases, the Japanese morphology and orthography provide a type of surface homogeneity not as readily available in the English equivalents.

We formed suffix classes of Japanese nouns by grouping all the items ending with the same kanji (i.e., the same character such as *-jin*, *-hin*, etc.). Although there are cases where the final character is not acting as the head (e.g., *satsujin* 'murder'), we were reasonably confident in the consistency afforded by this approach and did not try to eliminate such cases. Once the suffix classes were formed, we obtained the subject and object counts for the class. For example, given a suffix class of *-jin*, we incremented feature counts for this class any time *kajin*, *kyojin*, etc. appeared. We then applied this feature vector to each member of the class, so that *kajin*, *kyojin*, etc. have identical feature vectors. As with the average verb animacy ratio, this application of suffix classes

eliminates many item-level distinctions. However, recall and precision for both classes should increase, given much denser feature vectors.

Figure 1 shows the effect of forming suffix classes on the distribution of item frequency in our data set. The dashed line shows the cumulative probability distribution of items based on their subject or object counts before applying suffix classes. As the dashed line in Figure 1 indicates, the data set is initially sparse, with about 77% of the items occurring fewer than 10 times.

The solid line in Figure 1 shows the cumulative probability distribution of item frequency after forming suffix classes. The effect of the suffix classes on the cumulative probability distribution manifests itself in the graph in short regions of steep slope, which correspond to groups of identical feature vectors occurring at a particular frequency. We are able to account for 11% our data set that does not occur with any of our features by virtue of inclusion in a suffix class. Moreover, about 75% of our data set now occurs 250 times or fewer, as opposed to fewer than 10 times, indicating that the mass of the cumulative probability distribution has shifted considerably.

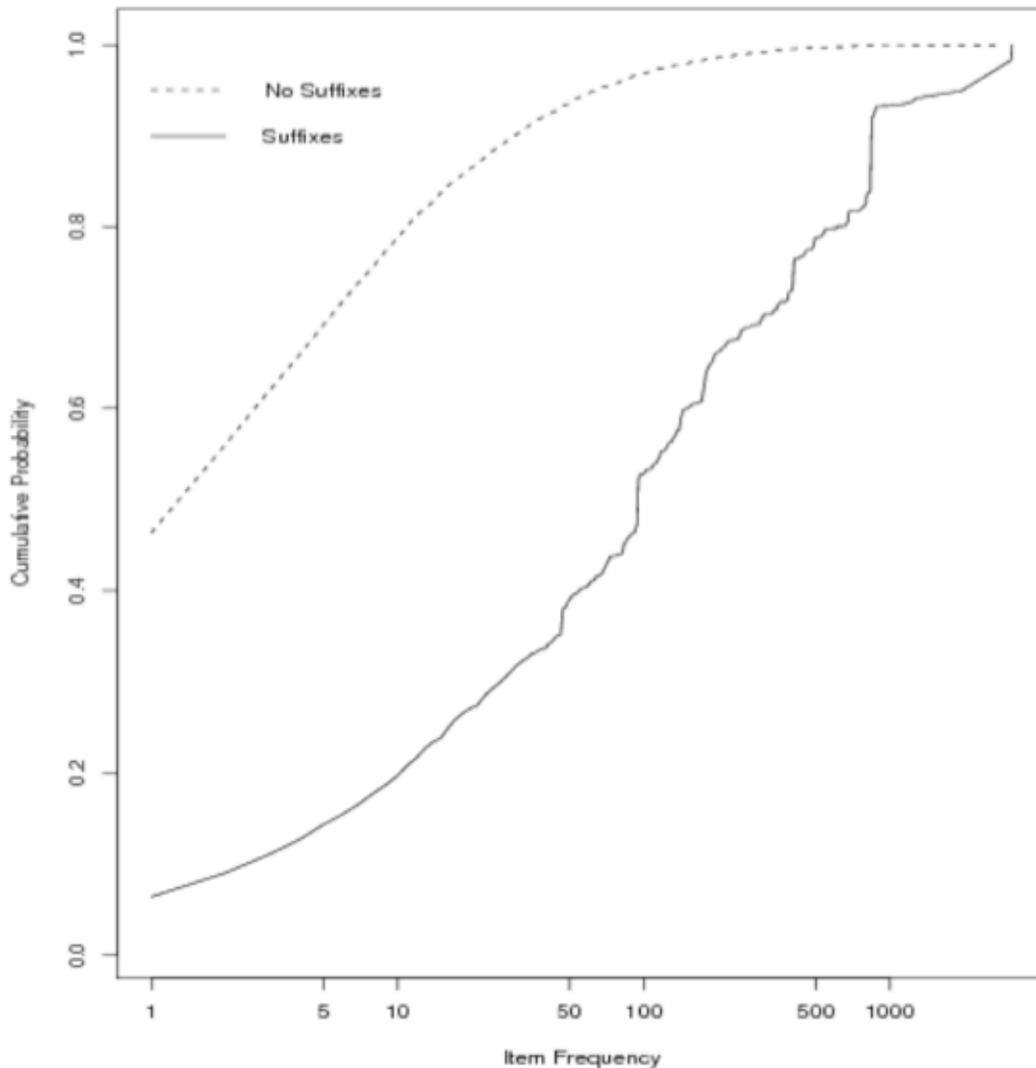


Figure 1. Cumulative probability distribution of item frequency with suffix classes

(dashed) and without (solid).

Table 7 shows the extent to which collapsing individual items into suffix classes reduced the size of the data set.

	Size	w/ suffix class	Reduction
Animate	4433	1892	57%
Inanimate	9956	8613	13%

Table 7. Effect of suffix classes on compressing the data set.

Applying suffix classes to the animate nouns resulted in a 57% reduction in the size of the class; for the inanimate items, we obtained a reduction of only 13%. The main reason for this disparity is the proportion of English origin items in the two classes: 74% of the inanimate nouns were English loanwords (hence, not members of a cohesive suffix class), but only 40% of the animate nouns were English loans.

Table 8 contains the classification results for the second experiment.

Feat	Cvg (%)	Acc (%)	Inanimate		Animate	
			Prec (%)	Rec (%)	Prec (%)	Rec (%)
Subject (Object) Frequency						
Subject	46	93.8	90.2	99.5	99.3	86.0
Object	49	94.3	91.2	99.5	99.3	87.7
Subj+Obj	51	93.6	90.4	99.5	99.3	85.6
Verb Animacy Ratio						
Subject	46	94.7	93.4	97.3	96.5	91.7
Object	49	95.3	94.4	97.5	96.6	92.6
Subj+Obj	51	94.6	93.5	97.2	96.1	91.0
Average Verb Animacy Ratio						
Subject	46	93.1	95.8	91.5	90.2	95.1
Object	49	93.8	95.4	93.4	91.8	94.2
Subj+Obj	51	93.4	95.4	93.0	90.8	94.0

Table 8. Classification results for Japanese using suffix classes.

The most frequent baseline for the covered portion of the data set is 58% (inanimate). Overall classification accuracy increases to above 95%, while coverage increases to 51% of the data set. Precision and recall of the animate class shows the largest improvement from Experiment One (up 16% and 34%, respectively) and is now on par with precision and recall for the inanimate class (93% and 96%, on average).

The effects of the ratios versus counts is visible in this less sparse data set. For the animate class, precision drops from 99% using counts to 96% using verb animacy ratios, and to 91% for the single-feature vector of average verb animacy ratio. Recall increases by a few points across each feature type from 87% (counts) to 94% (average verb animacy ratio). Precision and recall of the inanimate nouns shows the opposite effect: precision increases slightly from 91% (counts) to 96% (average verb animacy ratio) as recall decreases from 100% (counts) to 93% (average verb animacy ratio). We assume that the effect of using ratios is greater for the animate items, mainly because suffix classes resulted in a much greater reduction in the size of the animate noun class. As animate recall increases, inanimate precision increases because there are fewer incorrectly tagged animate items; as the number of incorrect animate predictions increases, recall of the inanimate class decreases.

4.3 Experiment Three

Even after forming suffix classes, we are able to cover only half of the data set. Most of the zero-frequency items are English loanwords. Since we have the English transliteration of each loanword, the purpose of Experiment 3 is to examine the feasibility of transferring animacy distinctions acquired in English onto Japanese data. We do not expect the English classifier to be as reliable as the Japanese one, because English is not particularly noted for robust sensitivity to animacy. However, we do expect performance to be better than chance.

For the English animacy classification, we extracted subject-verb pairs from the English Gigaword corpus (Graff 2003) using MiniPar (Lin 1995), a broad-coverage English dependency parser. Because data sparsity was less of an issue, we restricted our feature set to subject counts of transitive verb instances (i.e., verbs that occurred with a subject and object).

To create our training data, we translated the non-English words in the original data set into English using Babel Fish⁵. This resulted in 3302 training items (many items translated into the same word in English), two thirds of which were inanimate. There were 6917 test items (transliterations of the English loanwords); 5629 (81%) of these were inanimate and 1288 were animate. As with the Japanese suffix classes, we collapsed multi-word compounds into single categories before classification (e.g., *summer camp*, *day camp*, etc. → *camp*).

Table 9 contains the results of the English animacy classification. Overall, subject counts performed the best (88% correct), and precision and recall for the two animacy classes is similar to the results for Japanese using subject counts (Experiment 1, Table 6).

Accuracy	Inanimate		Animate	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
Subject Frequency				
88.0	89.2	97.0	78.7	48.8

⁵ <http://babelfish.altavista.com>

Verb Animacy Ratio

79.7	92.4	81.9	47.0	70.3
------	------	------	------	------

Average Verb Animacy Ratio

81.4	81.4	100	100	0
------	------	-----	-----	---

Table 9. English-only results on the loanwords.

The effect of substituting feature values with verb ratios is even more clearly visible on this less sparse data set (80% of the English items occurred more than 10 times, vs. 40% of the Japanese items with suffix classes applied).

The single feature vector containing a noun's average verb animacy ratio does not work for the English data. The most likely explanation for this fact is illustrated in Figure 2, which contains the distribution of verb animacy ratio for Japanese nouns versus the English test items.

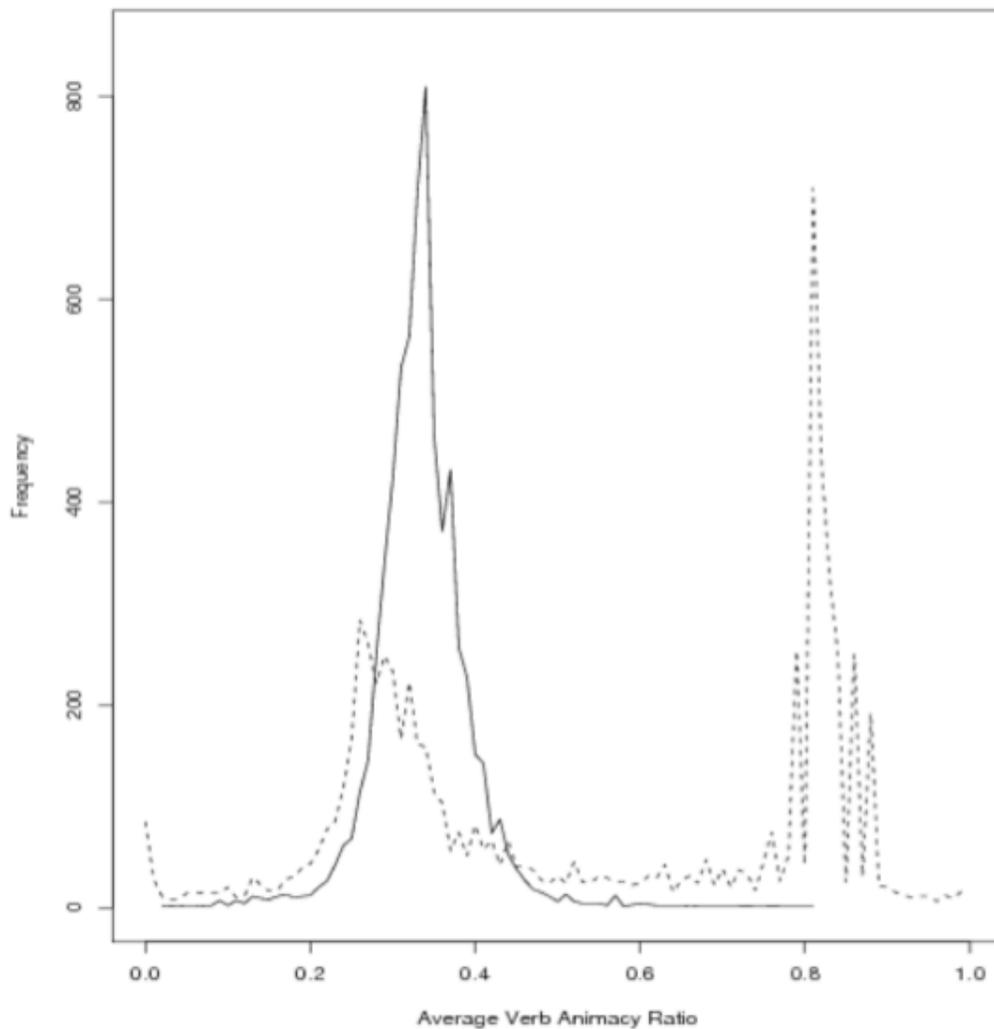


Figure 2. Distribution of Average Verb Animacy Ratio for Japanese nouns (dashed) and English test items (solid).

The English items are unimodally distributed around a mean of 0.36, meaning that most of their features are biased towards inanimates regardless of the item's true class. On the other hand, the Japanese data is bimodally distributed, indicating why this feature is a better predictor for the Japanese data.

Table 10 shows the classification results of the Japanese data with the English labels applied to the loanwords. The most frequent baseline is 70% (inanimate).

Feature	Coverage	Accuracy	Inanimate		Animate	
			Prec (%)	Rec (%)	Prec (%)	Rec (%)
Subject + Object Frequency						
Jp	97	87.6	84.9	99.8	99.1	60.2
Jp + En	97	86.6	91.1	89.3	77.1	80.5
Verb Animacy Ratio						
Jp	97	87.9	86.1	98.4	94.7	64.5
Jp + En	97	86.4	93.6	86.3	73.9	86.7
Average Verb Animacy Ratio						
Jp	97	86.8	87.1	95.0	85.8	68.5
Jp + En	97	85.8	94.4	84.5	71.9	88.7

Table 10. Japanese results with English transfer. Jp=Japanese baseline (inanimate), Jp+En=labels transferred from English.

We used subject and object occurrences for the Japanese feature set, as this performed the best for each feature type. Regardless of the Japanese feature type, the English labels were applied using the results of the subject counts.

With the English transfer, coverage increases to 97% of the data set (the remaining 3% are Japanese nouns that were not parsed as subjects or objects of verbs in the feature set). In every case, we get better inanimate precision and animate recall using transferred labels versus applying the default (inanimate) label to the same data set. Overall accuracy is not different in each feature type, but within-class precision and recall are.

5. Discussion

Overall, classification accuracy was higher for Japanese than English using comparable feature sets. In particular, classifying animate items is more reliable for Japanese (precision ≈ 96 , recall ≈ 91) than English (precision ≈ 79 , recall ≈ 49). This disparity may result from noise arising from the language transfer or differences in how the two languages lexicalize animacy.

In general, the reliability of the transferred animacy labels seems to be reasonable: 99% of the English translations that appear in WordNet2.0 of the Japanese inanimate nouns (2293) have only inanimate senses, and 97% of the English translations that appear in WordNet2.0 of the Japanese animate nouns (1008) are listed as unambiguously animate. This fact suggests that the difficulty lies in English verbs' relative lack of sensitivity to the animacy of their subjects. Figure 3 compares the distribution of verb animacy ratios for English and Japanese. An animacy ratio of 0.0 means that verb occurred exclusively with inanimate subjects, and an animacy ratio of 1.0 means that verb appeared only with animate subjects.

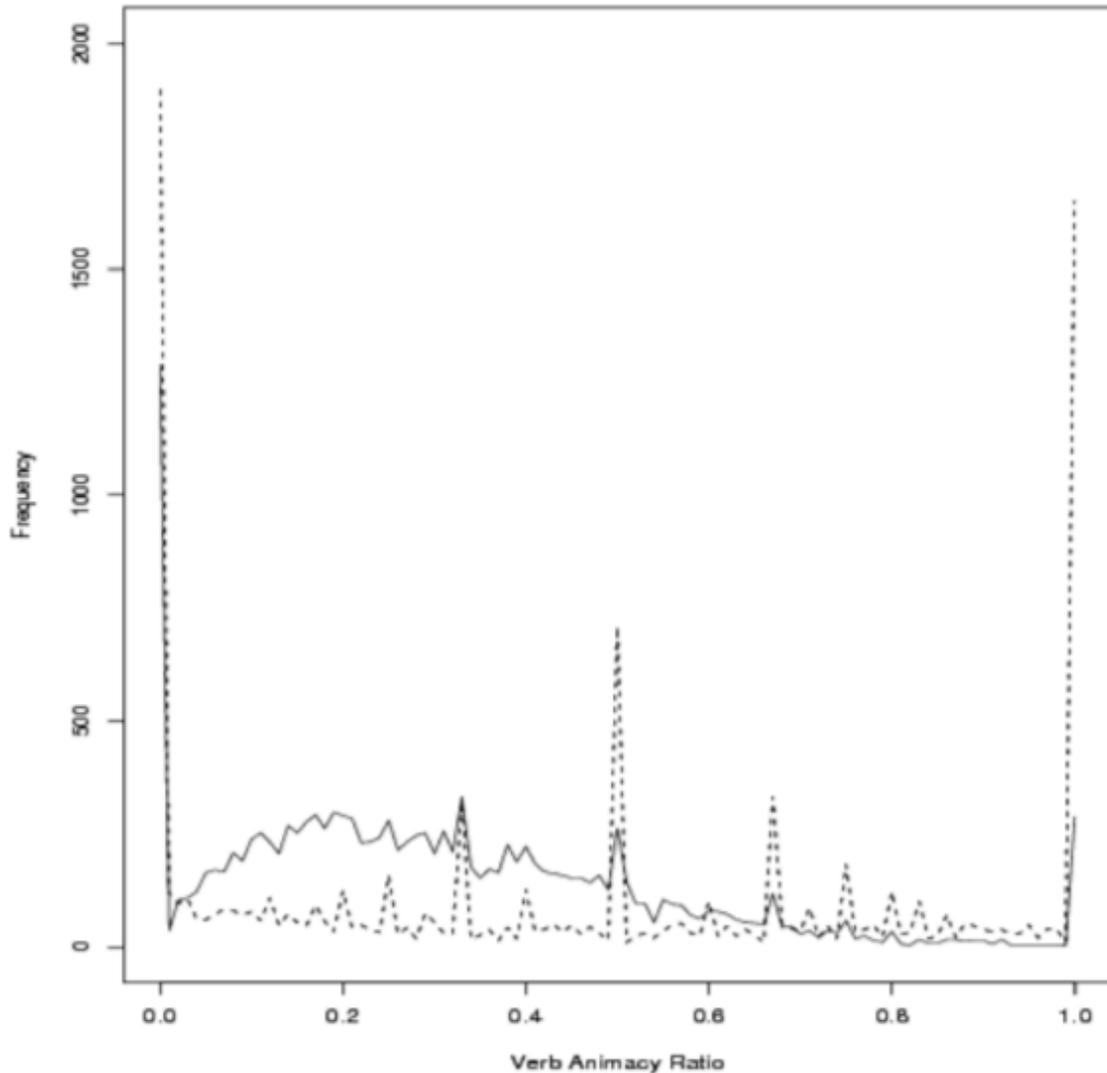


Figure 3. Distribution of Verb Animacy Ratios for Japanese (dashed) and English (solid) verbs.

Both languages exhibit peaks at the extremes of the scale, and for both languages the number of exclusively animate verbs is less than the number of exclusively inanimate verbs. In English, however, most verbs are biased towards inanimate subjects, with most of the frequency mass between 0.0-0.7. Japanese exhibits a third peak at 0.5, but most of the frequency mass is concentrated at the extrema. The distributions in Figure 3 indicate

that the Japanese feature set better partitions the data into the two animacy classes than the English feature set does. This fact calls into question the general cross-language applicability of relying solely on verbs as features: the criterion of good coverage seems to hold, but the animacy distribution afforded by the English feature set appears too weak for reliable classification.

Like Orăsan and Evans (2001), we found that animate items are harder to classify than inanimate ones, and offer an explanation for this phenomenon on the basis of the distribution of items in our data set. Intuitively, animate things are capable of a wider range of different actions than inanimate things are. This difference may be reflected in language data as a disparity between the number and dispersion of verbs associated with the two animacy classes.

Examination of the data set shows that across the two languages, there are approximately three times as many animate features as animate items, whereas the number of inanimate features is roughly equal to the number of inanimate items. For both languages, animate subjects are associated with a larger number of verbs than inanimate subjects are. Conversely, each verb is associated with fewer animate subjects than inanimate subjects. Animate nouns may be harder to classify because each item occurs with a larger set of different features, making each animate feature vector relatively unlike any other.

6 Conclusion and Future Work

This paper presented the results of three experiments on automatic animacy classification in Japanese and English in which we focused on classifying infrequent items. Animacy classification for Japanese was more reliable than for English, largely because English verbs are less sensitive to the animacy of their arguments. Replacing feature counts with verb animacy ratios resulted in improved classification accuracy for the harder-to-classify animate items. The biggest gains in classification accuracy resulted from placing items into suffix-based equivalence classes prior to classification. We further demonstrated the feasibility of language transfer from English to Japanese using loanwords as conduits for lexical semantic annotation. By exploiting lexical surface cues to animacy in Japanese that are not available in English, we were able to create a training set for an English classifier and transfer the acquired labels back onto the loanwords in Japanese.

Future work will look at aspects of multilingual lexical acquisition touched on in this paper; in particular, it will focus on exploiting the robust animacy lexicalization in Japanese for making improved animacy distinctions in English. We will examine the feasibility of classifying the relatively small set of frequent English loanwords using Japanese corpus data, and extending those labels to a larger set of English words via a semi-supervised learning technique such as manifold regularization (e.g., Belkin et al. 2004). Using loanwords is appealing for this task because their transliteration can be automated (e.g., Knight and Graehl 1998) lessening the dependence on external lexical resources.

Acknowledgments

We are indebted to James Unger, Hiroko Morioka, Brian Joseph, Peter Culicover and Adriane Boyd for their help and suggestions on this paper. All errors are ours alone. Funded by NSF CAREER Grant 0347799 to Chris Brew.

References

- Baker, Kirk. 2008. Multilingual distributional lexical acquisition. PhD Dissertation. The Ohio State University.
- Belkin, Mikhail, Partha Niyogi and Vikas Sindhwani. 2004. Manifold regularization: a geometric framework for learning from examples. *University of Chicago CS Technical Report TR-2004-06*.
- Bliss, Heather 2005. Topic, focus, and point of view in Blackfoot. In John Alderete (ed.), *Proceedings of the 24th West Coast Conference on Formal Linguistics*, 61-69. Cascadilla Proceedings Project. Somerville, MA.
- Bresnan, Joan, Anna Cueni, Tatiana Nikitina and Harald Baayen. 2005. Predicting the dative alternation. In *Proceedings of Royal Netherlands Academy of Science Workshop on Foundations of Interpretation*.
- Bresnan, Joan and Jennifer Hay. 2006. Gradient grammar: an effect of animacy on the syntax of give in varieties of English. Electronic manuscript. www.stanford.edu/~bresnan/anim-spokenyntax-final.pdf.
- Daelemans, Walter, Jakub Zavrel, Ko van der Sloot and Antal van den Bosch. 2003. TiMBL: Tilburg Memory-Based Learner version 5.0 Reference Guide. *ILK Technical Report – ILK 03-10*.
- Dahl, Östen and Kari Fraurud. 1996. Animacy in grammar and discourse. In Thorstein Fretheim and Jeanette K. Gundel, eds. *Reference and Referent Accessibility*, pp 47-64. John Benjamins.
- Dowty, David. 1991. Thematic proto-roles and argument selection. *Language* 67 (3): 547-619.
- Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.
- Gawronska, Barbara, Björn Erlendsson and Hanna Duczak. 2002. Extracting semantic classes and morphosyntactic features for English-Polish machine translation. In *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machines Translation*.
- Genkin, Alexander, David D. Lewis and David Madigan. 2004. Large-scale Bayesian logistic regression for text categorization. *DIMACS Technical Report*.
- Graff, David. 2003. English gigaword. *Linguistic Data Consortium*, Philadelphia. LDC2003T05.
- Graff David and Zhibiao Wu, 1995. Japanese business news text. *Linguistic Data Consortium, Philadelphia*. LDC95T8.
- de Ilarraza, Arantza Díaz, Aingeru Mayor and Kepa Sarasola. 2002. Semiautomatic Labelling of Semantic Features. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*.
- Iida Asako. 1999. A descriptive study of Japanese major classifiers. Electronic manuscript. <http://www5b.biglobe.ne.jp/~aiida/ephd.html>.

- Jäger, Gerhard and Anette Rosenbach. 2006. The winner takes it all – almost: Cumulativity in grammatical variation. *Linguistics* 44(5): 937-972.
- Joseph, Brian. 1979. On the animate-inanimate distinction in Cree. *Anthropological Linguistics* 21(7): 351-354.
- Kaji, Hiroyuki and Yasutsugu Morimoto. 2005. Unsupervised word sense disambiguation using bilingual comparable corpora. *IEICE Transactions on Information and Systems* E88 D (2): 289-301.
- Knight Kevin and Jonathan Graehl. 1998. Machine transliteration. *Computational Linguistics* 24: 599-612.
- Kudo, Taku and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proceedings of the CoNLL-2002*, 63-69.
- Li, Jianguo, Kirk Baker and Chris Brew. 2008. A corpus study of Levin's verb classification. *American Association of Corpus Linguistics (AACL-2008)*. Provo, Utah.
- Li, Jianguo and Chris Brew. 2008. Which are the best features for automatic verb classification? In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*. Columbus, OH.
- Lin, Dekang. 1995. A dependency-based method for evaluating broad-coverage parsers. In *Proceedings of the 1995 International Joint Conference on Artificial Intelligence (IJCAI)*.
- Madigan, David, Alexander Genkin, David D. Lewis and Dmitriy Fradkin. 2005. Bayesian multinomial logistic regression for author identification. *DIMACS Technical Report*.
- Orăsan, Constantin and Richard Evans. 2001. Learning to identify animate references. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL-2001)*, 129-136.
- Øvrelid, Lilja. 2006. Towards robust animacy classification using morphosyntactic distributional features. In *Proceedings of EACL 2006 Student Research Workshop*. Trento, Italy.
- Palmer, Martha, Dan Gildea and Paul Kingsbury. 2005. The proposition bank: a corpus annotated with semantic roles. *Computational Linguistics* 31 (1): 71-106.
- Redington, Martin, Nick Chater and Steve Finch. 1998. Distributional information: a powerful cue for acquiring syntactic categories. *Cognitive Science* 22: 425-469.
- van Nice, Kathy Y. and Rainer Dietrich. 2003. Task-sensitivity of animacy effects: evidence from German picture descriptions. *Linguistics* 5: 825-849.
- Woolford, Ellen. 1999. Animacy hierarchy effects on object agreement. In Paul F. A. Kotey (ed.), *New Dimensions in African Linguistics and Languages*, 203-216. Africa World Press, Inc.
- Yamamoto, Mutsumi. 1999. Animacy and Reference: *A Cognitive Approach to Corpus Linguistics*. John Benjamins Publishing, Cambridge, MA.
- Zaenen, Annie, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M. Catherine O'Connor and Thomas Wasow. 2004. Animacy encoding in English: why and how. In *Proceedings of the ACL Workshop on Discourse Annotation*, 118-125.