

# Analogical classification in formal grammar

Matías Guzmán Naranjo

Draft  
of May 23, 2019, 12:08

# Empirically Oriented Theoretical Morphology and Syntax

Chief Editor: Stefan Müller

Consulting Editors: Berthold Crysmann, Laura Kallmeyer

In this series:

1. Lichte, Timm. Syntax und Valenz: Zur Modellierung kohärenter und elliptischer Strukturen mit Baumadjunktionsgrammatiken
2. Bîlbîie, Gabriela. Grammaire des constructions elliptiques: Une étude comparative des phrases sans verbe en roumain et en français
3. Bowerman, Claire, Laurence Horn & Raffaella Zanuttini (eds.). On looking into words (and beyond): Structures, Relations, Analyses

ISSN: 2366-3529

# Analogical classification in formal grammar

Matías Guzmán Naranjo


Guzmán Naranjo, Matías. 2019. *Analogical classification in formal grammar* (Empirically Oriented Theoretical Morphology and Syntax 5). Berlin: Language Science Press.

This title can be downloaded at:

<http://langsci-press.org/catalog/book/186>

© 2019, Matías Guzmán Naranjo

Published under the Creative Commons Attribution 4.0 Licence (CC BY 4.0):

<http://creativecommons.org/licenses/by/4.0/>  <http://langsci-press.org/graphics/ccby.eps>

ISBN: 978-3-96110-186-3 (Digital)

978-3-96110-187-0 (Hardcover)

ISSN: 2366-3529

DOI: [10.5281/zenodo.3191825](https://doi.org/10.5281/zenodo.3191825)

Source code available from [www.github.com/langsci/186](https://www.github.com/langsci/186)

Collaborative reading: [paperhive.org/documents/remote?type=langsci&id=186](https://paperhive.org/documents/remote?type=langsci&id=186)

Cover and concept of design: Ulrike Harbort

Fonts: Linux Libertine, Libertinus Math, Arimo, DejaVu Sans Mono

Typesetting software: Xe<sub>La</sub>T<sub>E</sub>X

Language Science Press

Unter den Linden 6

10099 Berlin, Germany

[langsci-press.org](http://langsci-press.org)

Storage and cataloguing done by FU Berlin

 <http://langsci-press.org/graphics/storage/logo.pdf>

# Contents

<b>Acknowledgments</b>	<b>v</b>
<b>Abbreviations</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Remarks on analogy</b>	<b>3</b>
2.1 The many meanings of analogy . . . . .	3
2.1.1 Single case analogy . . . . .	3
2.1.2 Proportional analogies . . . . .	6
2.1.3 Analogical classifiers . . . . .	9
2.1.4 Summing up . . . . .	14
2.2 The mechanism for analogy . . . . .	14
2.2.1 Simple rules . . . . .	15
2.2.2 Schemata . . . . .	18
2.2.3 Multiple-rule systems . . . . .	20
2.2.4 Neural networks and analogical modelling . . . . .	23
2.2.5 Analogy or rules . . . . .	25
2.2.6 Mental representations vs grammatical relations . . . . .	28
2.2.7 Summing up . . . . .	29
2.3 Missing pieces . . . . .	30
2.4 Final considerations . . . . .	33
<b>3 Modelling analogy in grammar</b>	<b>35</b>
3.1 Basic assumptions . . . . .	35
3.1.1 Feature structures . . . . .	35
3.1.2 Type hierarchies . . . . .	36
3.2 Analogy as type constraints . . . . .	40
3.2.1 Analogy is categorical . . . . .	42
3.2.2 Analogy runs through the hierarchy . . . . .	43
3.3 The (semi-)formal model . . . . .	45
3.4 Final remarks . . . . .	51

<b>4</b>	<b>Methodological notes</b>	<b>53</b>
4.1	On the general methodology . . . . .	53
4.2	Statistical models and methodology . . . . .	54
4.3	Analogical models using neural networks . . . . .	57
4.4	Measuring variable importance . . . . .	67
4.5	Clustering and distances between classes . . . . .	68
4.6	Summing up . . . . .	72
<b>5</b>	<b>Gender systems</b>	<b>73</b>
5.1	Masculine-feminine syncretism: Latin . . . . .	75
5.1.1	The Latin third declension . . . . .	75
5.1.2	Data . . . . .	76
5.1.3	Results . . . . .	77
5.2	Gender vs inflection class: Romanian . . . . .	79
5.2.1	The Romanian gender and plural system . . . . .	79
5.2.2	Modelling the system . . . . .	85
5.2.3	Data . . . . .	91
5.2.4	Results . . . . .	93
5.3	Interim conclusion . . . . .	103
<b>6</b>	<b>Hybrid classes</b>	<b>105</b>
6.1	Overabundant inflection: Croatian singular instrumental . . . . .	105
6.1.1	Modelling the system . . . . .	107
6.1.2	Materials . . . . .	108
6.1.3	Results . . . . .	108
6.2	Frequency and analogical similarity: Russian diminutives . . . . .	111
6.2.1	Russian diminutives . . . . .	111
6.2.2	Modelling the system . . . . .	114
6.2.3	The dataset . . . . .	115
6.2.4	Results . . . . .	116
6.3	Interim conclusion . . . . .	117
<b>7</b>	<b>Morphological processes and analogy</b>	<b>119</b>
7.1	Prefixes and gender: Swahili noun classes . . . . .	119
7.1.1	Materials . . . . .	122
7.1.2	Results . . . . .	123
7.2	Prefixes and inflection classes: Eastern Highland Otomi . . . . .	127
7.2.1	Verb classes in Eastern Highland Otomi . . . . .	127
7.2.2	Materials . . . . .	128

7.2.3	Results	128
7.3	Stem changing processes: Hausa plural classes	130
7.3.1	The Hausa plural system	130
7.3.2	Materials	131
7.3.3	Results	132
7.4	Interim conclusion	135
<b>8</b>	<b>Complex inflectional classes</b>	<b>139</b>
8.1	Multiple inheritance and cross-hierarchies: Spanish verbal inflection	139
8.1.1	Spanish inflection classes	139
8.1.2	Previous takes on the Spanish verbal system	143
8.1.3	Modelling the system	146
8.1.4	Materials	152
8.1.5	Results	153
8.2	Cross-classifications between plural and singular: Kasem	164
8.2.1	ATR in Kasem	167
8.2.2	A simple analysis of Kasem noun classes	168
8.2.3	Materials	180
8.2.4	Modelling the system	181
8.2.5	Methodological considerations	186
8.2.6	Results	189
8.3	Interim Conclusion	206
<b>9</b>	<b>Concluding remarks</b>	<b>209</b>
9.1	The path forward	209
9.1.1	The limits of analogy	209
9.1.2	Analogical classifiers or proportional analogies	209
9.1.3	The features of analogy	210
9.1.4	Coverage	210
9.2	Final considerations	211
	<b>References</b>	<b>213</b>
	<b>Index</b>	<b>235</b>
	Name index	235
	Language index	235
	Subject index	237







## 8 Complex inflectional classes

So far I have only looked at systems with relatively few classes, and hierarchies with few types. This chapter looks at three examples where the systems are considerably larger, with many more inflection classes, and which require more complex type hierarchies. The main question here is what happens with the analogical relations, particularly the analogical similarities between classes, when the type hierarchies are made up of several interacting sub-trees.

### 8.1 Multiple inheritance and cross-hierarchies: Spanish verbal inflection

#### 8.1.1 Spanish inflection classes

In Spanish there are three clear inflectional classes given by the thematic vowel of the verb: *-a(r)* (e.g. *cantar*, ‘to sing’), *-e(r)* (e.g. *correr* ‘to run’) and *-i(r)* (e.g. *reír* ‘to laugh’), also referred to as first class, second class and third class, respectively. Depending on the variety, inflectional paradigms in Spanish consist of around 53 content cells, exemplified in Table 8.1 for *amar* (‘to love’). The 2PL forms given in Table 8.1 are only found in Spain, with Latin American Spanish using the 3PL form for the 2PL. Additionally, the future subjunctive is rare, and it is found mostly in fixed expressions: *sea lo que fuere* (‘whatever it may be’)<sup>1</sup>. Finally, the imperfect subjunctive exhibits overabundance (Thornton 2010a,b) between *-se* and *-ra*, with both forms having exactly the same morphosyntactic content (Cuervo & Ahumada n.d.; DeMello 1993; Kempas 2011; Rojo 2008; Schwenter 2013).

The literature recognizes two macroclasses in the inflectional system of Spanish based on their thematic vowel: verbs ending in *-ar* vs. verbs ending in *-er* or *-ir* (Aguirre & Dressler 2008 among many others). This distinction is easy to see from the partial inflectional paradigm of regular verbs in Table 8.2. The second and third person singular and the third person plural exponents of the second

---

<sup>1</sup>Notice however that it is easy to find uses of this form online: *Demos la vida si fuere necesario* (‘let us give our lives if it should be needed’). <http://portaluz.org/demos-la-vida-si-fuere-necesario-1570.htm>, consulted 12.11.2016.

Table 8.1: Complete paradigm for *amar* ('love'). Tense names are in bold.

Indicative				
	<b>Present</b>	<b>Imperfect</b>	<b>Preterite</b>	<b>Future</b>
1sg	amo	amaba	amé	amaré
2sg	amas	amabas	amaste	amarás
3sg	ama	amaba	amó	amará
1pl	amamos	amábamos	amamos	amaremos
2pl	amáis	amabais	amásteis	amaréis
3pl	aman	amaban	amaron	amarán
Conditional				
1sg	amaría			
2sg	amarías			
3sg	amaría			
1pl	amaríamos			
2pl	amaríais			
3pl	amarían			
Subjunctive				
	<b>Present</b>	<b>Imperfect</b>	<b>Preterite</b>	<b>Future</b>
1sg	ame	ama(se/ra)		amare
2sg	ames	ama(se/ra)s		amares
3sg	ame	ama(se/ra)		amare
1pl	amemos	amá(se/ra)mos		amáremos
2pl	ameis	ama(se/ra)is		amareis
3pl	amen	ama(se/ra)n		amaren
Imperative				
2sg	ama			
2pl	amad			
	<b>Infinitive</b>	<b>Gerund</b>	<b>Participle</b>	
	amar	amando	amado	

and third classes are the same, while these forms are different for the first class. The three classes are only clearly distinguished in the first and second person plural. There are no shared exponents between class 1 and one of the other two classes to the exclusion of the remaining class.

Table 8.2: Simple present paradigm of Spanish regular verbs.

Person/Number	cant-ar ('sing')	corr-er ('run')	aburr-ir ('bore')
1sg	cant- <b>o</b>	corr- <b>o</b>	aburr- <b>o</b>
2sg	cant- <b>as</b>	corr- <b>es</b>	aburr- <b>es</b>
3sg	cant- <b>a</b>	corr- <b>e</b>	aburr- <b>e</b>
1pl	cant- <b>amos</b>	corr- <b>emos</b>	aburr- <b>imos</b>
2pl	cant- <b>áis</b>	corré- <b>is</b>	aburr- <b>ís</b>
3pl	cant- <b>an</b>	corr- <b>en</b>	aburr- <b>en</b>
participle	cant- <b>ado</b>	corr- <b>ido</b>	aburr- <b>ido</b>
gerund	cant- <b>ando</b>	corr- <b>iendo</b>	aburr- <b>iendo</b>

Some alternative descriptions of the Spanish system have been proposed before. Boyé & Cabredo Hofherr (2006) suggest that thematic vowels seem to be a property of stems rather than verbs themselves. The authors base this claim on the fact that some irregular verbs show signs of having a different thematic vowel in certain stems: *andar* ('go, walk') - *anduve* (1sg preterite) and *anduviste* (2sg preterite). This might very well be the case, but it is a very rare phenomenon in Spanish, and it is currently eroding for *andar*, with speakers using the more regular forms: *andé* and *andaste*. I will exclusively focus on the infinitive stem of the verb, and its changes for the present singular, past participle and gerund. For these cells, even a verb like *andar* uses the same stem: *ando*, *andado*, *andando*, respectively.

A related criticism of the traditional analysis in three inflection classes comes from Boyé & Cabredo Hofherr (2004), who suggest that there are in fact only two inflection classes -*ar* and -*er/ir*: "the analysis of verbal morphology of Spanish should only recognize two groups of verbs, namely Conjugation 1, on the one hand, and Conjugation 2/3 on the other" (p. 29), as "the choice between the infinitive endings -*er* and -*ir* (and consequently membership of Conjugations 2 or 3) is predictable on the basis of the prethematic vowel" (p. 20) (Boyé & Cabredo Hofherr 2010, as cited by Roca 2010: 418). However, as Roca (2010: 418) shows, there are problematic cases like *competir* ('concern') - *competir* ('compete') which cannot be distinguished just by the prethematic vowel.

## 8 Complex inflectional classes

I will keep the traditional view of the Spanish system of having thematic vowels being a property of lexemes, and three main inflection classes based on said thematic vowels.

It should be clear, however, that three classes are insufficient to fully describe the inflectional behaviour of Spanish verbs. The main reason is that many verbs exhibit semi-regular conjugation patterns (some authors classify all these patterns under the umbrella of *irregular* Brovetto & Ullman 2005, but this kind of approach completely ignores that there are partial regularities within the different inflectional patterns Maiden 2001; 2005). The main process responsible for the minor conjugation patterns is diphthongization, but there are other stem changing processes. A few examples of different patterns found in the Spanish verbal paradigm are presented in Table 8.3.<sup>2</sup> In this table shows how the three principal parts of the Spanish system (first person singular, past participle and gerund) are taken to determine the inflection class of all regular and semi-regular verbs (cases like *ir* ‘to go’ are completely irregular and their inflection cannot be determined by their principal parts).

Table 8.3: Minor conjugation patterns of Spanish verbs.

verb	gloss	pattern	1sg	participle	gerund
escribir	write	/b~t/	escribo	escrito	escribiendo
elegir	choose	/e~i/	elijo	elejido	eligiendo
controvertir	controvert	/e~je/	controvierto	controvertido	controvirtiendo
descomponer	decompose	/g/	descompongo	descompuesto	descomponiendo
contraer	contract	/ig/	contraigo	contraído	contrayendo
adquirir	acquire	/i~je/	adquiero	adquirido	adquiriendo
fluir	flow	/j/	fluyo	fluído	fluyendo
aprobar	approve	/o~we/	apruebo	aprobado	aprobando
jugar	play	/u~ue/	juego	jugando	jugado
humedecer	humidify	/θ~θk/	humedezco	humedecido	humedeciendo

There are three macroclasses of verbal inflection: *ar*, *ir* and *er*, responsible for the inflectional endings, and multiple other minor (stem) patterns responsible for stem alternations in certain cells of the paradigm. The exact number of classes depends on how one classifies them and groups them. Mateo & Sastre (1995) find around 90 classes, but many of these are verb-specific. I take a more conservative approach where I only take into account classes with more than one lexeme.

<sup>2</sup>Notice that the actual realization of *j* depends on the dialect. Also, in American Spanish, the /θ/ would be an /s/, but the pattern remains the same.

Although different partitions of the stem patterns are possible, I will focus exclusively on those shown in Table 8.3.

An important point here is that many of the stem alternation classes in Table 8.3 can also apply to nouns and adjectives: *cuento* ('tale'), *vejez* ('old age'), *viejo* ('old'), *poblado* ('populated'), *población* ('population'), *pueblo* ('town'). Although I will only focus on verbs, the same hierarchy could be used for modelling stem alternations in nouns and adjectives. This is further evidence for the independence of thematic vowels from stem alternations.

### 8.1.2 Previous takes on the Spanish verbal system

Some older studies on the phenomenon of Spanish verbal inflectional classes considered the stem patterns to be the product of a sort of irregular or non-systematic inflection triggered by diacritics/features on the relevant verbs (Foley 1965; Brame & Bordelois 1973; Harris 1969; 1978) or by complex representations of the lexical entries which include the possible alternants a verb can exhibit (Hooper 1976). These analyses are phonological in nature, and assume a homogeneous morphological system. Brame & Bordelois (1973: 43) also claim that "it is impossible to predict whether any of these segments will alternate or not" and thus suggest hard-wiring whether a noun or verb will alternate or not.

Some recent approaches from a DM perspective (Arregi 2000), and an autosegmental OT perspective (Roca 2010), seem to make the same assumption that "[c]onjugation class membership is unpredictable" (Roca 2010: 412). Similarly, Bermúdez-Otero (2013: 3), talking about diphthongization in verbs, nouns and adjectives also claims that "[t]he choice of theme vowels in Spanish nouns and adjectives can be predicted neither from the phonological shape of roots nor from syntactic features like gender". He concludes that verbs are stored with their thematic vowel instead of having additional inflectional information.

Spanish verbal inflection has also been used in the debate between a dual and single route approach to morphological processing and acquisition (Brovetto & Ullman 2005; Clahsen et al. 2002; Costanzo 2011; Eddington 2009; Yaden 2003), language change (Galván Torres 2007; Wanner 2006), as well as to test different computational models of analogy (Albright 2009). Most of these studies focus on the nature of psycholinguistic processing and mental representations, but I will not focus on these issues (for a detailed review of the literature on the topic of mental representation of Spanish verbal inflection, see Eddington 2009).

There are multiple accounts of the diphthongization processes as shown Table 8.3 from a synchronic (Bellido 1986; Carreira 1991; Harris 1985; Kikuchi 1997) and diachronic (Wilkinson 1971) perspective, but these deal almost exclusively

with the phonological process itself, and do not actually discuss which verbs undergo the diphthongization process. Additionally, most of these accounts focus on the vocalic changes and ignore consonant alternations. Regarding possible regularities that might predict these patterns, Roca (2010: 423) claims that:

[...] contemporarily, diphthongization is lexically conditioned, non-diphthongising *e*, *o* being plentiful: cf. *vejár* ~ *vejo* ‘to ~ I slight’, *podar* ~ *podo* ‘to ~ I prune’, etc. Albright et al. (2001) report a number of frequency effects associated with contextual segmental correlations, but minimal pairs like *muelo* ‘I grind’ vs. *molo* ‘I am/look cool’, respectively from *moler*, *molar*, or *puedo* ‘I can’ vs. *podo* ‘I prune’, from *poder*, *podar*, confirm the unpredictability of lexical incidence. Note that conjugation class is also irrelevant: *vuelo* ‘I fly’, *ruedo* ‘I roll’, from 1<sup>st</sup> conj *volar*, *rodar*.

But the author confuses two things in a slightly disingenuous way. First, the minimal pairs for *podar* ~ *poder* and *molar* ~ *moler* look alike in their stem but belong to two different classes, while *volar* and *rodar* belong to the same class but do not look alike. The first example shows that major inflection class membership is not fully determined by the shape of the stem, but does not show that diphthongization is not predictable within classes.

In a similar vein, Harris (1985: 32) claims that:

[a]s has long been recognised [...] segmental phonological and morphological conditions do not suffice to predict the occurrence or non-occurrence of diphthongization. It follows that some otherwise unmotivated property of the representation - i.e. a lexical diacritic - must be employed to distinguish the alternating from the non-alternating cases, regardless of whether vowels or diphthongs are taken to underlie the alternation

However, Harris fails to provide any kind of evidence for the unpredictability of diphthongization.

A study by Eddington (1996) deals with the degrees to which different derivational processes make use of these diphthongs, but the author also claims that “of course, since not all mid-vowels are subject to diphthongization, those which are must be so designated by means of a diacritic or some other formal entity” (p. 9).

The first hints at an analogical relation holding between these stem alternation patterns, and specifically the diphthongization, was reported by Malkiel (1966). The author noted that *ie* tends to be changed to *i* in the presence of an *s* combined with an *r* or *v*. Malkiel does not present a full analogical model for all

conjugation patterns, though. A more elaborate model was proposed by Boyé & Cabredo Hofherr (2004), who observe that the thematic vowel and vocalic alternation of the stem is predictable, to some degree, from the prethematic vowel. The authors do not, however, provide a full model capable of accurately predicting inflection class. In their conclusion, they claim that the difference between *-ir* and *-er* is due to vocalic harmony, and both suffixes are really allomorphs of the same subjacent morpheme (p. 259).

The main work that deals with analogy in the Spanish system comes from studies by Albright (Albright et al. 2001; Albright 2008b; 2009). Albright (2008b) shows that *-er* verbs have no high vowels in their stem, and verbs in *ir* tend not to have the vowel /o/. He also shows that the rates of the types of vocalic changes are heavily conditioned by the main inflection class. But Albright (2008b: 3) still claims that “the choice of diphthongization vs. raising is not predictable”. One important point Albright (2008b) makes is that speakers seem to keep generalizations about verbs with regards to stem patterns internal to their main inflection class. That is, an *-ar* verb will not analogize to *-ir* or *-er* verbs. I will test this conclusion with the models below.

The most recent work on analogy in Spanish inflection is presented by Albright (2009). In this paper the author shows how a minimal generalization learner (Albright & Hayes 2002) can predict whether a verbal stem in Spanish would undergo diphthongization or not. As it was described before, a minimal generalization learner finds regular and semi-regular patterns, similar to schemas, that predict class membership, and weights them according to how frequent and how general or how specific these patterns are.

One of the main claims by Albright (2009) is that structural analogy is more predictive than pure surface similarity. This claim is tested against psycholinguistics data. Albright et al. (2001) tested 96 native Spanish speakers on new possible verbs (wugs) to see the rate of diphthongization these would have. Speakers were asked to produce the inflected forms of 33 wug items containing a mid vowel (e.g. *lerrar*). The analogical model proposed by Albright (2009) reached a correlation coefficient,  $r$ , of 0.77 when compared to experimental data. Additionally, Albright (2009) tested a less structured model, one that only takes into account surface similarity without structural similarity. The unstructured model reached an  $r$  of 0.56, clearly showing that the minimal generalization learner has better performance when predicting speaker’s behaviour. However, Albright (2009) only tackles the binary distinction: diphthong vs no diphthong. There is no attempt at modelling all inflectional patterns, or a significant subset of these. There are no previous attempts at modelling the full Spanish inflectional system with analogy.



### 8.1.3 Modelling the system

We need a way of classifying and relating stems to major inflection patterns for Spanish verbs. A simple alternative to capture the fact that the *er* and *ir* classes behave as a single class in opposition to the *ar* class, is with a hierarchy as in Figure 8.1.

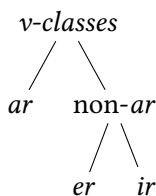


Figure 8.1: Basic hierarchy for Spanish theme vowels.

But this model is insufficient if we also want to capture semi-regular patterns. Table 8.4<sup>3</sup> presents the cross-tabulated distribution of stem and major patterns, as they appear in a list of around 3000 Spanish verbs (see below for the data description). From this table it is clear that there is no obvious systematicity to the patterns<sup>4</sup>. To which patterns a verb belongs has to be specified independently.

Boyé & Cabredo Hofherr (2006) suggest that the analysis of verbal inflection in Spanish should make use of the stem space (Bonami & Boyé 2003), that is, a list of stems that cover all cells in the paradigm of a verb: “lexemes should rather be associated with a vector of possibly different phonological representations” (Bonami & Boyé 2006). This stem space partitions the paradigm in a regular way, and it is a morphomic property of the paradigm. Boyé & Cabredo Hofherr (2006) show, as Maiden (2001) before, how certain tenses, with no apparent semantic connection, use the same stem (the authors identify eleven stems in total, p. 6). This proposal makes sense for the system. These patterns only affect the stems, and are independent of the thematic vowel of the verb. The implication is then that there is an independent hierarchy which captures the stem alternation system.

There are many ways to capture the patterns in Table 8.3, especially because this is not a complete list. Depending on what one considers to be an inflectional

<sup>3</sup>I use the letter *L* to indicate the *j* class, and the letter *z* to mark the /θ/ sound (as it is the norm in Spanish).

<sup>4</sup>Here, *non-alternating* stands for verbs with no special pattern, and *suppletion* for some verbs with patterns that only apply to them, stem suppletion (Boyé & Cabredo Hofherr 2006), and verbs derived from these (e.g. *decir* ‘to say’, and *bendecir* ‘to bless’).

## 8.1 Multiple inheritance and cross-hierarchies: Spanish verbal inflection

Table 8.4: Number of verbs by pattern and thematic vowel in a sample of 3054 Spanish verbs.

	Thematic vowel		
	a	e	i
b~t	0	0	9
e~i	0	0	23
e~ie	65	17	32
g	0	31	11
ig	0	11	0
i~ie	0	0	2
i~iet	0	0	6
suppletion	1	9	10
L (j)	0	0	31
o~ue	51	22	2
non-alternating	2409	79	143
u~ue	1	0	0
z (/θ/)-zc	0	73	16

pattern the list can be much longer (some lists mention up to 101 verbal patterns).<sup>5</sup> If we only focus on the patterns listed here the basic type hierarchy as in Figure 8.2<sup>6</sup> is sufficient.

Notice that there it is not necessary to list the specific position for the phonological process in the case of diphthongization because this process necessarily applies to the stressed syllable, except when the item appears with a derivational suffix that attracts stress like the diminutive *-ito* (*'poblar ~ 'pueblo ~ pue'blito*, ‘to populate’, ‘town’, ‘small town’) (Carreira 1991).

Combining the hierarchies in Figure 8.2 and Figure 8.1 produces a cross-classification as in Figure 8.3. Notice that in this hierarchy, the classes *theme-vowel* and *stem-space* refer to two different kinds of processes, or aspects of verb inflection that interact with each other.

Additional evidence for postulating cross-classification of two independent hierarchies comes from two observations. First, as mentioned before, some of the stem alternations are not exclusively restricted to verbs, but can also appear in

<sup>5</sup><http://www.verbolog.com/conjuga.htm>, visited 20.10.2016.

<sup>6</sup>The use of an *irregular* type is not really needed, however. Completely irregular verbs can be modelled by using lexical entries with a fully specified, and irregular, stem space.

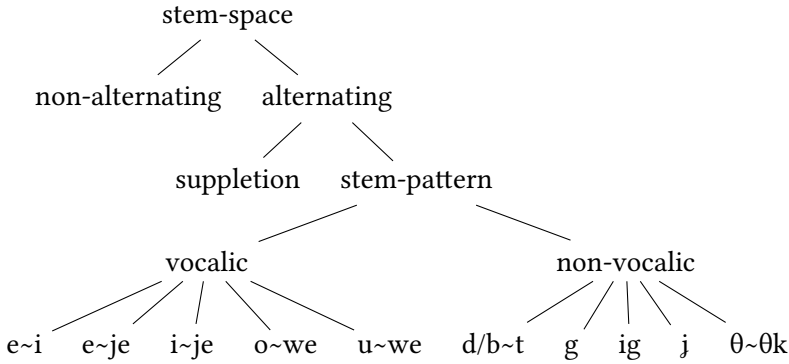


Figure 8.2: Hierarchy for Spanish verb stem alternations.

nouns: *dental* ~ *diente* ('dental', 'tooth'), *pernil* ~ *pierna* ('(animal) leg', 'leg'), *molar* ~ *muela* ('back tooth', 'tooth'), etc (Carreira 1991). Second, the case of *poner* ('to put') suggests that cross-classification can also occur within the stem alternation hierarchy, as it would belong to both types /g/ (1sg *pongo*) and /o-we/ (pp *puesto*). For the purposes of this study I will ignore these interactions due to their sparsity (see Fondow 2010 for a historical take on this particular class of verbs in Spanish).

The hierarchy concerning the thematic vowel in Figure 8.1 can be said to only be relevant for the actual endings in the inflected forms, but not so much for the stem alternations, besides specifying that *-ar* verbs do not seem to exhibit any non-vocalic stem alternation. At this point we cannot tell whether this is an accidental gap or a fact we should hardwire into the grammar. In contrast, the hierarchy in Figure 8.2 is about the stem alternations found in the different verbs.

Although Boyé & Cabredo Hofherr (2006) argue for the need of eleven stems for the Spanish paradigm, I will only focus here on the stems for the principal parts of verbs, since the other stems can be easily integrated into this system. I use a simplified stem specification as in (1) for Spanish verbs.

$$(1) \left[ \begin{array}{c} \text{STEMS} \left[ \begin{array}{c} \text{SLOT1} \\ \text{SLOT2} \\ \text{SLOT3} \end{array} \right] \end{array} \right]$$

In (1) SLOT1 is the stem of the 1sg present, SLOT2 is the stem of the past participle, and SLOT3 is the stem of the gerund. With this, a regular verb like *amar* ('love') would have a stem specification as in (2), but a completely irregular verb like *ir* ('to go') would have a stem specification as in (3).

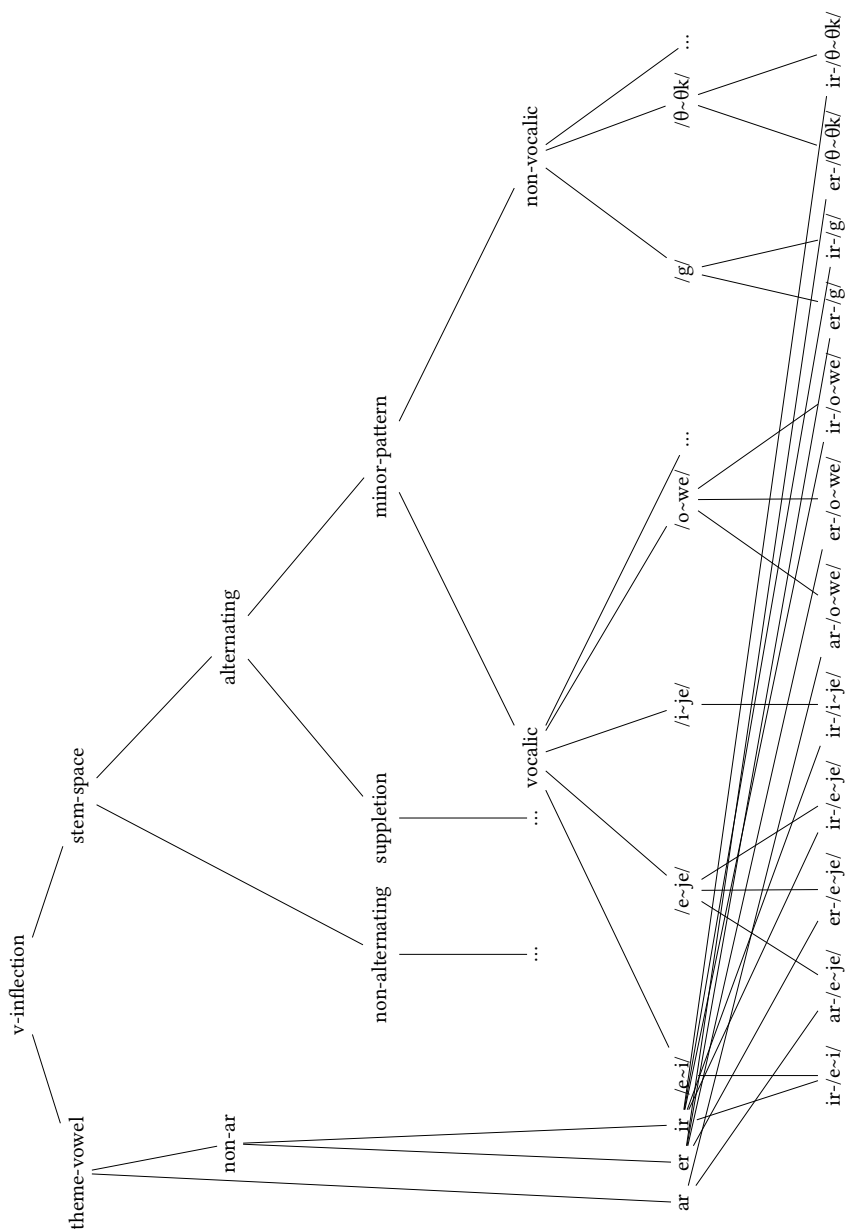


Figure 8.3: Complete hierarchy for Spanish verbs.

$$(2) \left[ \begin{array}{c} \text{STEMS} \left[ \begin{array}{c} \text{SLOT1 am} \\ \text{SLOT2 am} \\ \text{SLOT3 am} \end{array} \right] \end{array} \right]$$

$$(3) \left[ \begin{array}{c} \text{STEMS} \left[ \begin{array}{c} \text{SLOT1 voy} \\ \text{SLOT2 ido} \\ \text{SLOT3 y} \end{array} \right] \end{array} \right]$$

As pointed out before, however, the stem alternations of most verbs are not unsystematic, and we would like to capture these patterns. Additionally, we would like to avoid directional implicational relations, where one stem is used to derive all other stems, thus giving it some special status. I present here a very simple sketch that aims to achieve this. The point is to define the stem alternation types as constraints on the alternations seen for a verb of such a type. So, for the type  $b \sim t$ <sup>7</sup>, we have a constraint as in (4) where the co-indexed boxes indicate string identity.

$$(4) \quad b/d-t \Rightarrow \left[ \begin{array}{c} \text{STEMS} \left[ \begin{array}{c} \text{SLOT1 } \boxed{1}+b- \\ \text{SLOT2 } \boxed{1}+t- \\ \text{SLOT3 } \boxed{1}+b- \end{array} \right] \end{array} \right]$$

Similar constraints for all other alternations presented in Table 8.3 can be defined. Some examples are shown in (5) and (6).

$$(5) \quad e-i \Rightarrow \left[ \begin{array}{c} \text{STEMS} \left[ \begin{array}{c} \text{SLOT1 } \boxed{1}+i+\boxed{2}- \\ \text{SLOT2 } \boxed{1}+e+\boxed{2}- \\ \text{SLOT3 } \boxed{1}+i+\boxed{2}- \end{array} \right] \end{array} \right]$$

$$(6) \quad e-je \Rightarrow \left[ \begin{array}{c} \text{STEMS} \left[ \begin{array}{c} \text{SLOT1 } \boxed{1}+je+\boxed{2}- \\ \text{SLOT2 } \boxed{1}+e+\boxed{2}- \\ \text{SLOT3 } \boxed{1}+i+\boxed{2}- \end{array} \right] \end{array} \right]$$

The  $g$  pattern is only present in the verbs *poner* ('to put'), *venir* ('to come'), *tener* ('to have'), *valer* ('to be worth'), and *salir* ('to leave, exit') and all their

<sup>7</sup>Using actual phonological specifications is, of course, possible. I use orthography for simplicity, and because in the case of Spanish the orthographic representation does not hide important aspects of the morphology.

derivatives. In the case of *poner* shows that there is additional cross-classification with *o-we*: *puesto*. These can be seen in (7) and (8).

$$(7) \quad g \Rightarrow \left[ \text{STEMS} \begin{bmatrix} \text{SLOT1 } \boxed{1}+g- \\ \text{SLOT2 } \boxed{1}- \\ \text{SLOT3 } \boxed{1}- \end{bmatrix} \right]$$

$$(8) \quad o\text{-}we \Rightarrow \left[ \text{STEMS} \begin{bmatrix} \text{SLOT1 } \boxed{1}+we+\boxed{2}- \\ \text{SLOT2 } \boxed{1}+o+\boxed{2}- \\ \text{SLOT3 } \boxed{1}+o+\boxed{2}- \end{bmatrix} \right]$$

The pattern /ig/ is restricted to verbs ending in /a/ that belong to the *-er* conjugation: *traer*, *caer* and derivatives. At first sight, one could think this is a simple exception, but any new verb with this shape would also take this stem pattern. If given a wug verb like *saer*, the 1sg form would be *saigo*. The analogical constraint that specifies this pattern is simple enough: /a#/, but the complexity in the analogical specification is a matter of degree. The more productive cases are only partially specified, and this is precisely what makes them more productive (they have fewer restrictions on the shape of the stems that can appear with them). This constraint is shown in (9).

$$(9) \quad ig \Rightarrow \left[ \text{STEMS} \begin{bmatrix} \text{SLOT1 } \boxed{1}+ig- \\ \text{SLOT2 } \boxed{1}+i- \\ \text{SLOT3 } \boxed{1}+j- \end{bmatrix} \right]$$

The /i-je/ pattern is also very limited, only appearing in my corpus with *inquirir* ('to inquire') and *adquirir* ('to acquire'). Notice that in this case *-quirir* is not a verb, so neither verb is a derived form in itself, despite the presence of the prefixes *in-* and *ad-*. As with /ig/ before, any new verb that would take the form *-quir-* in its stem, would also inflect by the /i-je/ pattern: *sanquirir* - *sanquiero*. A structure like the one presented in (10) captures this pattern.

$$(10) \quad i\text{-}je \Rightarrow \left[ \text{STEMS} \begin{bmatrix} \text{SLOT1 } \boxed{1}+(k)+je+(r)- \\ \text{SLOT2 } \boxed{1}+(k)+i+(r)- \\ \text{SLOT3 } \boxed{1}+(k)+i+(r)- \end{bmatrix} \right]$$

I mark in parentheses the segments which will necessarily appear in the stem for clarity, but the constraint in (10) does not need to specify them. One might

be tempted to suggest that these extremely restrictive patterns should specify their restrictions directly on the lexical items themselves. This, however, would be missing out on the fact that these very restrictive patterns are just an extreme case of the more productive patterns. This is easily captured by using the analogical/form similarity function that licenses items being in particular types. For example, the difference between regular *ir* class verbs and *i-je* verbs is that regular *ir* class verbs have fewer formal restrictions than *i-je* verbs.

As stated before, these are simply sketches, and a more formal analysis could probably split these patterns into more basic processes, or collapse others based on more general phonological specifications. The important point here is that the definition of the minor patterns can be done in a way that is independent of whatever the major pattern of the stem is. This way the interaction between both types becomes straightforward. I will argue that the experimental results strongly support the observation that major and minor patterns are mostly independent of each other.

### 8.1.4 Materials

For this section I first extracted all verbs from a Spanish frequency list based on subtitle corpora.<sup>8</sup> From this list I extracted all lemmas using TreeTagger (Schmid 1995). This produced a list of 4271 lemmas, from which I removed all reflexive forms, verbs without complete conjugation paradigms, and verbs whose stem is too short to play a role in an analogical model (e.g. *ir*). The final list was comprised by 3052 verb lemmas, for which I produced all three principal parts.

Extracting the stem of the verbs was relatively easy in this case, because we define the infinitive stem as the verb minus the thematic vowel and final *r*. Additionally, to control for orthography I replaced all letter pairs that represent a single phoneme with a single symbol (e.g. *ch* → *C*, *ll* → *L*, etc.). Because of the imbalance seen in the proportion of *ar* verbs vs all other verbs, I left only in the dataset the 300 most frequent *ar* verbs, which produced a 808-verb dataset<sup>9</sup>. I present side by side statistical results from the smaller dataset and the complete dataset, but focus on the distributions obtained with the smaller dataset.

---

<sup>8</sup>Found at: <https://invokeit.wordpress.com/frequency-word-lists/>, visited 8-11-2016.

<sup>9</sup>It is worth mentioning here that leaving all verbs in the dataset did not produce significantly worse results in the models, but did introduce a confound when interpreting the role of *ar-non-alternating*. The accuracy metrics used are somewhat sensitive to these imbalances, and the accuracy of a model will be very high if the model always predicts the most frequent class. This sometimes makes models over-generalize towards the more frequent class and ignore patterns in the less frequent classes. Ultimately this is a weakness of the models I am using which could possibly be overcome with a different approach.

### 8.1.5 Results

There are three interesting models to look at. First, we test how well our analogical model can predict the thematic vowel of the verb. This is the basic model, which should basically capture insights mentioned before (Boyé & Cabredo Hofherr 2004). The second model should predict the minor patterns. Finally, the third model will deal with the combination of both dimensions, giving us a the full predictions of verb inflection classes.

We start with the model predicting the major inflection pattern. This model only looks at the final three segments of the stems thematic vowel ~ final.1 + final.2 + final.3<sup>10</sup>. The results are presented in Table 8.5, and the corresponding statistics in Table 8.7.

Table 8.5: Confusion matrix for the model predicting thematic vowel of Spanish verbs.

	Reference		
Prediction	ar	er	ir
ar	302	19	42
er	25	208	9
ir	51	7	225

Table 8.6: Confusion matrix for the model predicting thematic vowel of Spanish verbs with full dataset.

	Reference		
Prediction	ar	er	ir
ar	2400	48	118
er	37	182	3
ir	89	3	154

<sup>10</sup>The model had eight hidden nodes, and a decay rate of 0.09. There was no noticeable improvement from using more structured predictors.



Table 8.7: Statistics for Table 8.5.

Overall Statistics			
Accuracy : 0.8277			
95% CI : (0.8012, 0.852)			
No Information Rate : 0.4257			
Kappa : 0.737			
Statistics by Class:			
	Class: ar	Class: er	Class: ir
Sensitivity	0.799	0.889	0.815
Specificity	0.880	0.948	0.905
Neg Pred Value	0.854	0.957	0.904
Balanced Accuracy	0.839	0.919	0.860

Table 8.8: Statistics for Table 8.6.

Overall Statistics			
Accuracy : 0.9019			
95% CI : (0.8906, 0.9121)			
No Information Rate : 0.8326			
Kappa : 0.6528			
Statistics by Class:			
Class: ar	Class: er	Class: ir	
Sensitivity	0.950	0.781	0.560
Specificity	0.673	0.985	0.966
Neg Pred Value	0.731	0.981	0.956
Balanced Accuracy	0.812	0.883	0.763

## 8.1 Multiple inheritance and cross-hierarchies: Spanish verbal inflection

Table 8.9: Distance Matrix for Table 8.5.

	ar	er
er	2.25	
ir	1.21	2.89

Table 8.10: Distance Matrix for Table 8.6.

	ar	er
er	2.35	
ir	1.06	2.92

First of all, the model has a very high accuracy and kappa score. It is clear that the prediction of the thematic vowel is possible from the stem of the verb. Somewhat worrying, however, is that the confusion between the three classes does not follow the predictions made by the hierarchy in Figure 8.1. In the model *er* and *ir* show less confusion with each other than with *ar*. This seems to go against the hierarchy proposed to model their morphological asymmetries. Just looking at this case it appears as a strong counter example for the thesis of this book. However, if instead of measuring the distance based on the errors made by the model, we measure this distance directly on the probability matrix, the result is very different. The distance matrices can be seen in Table 8.11 and Table 8.12. In the reduced dataset the distances are pretty much the same between the three classes (with minor variations), while in the complete dataset there is a strong effect in the expected direction, that is, *class-er* is closer to *class-ir*. The problem here is that this effect is caused by the frequency imbalance between the classes. Because *class-ar* has so many more members that are correctly predicted, the overall distance of this class from the other two increases. At best this particular case remains inconclusive.

Next, we try to predict the minor inflectional pattern only. We fit the same model as before: `minor pattern ~ final.1 + final.2 + final.3`.<sup>11</sup> The results are shown in Table 8.13 and Table 8.14 (the overall results for the full dataset are in Table 8.15).

<sup>11</sup>With eight hidden nodes and a decay rate of 0.01.

Table 8.11: Distance Matrix on probabilities for the reduced dataset.

	ar	er
er	2.12	
ir	2.05	2.19

Table 8.12: Distance Matrix on probabilities for the complete dataset.

	ar	er
er	2.46	
ir	2.41	1.55

Table 8.13: Confusion matrix for the model predicting minor inflection patterns of Spanish verbs.

	Reference											
Prediction	b~t	e~i	e~ie	g	ig	i~ie	i~iet	L	o~ue	non-alt.	z~zc	u~ue
b~t	9	0	0	0	0	0	0	0	0	0	0	0
e~i	0	13	0	0	0	0	0	1	0	9	0	0
e~ie	0	0	31	0	0	0	0	0	0	8	1	0
g	0	0	1	40	0	0	0	0	3	4	0	0
ig	0	0	0	0	11	0	0	0	0	0	0	0
i~ie	0	0	0	0	0	2	0	0	0	0	0	0
i~iet	0	0	0	0	0	0	6	0	0	0	0	0
L	0	0	0	0	0	0	0	28	0	3	0	0
o~ue	0	0	0	0	0	0	0	0	31	11	0	0
non-alt.	0	10	28	2	0	0	0	2	8	452	3	1
z~zc	0	0	1	0	0	0	0	0	0	3	85	0
u~ue	0	0	0	0	0	0	0	0	0	1	0	0

## 8.1 Multiple inheritance and cross-hierarchies: Spanish verbal inflection

Table 8.14: Overall and by class statistics for Table 8.13.

Overall Statistics				
Accuracy : 0.8762				
95% CI : (0.8515, 0.8982)				
No Information Rate : 0.6077				
Kappa : 0.792				
Statistics by Class:				
	Class: b~t	Class: e~i	Class: e~ie	Class: g
Sensitivity	1.000	0.565	0.508	0.952
Specificity	1.000	0.987	0.988	0.990
Neg Pred Value	1.000	0.987	0.961	0.997
Balanced Accuracy	1.000	0.776	0.748	0.971
	Class: ig	Class: u~ue	Class: i~ie	Class: i~iet
Sensitivity	1.000	0.000	1.000	1.000
Specificity	1.000	0.999	1.000	1.000
Neg Pred Value	1.000	0.999	1.000	1.000
Balanced Accuracy	1.000	0.499	1.000	1.000
	Class: L	Class: o~ue	Class: non-alt	Class: z~zc
Sensitivity	0.903	0.738	0.921	0.955
Specificity	0.996	0.986	0.830	0.994
Neg Pred Value	0.996	0.986	0.870	0.994
Balanced Accuracy	0.950	0.862	0.875	0.975

Table 8.15: Overall and by class statistics for model predicting minor patterns on the full dataset.

Overall Statistics
Accuracy : 0.9268
95% CI : (0.917, 0.9358)
No Information Rate : 0.8672
Kappa : 0.6888

Once again, the model has a good accuracy in predicting these minor patterns, even those claimed to be unpredictable. This is not too surprising given the previous studies that have already found strong phonological regularities that correlate with diphthongization. Some of the consonant patterns are in fact (almost) fully predictable by simple rules. Most verbs ending in /n/ are of *class-g*, while all verbs that end in /a/ are of *class-ig*. This is interesting because it means that this particular tree is a mix of fully and partially predictable classes, which lends support to the claim that the filter that assigns stems to types can go from a fixed simple constraint to a more complex pattern. Finally, *non-alternating* is indeed the default class, with the lowest negative predictive value. Remember that the negative predictive value represents how many false positives are in a given class. The class with the lowest negative predictive value is the class where most errors from other classes are grouped. Whenever the model does not know what class an item should be assigned to, it assigns it to the default class.

For the last case we try to predict the complete conjugation of the verb (i.e. the thematic vowel and minor inflection pattern together). The model is once more the same:  $\text{conjugation} \sim \text{final.1} + \text{final.2} + \text{final.3}$ <sup>12</sup>. The corresponding heat map is shown in Figure 8.4, and the corresponding statistics in Table 8.16.

These results show that *ar-non-alternating* is still the class with lowest negative predictive value, which means it is the default class for our model, as predicted. Most of the other classes are relatively more or less predictable, with some diphthongization classes having little predictability, like *ir-o~ue* and *ar-u~ue*. These are, however, extremely infrequent, with 2 and 1 frequency counts, respectively. It is not surprising that such low-frequency classes should be hard or impossible to predict. It is also expected that combining both dimensions causes some classes to have low predictability. After all, we use the same three predictors to predict sixteen classes, instead of the three and eight from before. The validation results of this final model are presented in Figure 8.5.

The results of the MDS and clustering are shown in Figure 8.6. These clusters exhibit several interesting properties. First, the types *ar-non-alternating*, *er-non-alternating*, and *ir-non-alternating* are all three in the corners of the space. These are maximally different from each other. The color clustering seems less insightful in this case than the MDS, but some groups do form nicely. The least insightful cluster is probably the lila one in the lower left quadrant with the patterns *ir-b~t* and *ir-z~zc*, and directly besides this one (in light orange) the alternations *er-ig* and *ir-L*. These two clusters do not seem to follow any pattern, but then again, there is little organization to them. In red we have a clear cluster of *ir-g* and *er-g*,

<sup>12</sup>With eight hidden nodes and a decay rate of 0.01.

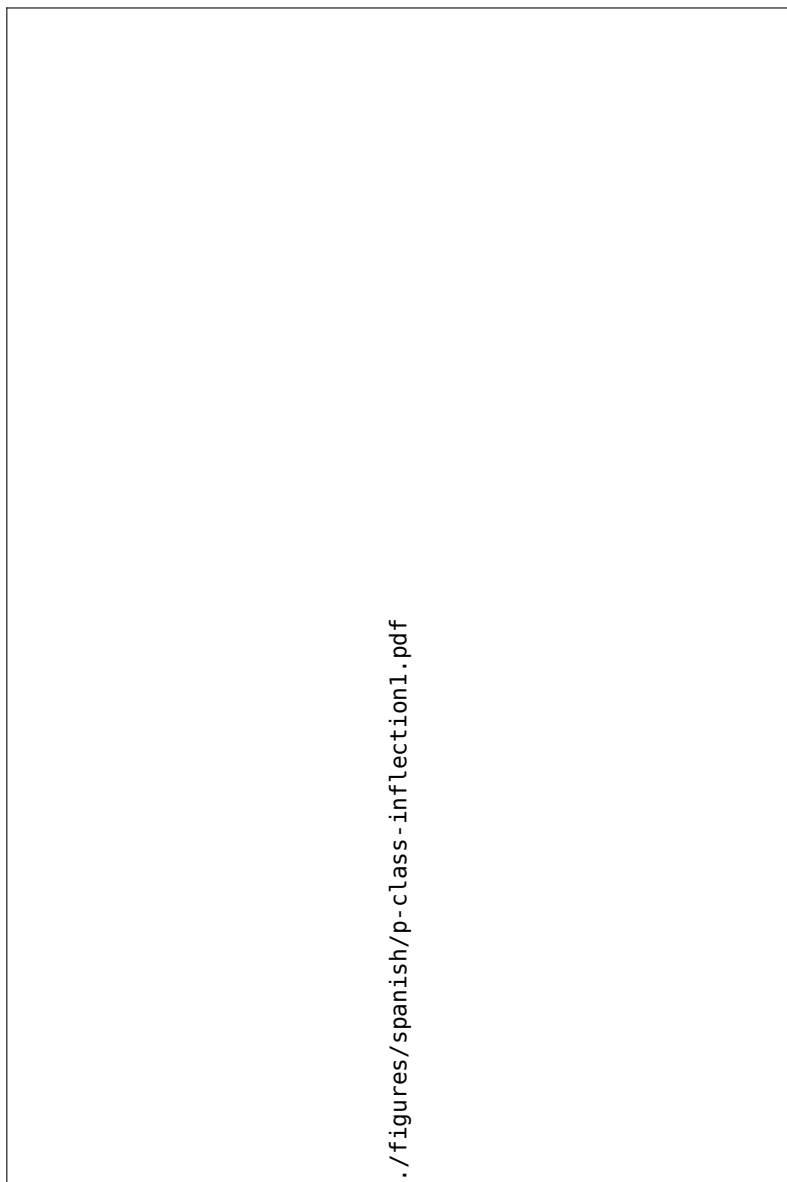


Figure 8.4: Heat map for the model predicting inflection class of Spanish verbs.

Table 8.16: Overall and by class statistics for Figure 8.4.

Overall Statistics				
Accuracy : 0.7772				
95% CI : (0.7469, 0.8055)				
No Information Rate : 0.3329				
Kappa : 0.7313				
Statistics by Class:				
	Class: ar-e-ie	Class: ar-non-alt	Class: er-e-ie	Class: er-g
Sensitivity	0.2500	0.7695	0.0588	0.9355
Specificity	0.9975	0.8794	0.9924	0.9961
Neg Pred Value	0.9888	0.8843	0.9800	0.9974
Balanced Accuracy	0.6237	0.8245	0.5256	0.9658
	Class: er-ig	Class: er-o~ue	Class: er-non-alt	Class: er-z-zc
Sensitivity	1.0000	0.6818	0.7595	0.9589
Specificity	1.0000	0.9924	0.9561	0.9918
Neg Pred Value	1.0000	0.9911	0.9735	0.9959
Balanced Accuracy	1.0000	0.8371	0.8578	0.9754
	Class: ir-e-i	Class: ir-e~ie	Class: ir-g	Class: ir-i~iet
Sensitivity	0.4348	0.6562	0.9091	1.0000
Specificity	0.9898	0.9910	0.9912	0.9975
Neg Pred Value	0.9835	0.9859	0.9987	1.0000
Balanced Accuracy	0.7123	0.8236	0.9502	0.9988
	Class: ir-L	Class: ir-non-alt	Class: ir-z-zc	Class: ar-o~ue
Sensitivity	0.9355	0.8462	1.0000	0.5000
Specificity	0.9974	0.9639	0.9987	0.9899
Neg Pred Value	0.9974	0.9668	1.0000	0.9886
Balanced Accuracy	0.9665	0.9050	0.9994	0.7449
	Class: ir-b~t	Class: ir-i~ie	Class: ir-o~ue	Class: ar-u~ue
Sensitivity	1.0000	0.5000	0.0000	0.0000
Specificity	0.9987	1.0000	1.0000	1.0000
Neg Pred Value	1.0000	0.9988	0.9975	0.9988
Balanced Accuracy	0.9994	0.7500	0.5000	0.5000

and in dark blue we see a similar situation with the cluster *ar-o~ue* and *er-o~ue*. These two clusters organize according to the stem patterns, and not according to the thematic vowels. The class *ir-o~ue* is very close in the plane to the other two *o~ue* patterns, but by the hierarchical clustering analysis grouped together with the *ir* alternations *ir-i~ie* and *ir-i~iet*. In this case the thematic vowel seems to be more important for the organization of these three patterns. In light blue we have the classes *ir-a~ie*, *ir-e~i* and *ir-e~ie*. Here we see again three classes that basically cluster around stem patterns.

The clusters are by no means perfect, but they do match the proposed hierarchy to some extent: there are three major inflection patterns that correspond to the thematic vowel, and there are some minor conjugation patterns that cross-classify with these.

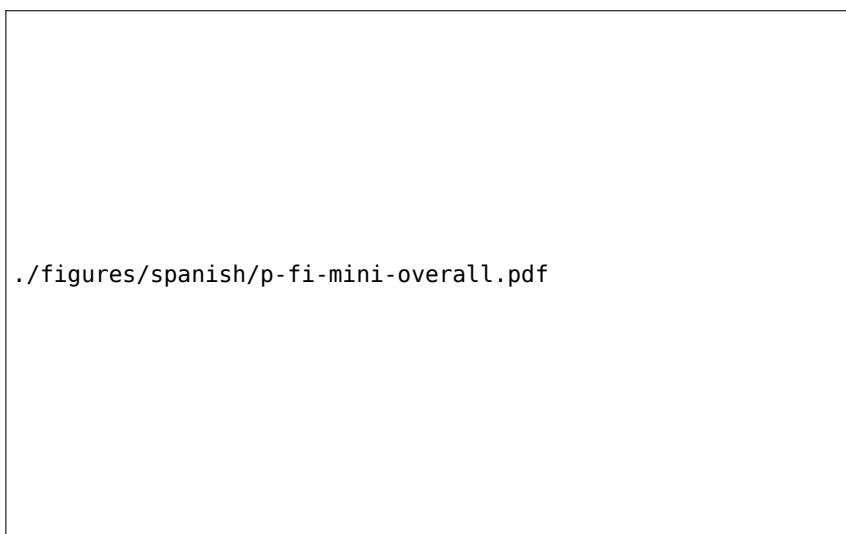


Figure 8.5: Overall validation for the model predicting inflection class of Spanish verbs.

Some of these effects seem to contradict the claim by Albright (2009) that analogical effects are local to the three major classes. These results show that analogical effects between minor patterns run across these three main classes. Although there is no clear explanation for why some clusters prefer to form around the thematic vowel, while others group around the stem patterns, it seems clear that there must be analogical relations that run through both subtrees of the inflection hierarchy of Spanish verbs. In the model I propose here, all dimensions of the hierarchy can carry some analogical information. However, which dimensions will matter most, or where the strongest similarities will be found, cannot be determined by any particular property of the hierarchy.

For Spanish, it is also interesting to compare the model to the experimental results of Albright (2009) mentioned above. As already described, in the original experiment, Albright et al. (2001) tested 96 native Spanish speakers on wugs to see whether these wugs would be prone to diphthongization or not. The author used 33 wugs with forms like *lerrar*. Speakers were presented with the verb used in a non-alternating context, like the first person plural (*lerramos*), and then asked to fill in a dialog where the wug appeared in non-alternating and alternating contexts. The authors then calculated the probability of a wug diphthongizing as: the number of speakers who produced a diphthongized form for said wug, over the total number of speakers.



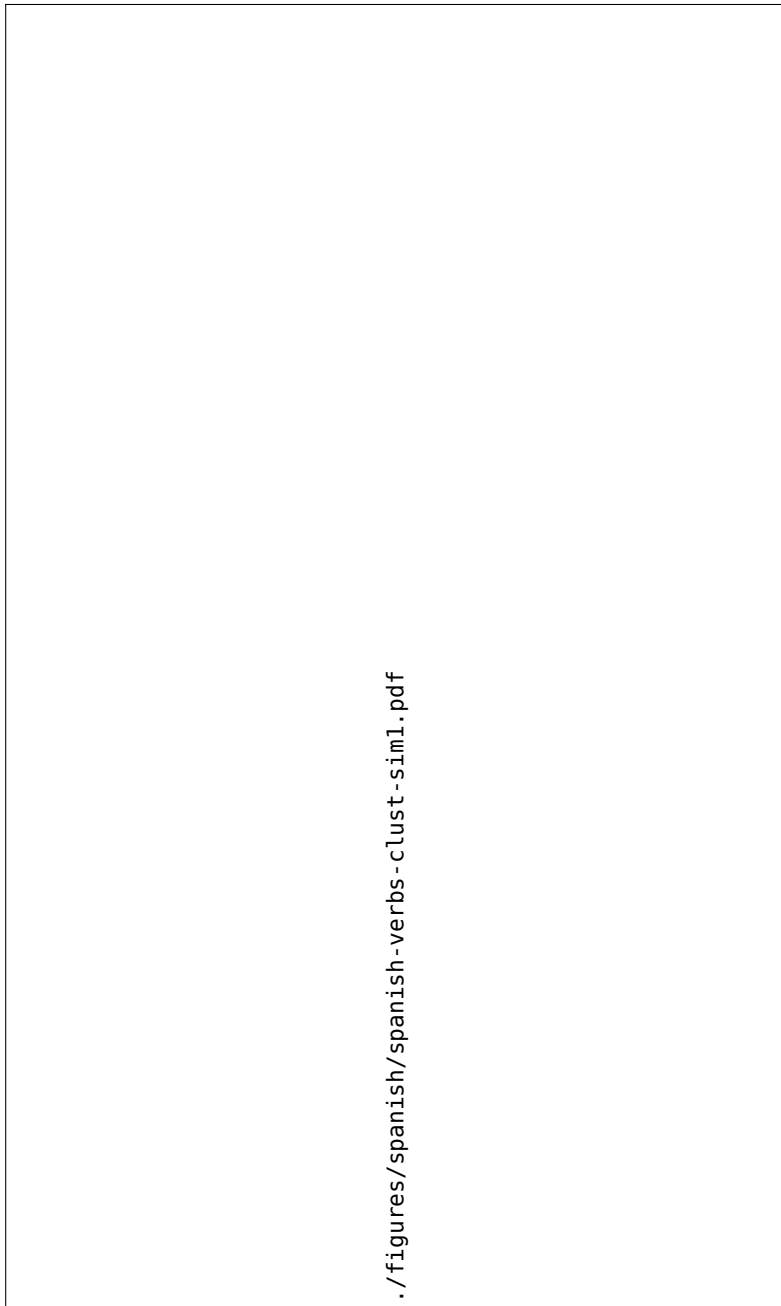


Figure 8.6: Multidimensional scaling with hierarchical clustering for label colors.

Since we are now predicting experimental data, we can use the complete dataset (with 3000 verbs) without doing any splitting. As the experimental dataset only contains information about mid vowel diphthongs, we have to fit a model trained to predict only this factor. In this case, the previous formula for fitting the model did not perform as well. A more structurally defined model did a much better job: `diphthong ~ final.1 + final.2 + pre-theme vowel * theme vowel + n_clusters`<sup>13</sup>.

This model also takes the final and prefinal segments of the stem, but additionally identifies the pre-thematic vowel interacting with thematic vowel, and the number of consonant clusters<sup>14</sup>. The reason for also adding the thematic vowel is simple. Albright presents a model trained exclusively on *ar* verbs. Adding the thematic vowel in this case means that the model knows what the main portion of the dataset it should look at is when making the predictions, but also has the rest of the dataset to learn from. This is important because our model is less capable of making large phonological generalizations than Albright's is, and every bit of data matters.

When predicting the wugs, the model achieved a correlation of  $r = 0.59$  ( $p < 0.05$ ), which is quite close to the generalized context model Albright (2009) reports on ( $r = 0.56$ ). It is, however, considerably below the minimal generalization learner ( $r = 0.77$ ). The predicted probabilities in Figure 8.7 show where the problem lies. The analogical model has difficulties with some wugs ending in complex clusters. This is because these particular combinations are either not present in the data (*etC* is missing) or very rare (*otr* has a frequency of 1). This shows that the generalizations the model makes are too local, and not general enough to capture weird looking wugs correctly. Nevertheless, this is not a bad performance in the sense that the model seems to have some sort of correlate with speaker's intuitions, particularly regarding wugs that do look like observed words. Those cases where speakers were much less likely to allow for diphthongs are also completely disallowed by the model.

The fact that the minimal generalization learner outperforms the analogical model means that the latter is a rougher approximation to what speakers actually do than the former. It is likely that the analogical model better captures the regularities of the synchronic system, but fails to distinguish between truly productive patterns and unproductive patterns. On the other hand, a big downside of Albright's approach is that it only predicts one categorical distinction.

<sup>13</sup>Because we are now predicting probabilities, using the `linout` linking function produces better results. The model had no hidden nodes and only a skip layer.

<sup>14</sup>I take any two consonants appearing together to be a consonant cluster.



Figure 8.7: Predicted vs. observed probabilities of diphthong stems.

In contrast, our model is capable of precise class assignment. Ultimately a more sophisticated system would be required to be able to perform both tasks well: simulate speaker’s performance and fine class predictions. Finally, wugs ending in *otr* all get different empirical probabilities. This shows that the initial models that only consider the last three segments are missing something.

## 8.2 Cross-classifications between plural and singular: Kasem

Kasem is a Gur Language, of the Grusi branch, spoken mostly in Ghana and Burkina Faso (Naden 1988). Kasem featured prominently during the seventies and eighties in phonological debates (Phelps 1975; 1979; Halle 1978; de Haas 1987; 1988) because of coalescence phenomena (see also Zaleska 2017). Like other Gur languages (Naden 1989), Kasem exhibits a complex system of genders and classes that has received relatively little attention in the literature (see Awedoba (2003) for some recent discussion of the Kasem gender system, and Niggli & Niggli

(2007) for an electronic dictionary of Kasem). Kasem is traditionally analyzed as having 5 genders and 9 nominal classes:

a class is considered singular if the majority of its members are singular, semantically and grammatically; and plural if the majority of its membership is grammatically and semantically plural. There are four singular classes and five plural classes. A pairing of a singular and a plural class constitutes a gender (Awedoba 2003: 4)

Gender is defined with relation to the agreement of the noun with the determiner (most adjectives do not agree at all, and those which do have inherent markers). Awedoba (2003: 3) proposes the classification shown in Table 8.17 (adapted from the original, and with additional information from Awedoba 1980 and Awedoba 1996)<sup>15</sup>. I will show in the following sections that this approach is insufficient to properly capture the complexity of the Kasem system. Nonetheless, the organization in Table 8.17 already gives us an idea of what the problem is (for work on the noun class systems of related languages see Brindle 2009, Bodomo 1994, Bodomo 1997, and Dakubu 1997): there are five genders based on agreement patterns with pronouns<sup>16</sup>, and many more number markers that do not correspond 1 to 1 with said genders. In a way, this is a similar situation to the Romanian system discussed in Chapter 5.

Table 8.17: Gender and classes in Kasem.

Gender	sg. classes			pl. classes		
	noun class	marker	Det	noun class	marker	Det
1	I	u, i, a	wɔm	II	a	bam
2	III	i	dɔm	IV	a	yam
3	V	a	kam	VI	i	sum
4	VII	u	kɔm	VIII	0, du	tum
5	VII	u	kɔm	IX	0, ni	dum

Awedoba (1980: 249) admits that the markers in Table 8.17 are only the ones he considers to be the most frequent in the language, and that there are other less frequent ones (I will present several additional markers in the following sections).

<sup>15</sup>Other sources label the genders with letters from A to E (Callow 1965).

<sup>16</sup>The literature does not present any clear examples, but it is mentioned.

Since the author does not provide an explicit list of all the markers and the genders they define, and because gender assignment is defined by the combination of a noun's singular and plural markers, I will not focus on gender, but rather on the question of how number markers get assigned to nouns<sup>17</sup>. This question has been studied before. Some semantic regularities seem to be present in the gender assignment patterns. Gender 1 mostly contains human nouns. Gender 2 contains fruit names and body parts, among others. Gender 3 also contains body parts and names of fruits, but also animals, trees and other plants. Gender 4 seems to be the default class, and Gender 5 is claimed to only contain some 20 nouns mostly related to domestic items. The author concludes that

Kasem Genders are not based on a grouping of homogeneous items. While a gender may contain items from several semantic categories, no gender can be said to monopolise absolutely nouns belonging to any one semantic category (Awedoba 2003: 7)

A further complication for the semantic analysis is that stems can belong to multiple genders. So, while the term for a Kasem person *kasinɔ* belongs to Gender 1, the term for the language *kasinɪ* belongs to Gender 2. Similarly, diminutives belong to Gender 1, even if the stem belongs to any of the other genders.

Some hints towards the possibility of formal analogical relations are already present, although not spelled out (notice that the author does not tell us what the underlying forms would be in the proposed examples):

While the semantic bases of the genders cannot be denied, as the discussion so far illustrates, phonology can also play a role in the allocation of nouns to classes and gender (Awedoba 2003: 12)

and

The final syllable of a noun, especially the quality of the final vowel, plays some role in the allocation of nouns to their genders. For example, although *bugə* 'river', Gender 3 and *bugə* 'tiredness', Gender 2 appear to be homophones they are assigned to different classes and genders not necessarily on semantic grounds but perhaps on account of their suffixes, which happen on the surface to be identical but not in deep structure (Awedoba 2003: 13)

---

<sup>17</sup>There is also the more practical problem that the dictionary does not contain gender information. This means that gender can only be inferred from the markers themselves.

Similarly, Awedoba (1980: 250) had already observed (informally) that gender assignment for loan words in Kasem follows semantic and phonological analogy.

Another important data point mentioned by Awedoba (2003: 13) (first found in Awedoba (1996)), but not discussed with relation to the analogical relations in the system, is the fact that noun-adjective compounds can have different genders independently of the head noun in the compound. So, while *ka-balaṇa* (woman-small, ‘small woman’) belongs to Gender 3, *kə-kəṃumu*<sup>18</sup> (woman-big, ‘big woman’) belongs to Gender 4. This indicates that the adjective assigns the gender of the compound, and not the head noun. This is interesting because it means that formal features can easily overcome semantic features in gender assignment in Kasem.

Kasem also has a complex tone system. However, because the dictionary I am relying on (Niggli & Niggli 2007) only lists the tones for the singular form, and it is not clear what happens to those tones in many plurals (especially when the number of syllables of the singular and plural are different), I will not consider tone in this study.

### 8.2.1 ATR in Kasem

In Kasem, as in many West African languages, there is an alternation between [+ATR] (advanced tongue root: /u/, /i/, /ə/, /e/, /o/) and [-ATR] (/ʊ/, /ɪ/, /a/, /ɛ/, /ɔ/) vowels (Casali 2008): /ʊ-u/, /ɪ-i/, /a-ə/, /ɛ-e/, /ɔ-o/. Words (with the exception of compounds), have the same [ATR] specification for all their vowels:

In the simplest and most general case of ATR harmony, the vowels in any given word are either all [+ATR] or all [-ATR]. Thus, words in which some vowels are [+ATR] and others [-ATR] do not ordinarily (setting aside certain common classes of exceptions) occur (Casali 2008: 496)

This feature, as can be seen in (11),<sup>19</sup> creates minimal pairs and seems to be lexically specified.

<sup>18</sup>See the following subsection for an explanation on ATR in Kasem.

<sup>19</sup>All Kasem examples are taken from Niggli & Niggli (2007).

## 8 Complex inflectional classes

(11)	singular	plural	gloss	
a.	colo	cwəlu	‘kilogram’	+
b.	cɔlɔ	cwaalɔ	‘girl that likes going out with men’	-
c.	peeli	peelə	‘shovel, spade’	+
d.	pɛli	pɛla	‘bean cake’	-
f.	vəlu	vələ	‘traveller’	+
e.	valɔ	vala	‘farmer’	-
g.	yiri	yirə	‘type, kind’	+
h.	yiri	yira	‘name’	-

There are, however, some cases in the dictionary where it is not completely clear whether we are dealing with exceptions to this rule or errors in the dictionary itself:

(12)	singular	plural	gloss
a.	tanti	tantiə	‘aunt’
b.	yukwala	yukwalɔ	‘headscarf’
c.	yukwələ	yukwəli	‘small skull’

In (12) there is a supposedly impossible combination of /i/ and /a/, while in the other two examples /u/ appears with both /ə/ and /a/. It is recognized that in ATR harmonizing languages some words may fail to show any harmony, or only present partial harmony (Casali 2008), but it is hard to check any of the particular cases in the dictionary.

For Kasem, it is claimed that the [+ATR] feature is carried by the root, and it then extends to the affix (Casali 2008: 501). It is mainly for this reason that I will not consider ATR as a predictor or predicted feature of Kasem noun classes. I do not claim that it does not play a role, but counting it in would make an already complex system even more complex.<sup>20</sup>

### 8.2.2 A simple analysis of Kasem noun classes

There are different takes on what the number markers in Kasem are. The ones I propose here are based on my own analysis of the system. Alternative models are of course possible, but should have little impact on the analogical system. As a guiding principle for my analysis, I tried to maximize morphology and minimize phonology. Whenever there is enough evidence for a marker to be morphologically motivated, I rejected the phonological explanation for it. This is a

<sup>20</sup>In the models I neutralized ATR by converting all [-ATR] vowels to [+ATR].

conservative approach. In the worst case scenario, I am proposing more markers than there are in the system, which means that the analogical model will have a harder time to predict the classes. A smaller set of markers would result in a better model.

Kasem has many different number markers, and some of these seem to be more clearcut than others. First, I will introduce the markers where there should be less room for an alternative analysis, and in the following subsection I will introduce those cases where different approaches are possible. This runs counter to the standard way of analyzing Kasem. Previous takes on Kasem have tried to minimize the number of exponents by way of using phonological rules and underlying representations based on some further assumptions. So, for example, [de Haas \(1987: 184\)](#) analyses the example in (13) as having a marker *-i* which coalesces with the underlying vowel in the stem and turns into /e/, instead of there being a marker *-e*.

- (13) a. /zwa + i/ → /zwe/ ‘ear’  
 b. /čwa + i/ → /čwe/ ‘liver’

However, this approach relies on the assumption that *čwe* and *zwe* belong to Gender 2 (Class B in the original) based on the agreement with the determiners, and that all nouns of Gender 2 have a singular marker *-i*. This would make sense if there were compelling evidence from some other morphological process that shows that the stem of these words ends with /a/. In a few cases like *zwe*, one can propose that compounds provide such evidence. The example in (14) shows /zwa/ as a stem in three noun-adjective compounds (these are all right-headed compounds, in that order: noun-adjective):

(14)	singular	plural	gloss
a.	zwa-bɔɔ	zwa-bɔɔrɔ	‘hole in the ear’
b.	zwa-kɔgɔ	zwa-kwarɔ	‘deaf person’
c.	zwa-kwana	zwa-kwana	‘earring’

However, there is no such evidence for any of the other 52 nouns that end in /e/ in the singular in the dictionary, and there is even counter evidence for a general rule. In (15) we see what could be thought to be examples just as *zwe*, where a noun belongs to Gender 2, takes the singular marker *-i* and the plural marker *-ə*, but because the stem ends in /ə/, the /i/ surfaces as /e/.



## 8 Complex inflectional classes

(15)	singular	plural	gloss
a.	kalwe	kalwə, kali	‘monkey’
b.	kandwɛ	kandwa	‘stone, rock’

However, compounds built from these nouns do seem to have a /ə/ in the stem, as shown in (16) below.

(16)	singular	plural	gloss
a.	kalwe-faa	kalwe-faarʊ	‘baboon’
b.	kalwe-sɪja	kalwe-sɪna	‘Red Patas Monkey’
c.	kalwe-zwənə	kalwe-zwəm	‘Green Monkey’
d.	kandwɛ-gara	kandwa-garɪ	‘dike’
e.	kandwɛ-nyuni	kandwa-nyuna	‘bright / shiny stone’

What this means is that even if the phonological analysis is right in the case of *zwe*<sup>21</sup>, we cannot automatically assume that this analysis applies to all nouns ending in /e/. A systematic study of each case would have to be undertaken, but because of the limitations of the dataset I am using, this is not feasible. For this reason, I will take markers to be what they appear to be in their surface form, unless there is clear and strong evidence to the contrary.

### 8.2.2.1 Basic number markers

An important feature of Kasem is that the same number markers can appear as singular markers in some nouns, and as plural markers in other nouns. The main markers (i.e. the most common ones) are: *-e*, *-ə*, *-i*, *-o*, *-u*, *-nə* and *-nu*. We see in (17) examples of the *-i* marker in the plural, with the *-ə* marker in the singular. In (18) we have the inverse situation. In both examples there is an assumption of coalescence between the /i/ in the stem and the /i/ in the marker (/i+i/ → /i/). In following sections I will discuss the possibility of an *-iə* marker instead.

(17)	singular	plural	gloss
a.	afɪɪa	afɪɪ	‘sugar cane’
b.	bordɪə	bordɪ	‘plantation’

<sup>21</sup>Even in this case it is unclear that this is the right analysis. It is not obvious that the form found in the compound is the stem, since the head noun of a compound can show some variation: *tu-mwen* (‘shrub, bush, small tree’) in the singular has the form *twe-mwan* in the plural.

(18)	singular	plural	gloss
a.	bi	biə	‘counter’
b.	pɔmpɪ	pɔmpia	‘water pump’

This is not the only possible analysis of these examples. One could also postulate a zero marker for the singular and a *-ə* marker in the plural. In this case the data are not enough to clearly distinguish between all the alternatives. I have tried to always take the most conservative approach.<sup>22</sup>

Examples in (19) and (20) show the alternation between the *-e* marker and the *-ə* marker for both singular and plural.

(19)	singular	plural	gloss
a.	cicwe	cicwə	‘spear’
b.	nafɔzwɛ	nafɔzwa	‘chapped fingers’

(20)	singular	plural	gloss
a.	gungwəŋə	gungwe	‘hour-glass drum’
b.	payaa	payɛ	‘jaw’

The examples in (21) and (22) show the *-o* and *-u* markers. While the *-o* marker rarely appears in the plural (and then only with another *-o* marker in the singular), the *-u* marker can be found both for plural and singular.

(21)	singular	plural	gloss
a.	bolo	bwəlu, bwəllu	‘valley, low land’
b.	tasɔɾɔ	taswaarɔ	‘flint lighter, lighter’

(22)	singular	plural	gloss
a.	yukolo	yukollo	‘skull’
b.	yɪrɪnɔ	yɪrɪna	‘security guard, warden’
c.	tɪəbu	tɪəbiə	‘cat’

Finally, in (23) and (24) we see some examples of the *-nu* and *-nə* markers. Both markers are almost exclusively found in the plural. The marker *-nu* always appears with lengthening of either the vowel or the consonant, and can only co-occur with either *-ŋɔ* or *-ŋu* in the singular, while the marker *-nə* can appear

<sup>22</sup>For purposes of the models, in these cases the stem was taken to be *pɔmp* or *b*, without an additional *-i*.

without lengthening in certain cases and is less restricted in terms of the singular markers it can combine with, although it tends to be pair with *-m*.

(23)	singular	plural	gloss
a.	dɔŋɔ	daanɔ	‘sticks to support a flat roof’
b.	luluŋu	lulunnu	‘perspiration’

(24)	singular	plural	gloss
a.	jazum	jazuna	‘right hand’
b.	zuŋə	zunə	‘bird’

These are the simple, straightforward number markers in Kasem. These examples show that the language allows for reversals (Baerman 2007), where pairs of markers flip their value depending on the noun. This will be one important point in the analysis.

#### 8.2.2.2 The *-ŋ*- and *-g*- markers

I now turn to less straightforward cases. Many words show a /ŋ/ segment in the singular that does not appear in the plural. Sometimes this segment is the final segment in the word, but it is mostly followed by what appears to be a regular singular marker like those discussed above. For this reason it has been claimed that the /ŋ/ is part of the singular stem, and that it tends to disappear in the plural (Callow 1965; Awedoba 1980). Thus, examples like those in (25) are analyzed as having an *-ə* marker in the singular and an *-e* marker in the plural. This, however, is no different from claiming that /ŋ/ is a singular marker which alternates with other markers for the plural, with the caveat that it can then somewhat freely combine with additional singular markers. There does not seem to be anything special about these examples that make them different from others.

(25)	singular	plural	gloss
a.	wu-saŋa	wu-sɛ	‘second flute’
b.	baya-pwəŋə	baya-pwəɛnu	‘illness where the eyes, feet and hands are swollen’
c.	bugəni-zuŋə	bugəni-zunə	‘stork’

It is then worth asking whether we are dealing with two co-occurring markers *-ŋ*- and *-ə* (in a case of multiple exponence), or if there is an additional, independent marker *-ŋə*. Looking more closely it becomes clear that *-ŋ*- can appear with

## 8.2 Cross-classifications between plural and singular: Kasem

-ə, -o and -u. Some examples are given in (26). These examples show that the marker -ηV often alternates with -nu, but not necessarily, which is evidence that these are co-occurring markers.

(26)	singular	plural	gloss
a.	nyɪŋa	nyia, nyɪ	‘horn’
b.	bwəŋə	bwe	‘adultery’
c.	lɔŋɔ	lwaanɔ	‘distance, length, surface’
d.	bulɔŋo	bulwənnu	‘liana’
e.	kuŋu	kunnu	‘Bohor Reedbook’
f.	bɔŋɔ	bənnɔ	‘root’

An additional argument against the phonological analysis that states that /ŋ/ is in the stem and gets deleted in the plural can be seen in (27), where an apparent -ηV alternates with a -ŋa marker, or an -i or -ia. Although it is hard to distinguish between both alternatives, /ŋ/ is not simply deleted in the plural.

(27)	SG	tɪtɔŋɪ	PL	tɪtɔŋa, tɪtwia	‘work, occupation’
------	----	--------	----	----------------	--------------------

The existence of the five examples in (28) makes things more complex, because here -ŋ appears as a marker on its own. As we will see later, there is a ∅ marker in Kasem, which means this could be a case of -ŋ-∅, but also simply a -ŋ final marker.

(28)	singular	plural	gloss
a.	doŋ	donnə	‘mate, fellow, friend’
b.	badoŋ	badonnə	‘friend, colleague, comrade’
c.	ciloŋ	cilonnə, ciloona	‘friend’
d.	ka-doŋ	ka-donnə	‘fellow wife’
e.	yuudoŋ	yuudonnə, yuudwənnə	‘mate, friend of same age, comrade’

A similar marker to the -ŋ- marker just discussed, is the -g- marker. Like -ŋ-, this marker can also only appear with -ə, -o and -u, and it exclusively marks singular. Some examples are given under (29).

(29)	singular	plural	gloss
a.	gar-digə	gar-di	‘mosquito net’
b.	juga	ju, je	‘place, location’
c.	pogo	pwəru	‘spider’s web’
d.	sɔgɔ	sɔm, sɔnɪ	‘knife, razor’
e.	kajugu	kajuru	‘head pad for carrying loads’

The distribution of theses -gV markers with the corresponding plural markers is also not very restricted, particularly for -gə. *Callow (1965)* also claims that this marker is a stem phoneme that undergoes a phonological deletion process.

The claim that *ŋ* and *g* are part of the stem is not well argued for in the literature. One argument in favour of this kind of analysis seems to be based on evidence from compounds like those in (30). The assumption is that singular markers cannot appear inside compounds.

(30)	singular	plural	gloss
a.	zɔŋa	zunə	‘bowl, calabash’
b.	zɔŋ-biə	zɔŋ-bi	‘calabash used for measuring’
c.	zɔŋ-diə	zɔŋ-di	‘calabash for eating food, eating bowl’

This kind of evidence is rather weak and not very systematic, however. For example, in cases like those in (31), the /g/ segment does not appear in the compounds of the noun, so one could just as well say that based on this evidence, -g- has to be a marker.

(31)	singular	plural	gloss
a.	digə	di	‘hut, room, house’
b.	di-niə	di-ni	‘married woman’s principal room’
c.	di-yuu	di-yum	‘woman’s annex room, inner kitchen in the rainy season’

Similarly, some compounds use the complete singular form of the noun, like those in (32).

(32)	singular	plural	gloss
a.	sɔŋɔ	swannɔ	‘shea-nut tree’
b.	sɔŋɔ-sabara	sɔŋɔ-sabarɪ	‘tree species’

Thus, evidence from compounds to infer stems is contradictory.

Finally, whether we should consider *-ŋ-* and *-g-* as independent markers or postulate at least six *-[+velar]V* markers seems to be a secondary issue. As a middle ground, I posit a system where *-ŋ-* and *-g-* can combine with other singular markers, while being markers on their own. Unlike the *-a*, *-o* and *-u* markers *-ŋ-* and *-g-* can combine with, *-ŋ-* and *-g-* are (almost) exclusively singular markers. In the end, however, this will not make any difference for the analogical models.

### 8.2.2.3 The *-r-* marker

A similar situation arises in the plural with the *-rV*<sup>23</sup> markers. The examples in (33) show the *-r-* marker, which almost exclusively appears in the plural (with the exception of the two words in (34)). We find *-r-* appearing mostly with *-a* and *u*, and only in a few cases with *-o*. Additionally, the *-ru* combination is found co-occurring with quite a few different singular markers.

(33)	singular	plural	gloss
a.	ba-dəgɔ	ba-dərɔ	‘sterile man’
b.	cibu-pogo	cibu-pwəru	‘chick of about one month’
c.	dudu	duduurə	‘musical instrument’
d.	tabulo	taabuloro	‘black board’

The example in (34) shows that there are at least two apparent exceptions where *-ru* appears in the singular. It is hard to know how to interpret these cases. It could be that in fact *-r-* can appear in the singular but is dispreferred, or it could be that these are special cases that require some different kind of analysis.

(34)	singular	plural	gloss
a.	barɔ	banna	‘husband, partner’
b.	kan-barɔ	kan-banna	‘husband’

### 8.2.2.4 The *-m* marker

A particularly hard case is found in the *-Vm/-nV* pairs, like those shown in (35).

(35)	singular	plural	gloss
a.	badəm	badənə	‘bachelor’
b.	banı-nyım	banı-nyına	‘disrespectful person’
c.	dəm	dəna	‘enemy’

<sup>23</sup>In earlier works it is common to find a reference to a marker *du* instead. This seems to be because /r/ and /d/ are allophones in the language. Since the source I am using uses /r/, I will use this notation.

There are several possible analyses for these examples. The more phonological one would suggest a sort of coalescence process between an /m/ segment of the stem and the *-nV* marker. Alternatively, one could argue that the fact that the sequence /mV/ is not found in singular forms suggests that the vowel is turning the /m/ into an /n/, and the fact that the final vowel of the singular is often kept in the plural strongly suggests that the stem ends in /m/, and these are examples of nouns without a singular marker. There are, however, several facts that speak against a phonological explanation. First of all, pairs like these can be found for the plural (with lower frequency, however):

(36) SG *balojana* PL *balejam* ‘Buzzard’

If these were a purely phonological process, the symmetry would be a bit suspicious. Particularly, cases like those in (37) are more in line with an *-m* marker, rather than an /m/ stem and coalescence.

(37)	singular	plural	gloss
a.	bɛɛsum	bɛɛsa	‘torment, torture, oppression’
b.	kadagum	kadagwi	‘kind of sorghum’

Although one could postulate some kind of /m/ deletion rule, this overly complicates what could be a straightforward system. This is even more clear from the perspective of the plural, especially cases with overabundance as those shown in (38).

(38)	singular	plural	gloss
a.	di-yuu	di-yum	‘woman’s annex room, inner kitchen in the rainy season’
b.	ga-sugu	ga-sum	‘wild Guinea fowl’
c.	sɔŋɔ	sum, sanɪ	‘house, compound’
d.	sugu	sum, suni	‘guinea-fowl’
e.	sɔɡɔ	sɔm, sɔni	‘knife, razor, cutlass’

These examples are strong evidence that this is not a phonological process, but rather a morphological one. I will thus consider *-m* to be a marker in its own right.

#### 8.2.2.5 The *-iə* marker

This particular marker is even harder to argue for, particularly in the light of the *-ə* marker (discussed above). For most cases, it is not completely clear whether

## 8.2 Cross-classifications between plural and singular: Kasem

we are dealing with a *-iə/-i* class, or with a *-ə/-0* class, where either the plural or singular is expressed by a zero marker. In (39) we see a couple of examples:

(39)	singular	plural	gloss
a.	manjɪ	manjɪa	‘matches’
b.	miamɪ	miamɪ	‘imported body creams/lotions’

This is especially difficult in cases where the opposing marker is an *-e*, since one could just as well postulate a phonological rule which reduces /ie/ into /e/.

(40)	singular	plural	gloss
a.	kwər-dɪa	kwər-dɛ	‘loud voice’
b.	kunku-bɪa	kunku-bɛ	‘soldier termite’

For both examples either analysis would work. The only clear evidence we have for an *-iə* marker comes from a few examples where nouns have a /iə/ in the plural and something else in the singular, or where we get a clearly different plural marker:

(41)	singular	plural	gloss
a.	dɪndwɛ	dɪndwɪa	‘dream’
b.	ga-digəbu	ga-digəbiə	‘African wild cat’
c.	kabəl-bu	kabəl-biə	‘small soup-bowl for sauce’
d.	naniə	naniinə	‘cow’

I will assume an *-iə* marker, but acknowledge that there are many cases where it is not completely straightforward, from the dictionary alone, to determine whether we are actually dealing with a *-iə* marker or a *-ə* marker.

### 8.2.2.6 The *-n* marker

Some examples like those in (42) show for both singular and plural what appears to be an *-n* marker.

(42)	singular	plural	gloss
a.	bugə-nyɔan	bugə-nywin	‘plant’
b.	gwiən	gwin	‘Yellow-billed Shrike’
c.	bɔcwən	bɔcwan	‘goat that has not yet given birth’
d.	bu-kwɪn	bu-kwɪn	‘adolescent’
e.	baŋa	bɛn	‘bracelet, bangle, metal ring’



In this case one could, as before, postulate an additional series of *-Vn* markers, or a *-n* marker which can co-occur with other singular and plural markers. Since there does not appear to be evidence that could distinguish between either hypothesis, I will assume that this is again a case of multiple exponence, but the alternative should not have any impact on the implementation of the model.

### 8.2.2.7 Three minor markers: the *-iine*, *-si* and $\emptyset$ markers

The final two segmental markers are the marker *-iine*, shown in (43), and the *-si* marker in (44).

(43)	singular	plural	gloss
a.	bar-nu	bar-niinə	‘mother-in-law’
b.	fitə-tu	fitə-tiinə	‘mechanic, fitter’

(44)	singular	plural	gloss
a.	dʊ-baga	dʊ-bagsɪ	‘thunder’
b.	ga-cawaka	ga-cawagsɪ	‘shrub species’

These two markers are infrequent and are not featured in the literature, but it seems unlikely that they could be analyzed as resulting from phonological processes.

Finally, there is a  $\emptyset$  marker. This marker is rather rare, with only 15 examples, 12 of which end in  $/[+velar]ə/$  in the singular. Of course, a ‘no marker’ alternative works equally well and makes no real difference for the analysis. A phonological explanation could work for those cases where there is a final vowel (like in (45)d.), in which one could postulate coalescence between the vowel in the stem and the marker, and thus we do not see any extra marker. But this explanation is much less likely for the examples with a consonant ending.

(45)	singular	plural	gloss
a.	kən	kɔɔna	‘Roan Antelope, Kob’
b.	kwan	kwan	‘water-lily’
c.	plan	plaanrɔ	‘plan, map’
d.	mancɪga	mancɪ	‘manioc, cassava’
e.	gar-digə	gar-di	‘mosquito net’
f.	bancɪga	bancɪ	‘manioc, cassava’

In these examples it is clear that forms like *mancɪ* or *bancɪ* have no plural marker because the singular contains them entirely, and adds some additional

marker which does not otherwise combine, or follow a vocalic marker (i.e. *-gə* does not follow an *-ɪ* marker).

### 8.2.2.8 Lengthening and diphthongization

There are two phonological processes found in Kasem which seem to mark plurality in addition to the individual segmental markers presented before. These are: lengthening of the stem and diphthongization of the last vowel of the stem.

(46)	singular	plural	gloss
a.	logo	lwəru	‘hole dug for planting seed, seed-hole’
b.	ɲwɪɔ	ɲwɪuɔ	‘wage, payment’
c.	pulu	pullu	‘granary made of straw’

In (46)b we see that the lengthening can be of the last vowel and in (46)c we see that it can be of the last consonant. This strongly speaks for a mora insertion which can either attach to the consonant or vowel. This analysis is supported by some overabundant examples where both effects are found. In (47) we see that this phenomenon is even independent of the additional segmental plural marker chosen.

(47)	singular	plural	gloss
a.	cɔɔɔ	cɔɔɔɔ, cɔɔɔɔ	‘black make-up’
b.	vɔɔɔ	vannɪ, vaanɔ	‘hoe’

Especially interesting are the cases where both processes (i.e. lengthening and diphthongization) occur on the same word as shown in (48).

(48)	singular	plural	gloss
a.	bugə-kanyɔɔ	bugə-kanywannɔ	‘kind of tree’
b.	yolo	ywəllu	‘empty area / field, empty space outside village’
c.	cɔɔ	cwaalɔ	‘girl that likes going out with men’
d.	war-boro	war-bwəəru	‘brick mould / mold’

### 8.2.2.9 Other stem changes

Some nouns show some sort of unpredictable stem changes, mostly in velar segments as seen in (49)

## 8 Complex inflectional classes

(49)	singular	plural	gloss
a.	coro	ceeni, ceenu	‘hen, fowl, chicken’
b.	boŋo	bənnu	‘dung, shit’
c.	biboku	bibəgəru	‘stutterer’
d.	cɪkɔ	cɪgɪɔ	‘trap’
e.	cɪɔgɔ	cɪɔɪɔ	‘feather of fowls’

I do not consider suppletion among the classes for the analogical model, but in principle this could also be a dimension of noun inflection.

### 8.2.2.10 Compounds

For most compounds, the only part that changes is the rightmost (the adjective). There are, however, exceptions with compounds with the word *kandwɛ* (‘stone’), among some others as in (50).

(50)	singular	plural	gloss
a.	kandwɛ-nyuni	kandwa-nyuna	‘bright / shiny stone’
b.	kandwɛ-ŋɔni	kandwa-ŋɔna	‘precious / bright stone, jewels, pearl.’
c.	kandwɛ-pɪsuni	kandwa-pɪsuna	‘pile / heap of stones’
d.	kandwɛ-pɔlɔɔ	kandwa-palwaarɔ	‘rock’
e.	kandwɛ-pɔpɔɔ	kandwa-pɔpɔrrɔ	‘stone bracelet’
f.	kunkwən-poŋo	kunkwəŋ-pwəənu	‘Red-eyed Dove, collared dove’

I will leave this case as an open problem since the data are not conclusive as to why some compounds can inflect for their head noun and others do not.

### 8.2.3 Materials

The dataset, as well as all examples cited here, come from the Kasem Burkina Faso Dictionary (Niggli & Niggli 2007) in its online version.<sup>24</sup> The dictionary lists for each noun its singular and plural forms, as well as the tones for the singular form. The tones for the plural form are only listed in a few exceptional cases, which seems to suggest that the plural and singular forms have the same tones. This, however, is hard to extrapolate to words where the plural is longer or shorter than the singular. From 2000 nouns listed in the dictionary, I removed

<sup>24</sup>[kassem-bf.webonary.org/](http://kassem-bf.webonary.org/), visited on 10-11-2016.

30 cases where either the marker was completely unclear, the plural showed unpredictable suppletion, or where there was reason to suspect an error (i.e. nouns where the ATR feature did not match across all their vowels, etc.), and ended up with a total of 1970 nouns.

For the two nouns in (51) the dictionary presented an alternative in the singular. For both these cases I only considered the main form.

- (51) a. kwɪan (kwɛ) ‘Stripped Ground Squirrel’  
b. sɛ (swɛ) ‘ivory bracelet’

In the cases of polysemy I left all entries in the table:

- (52) a. ni ‘opening of a room/house, gate’  
b. ni ‘mouth, beak’  
c. ...

As we have seen in multiple examples already, Kasem, just like Hausa, presents some overabundance in the plural forms:

(53)	singular	plural	gloss
a.	bwana	bwani, bwam	‘mosquito’
b.	bɔŋɔ	bɔni, bɔm	‘goat’

In all these cases I only considered the first plural listed. The reason is that the dictionary only lists 108 nouns with overabundant plurals. This is not enough to be able to reliably model overabundance in this case.

For roughly half of the nouns, the dictionary included a semantic annotation which consists of some basic groupings like ‘animal’, ‘human’, ‘animate’, etc., coded with numbers. I use this semantic annotation in the analogical models. As for the nouns without semantic coding, I assigned them to a default class.

#### 8.2.4 Modelling the system

After the previous discussion it is useful to look at the pairings between segmental singular and plural markers. Table 8.18 shows the number of nouns for which a given pairing holds (ignoring overabundant cases), after neutralizing ATR. The table also ignores lengthening and diphthongization. Table 8.19 shows the co-occurrences of plural markers with either lengthening of the vowel (VV), the consonant (CC), and with the presence or absence of diphthongization.

Table 8.18: Co-occurrence of singular and plural markers.

Plural																					
Singular	0	e	ə	en	ən	i	iə	iinə	in	m	n	ne	nə	ni	nu	o	rə	ro	ru	si	u
0	2	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
e	0	1	48	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ə	3	28	33	0	0	236	0	0	0	0	0	0	2	2	0	0	0	0	41	10	2
en	0	0	1	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ən	0	1	0	0	1	0	0	0	7	0	1	0	0	0	0	0	0	0	3	0	0
gə	19	20	1	0	0	17	0	0	0	0	0	1	0	0	0	0	1	0	2	1	0
go	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	53	0	0
gu	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	34	0	0
i	0	0	321	0	0	9	27	0	0	0	0	0	1	0	0	0	0	0	1	0	0
iə	0	14	0	0	0	32	4	3	1	0	0	0	1	2	0	0	3	0	1	2	0
in	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0
m	0	0	12	0	0	1	0	0	0	1	0	0	64	0	0	0	0	0	0	0	0
nə	0	0	0	0	0	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	0
ŋ	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0
ŋə	8	42	1	9	0	9	0	0	9	0	1	0	7	0	2	0	0	0	0	1	0
ŋi	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ŋo	1	0	0	0	0	0	0	0	0	1	0	0	3	0	91	0	0	0	0	0	0
ŋu	0	0	0	0	0	0	0	0	0	0	0	0	1	1	35	0	0	0	0	0	0
no	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
o	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
on	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	5	0	1	42	0	137
ru	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
u	1	1	160	0	0	0	0	12	0	1	0	0	2	0	0	0	0	0	0	0	0
un	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0

Table 8.19: Co-occurrence of plural markers with lengthening and diphthongization.

		Plural marker																			
	0	e	ə	en	ən	i	iə	iinə	in	m	n	ne	nə	ni	nu	o	rə	ro	ru	si	u
no-lengthening	32	106	411	9	9	297	73	0	17	30	2	1	74	4	0	0	4	2	107	14	8
CC-lengthening	0	0	6	0	0	0	0	0	0	0	0	0	8	1	48	5	0	0	1	0	72
VV-lengthening	2	1	175	0	0	7	1	15	0	0	0	0	6	10	81	0	1	1	154	0	137
no-diphthongization	33	101	565	9	9	277	74	15	17	29	2	1	84	15	39	5	5	3	192	14	83
diphthongization	1	6	27	0	0	27	0	0	0	1	0	0	4	0	90	0	0	0	70	0	134

Table 8.20: Co-occurrence of lengthening and diphthongization.

	diphthongization	no-diphthongization
CC-lengthening	24	117
no-lengthening	108	1092
VV-lengthening	228	362

If we cross-classify all factors the result are 144 nonempty classes (ignoring ATR), with most classes having less than 50 members, and 63 classes of only 1 member. Because of this, a flat list of inflection classes looks particularly unconvincing. A more straightforward approach is to use cross-classification as with the Spanish systems.

To model the complete space of inflectional classes several trees are required. The first thing we have to recognize is that markers like *-i*, are not in themselves plural or singular markers, but simply number markers. Whether they indicate plural or singular depends on their distribution with other markers. There are two alternatives at this point, either overspecification as in Figure 8.8, or underspecification as in Figure 8.9.

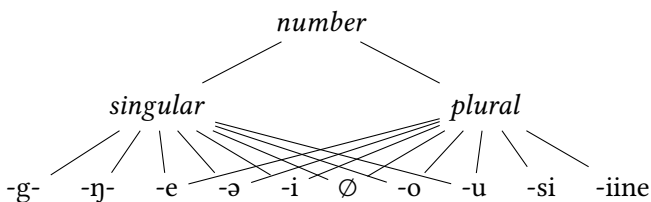


Figure 8.8: Kasem number markers with overspecification.

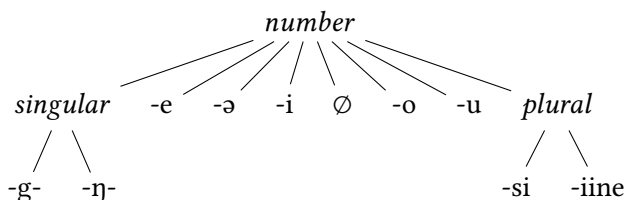


Figure 8.9: Kasem number markers with underspecification.

For the purposes of this study either alternative would work equally well. For simplicity I will go with the underspecification approach in Figure 8.9.

Lengthening and diphthongization are processes which are completely independent of the segmental markers, but from Table 8.19 it should be clear that the distribution of plural markers is not random with regards to the classes they co-occur with. Both are much more likely with *-u* markers, and lengthening of the vowel is also very likely with *-a*. Similarly, we see that while *-ru* is very likely to co-occur with lengthening of the vowel, it only co-occurs once with lengthening of the consonant, as shown in (54).

(54) SG *ɲwam-pɔ̃gɔ* PL *ɲwam-pɔ̃rrɔ* ‘scale of wound’

Similarly, as can be seen in Table 8.20, the proportion of words with no lengthening in the plural but diphthongization is around 10%, while that of CC-lengthening and diphthongization is around 20%, and the proportion of nouns with diphthongization and VV-lengthening is of almost 40%. These are clearly not random distributions<sup>25</sup>. What this means is that our model for cross-inheritance should consider all four factors: segmental markers of the singular, segmental markers of the plural, lengthening and diphthongization.

Because lengthening and diphthongization only occur on the stem, these two dimensions can also be modelled with a stem space. For this, we have to postulate that Kasem nouns have a singular and a plural stem. Alternatively, nonconcatenative morphological processes could also be used to account for these changes. In the end, the important thing is that all nouns must be specified for whether they undergo these processes or not. The partial trees for lengthening and diphthongization can be trivially defined as in Figure 8.10 and Figure 8.11.

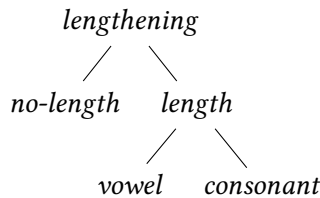


Figure 8.10: Hierarchy for lengthening in Kasem.

Figure 8.12 shows a partial hierarchy with all dimensions of Kasem noun inflection class. Segmental markers constitute a hierarchy of their own, which specifies

<sup>25</sup>I skip statistical tests here because I will show this is the case with the models in the next section.



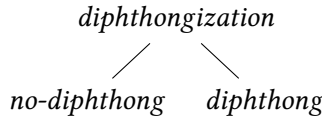


Figure 8.11: Hierarchy for diphthongization in Kasem.

which markers combine with which other markers. Underspecified markers can mark either singular or plural, and the combination of two of these underspecified markers means that both alternatives are available<sup>26</sup>. The complete inflection class of a noun is given by the *sg-pl-diphth-length*.

Every noun in Kasem must be typed for its complete inflection class. In Figure 8.12 the lexeme *alapl* (‘aeroplane’) belongs to class *i-ə-ndiphth-nl*, which means it takes an *i* in the singular, a *ə* in the plural, and its stem does not undergo diphthongization or lengthening. How different theories chose to realize these properties, is an independent problem.

## 8.2.5 Methodological considerations

### 8.2.5.1 Predictability between subtrees

In several of the models below, when predicting a subtree (e.g. *lengthening*), I will include information from another subtree (e.g. *diphthongization*). From a theoretical perspective, this works in a different way than the stem information. Adding information about a cross-classifying tree is equivalent to removing a subset of the possible classes. In the toy example in Figure 8.13, two subtrees,  $\tau$  and  $\sigma$ , cross-classify to build the inflection classes for the lexemes  $w_1$  to  $w_9$ . If an analogical model predicting  $\tau$  for the words  $w_1$  to  $w_9$ , knows  $\sigma$ , it will not have to decide between three classes, but at most two. For words  $w_7$  to  $w_9$ , the type *s2* uniquely determines that these words belong to type *t3*, because it removes the possibility that these words could belong to either *t1* or *t2*. For words  $w_1$  to  $w_6$ , the type *s1* removes the possibility of *t3*.

### 8.2.5.2 Compounds

We now turn to the analogical modelling. A difficult decision regarding this particular dataset is whether to include compounds or not. Including them means

<sup>26</sup>It is however unclear if for all combinations of underspecified markers reversals are found. In other words, if  $x$  and  $y$  are underspecified, it is not clear whether  $x$ - $y$  and  $y$ - $x$  necessarily exist, or that it could exist.

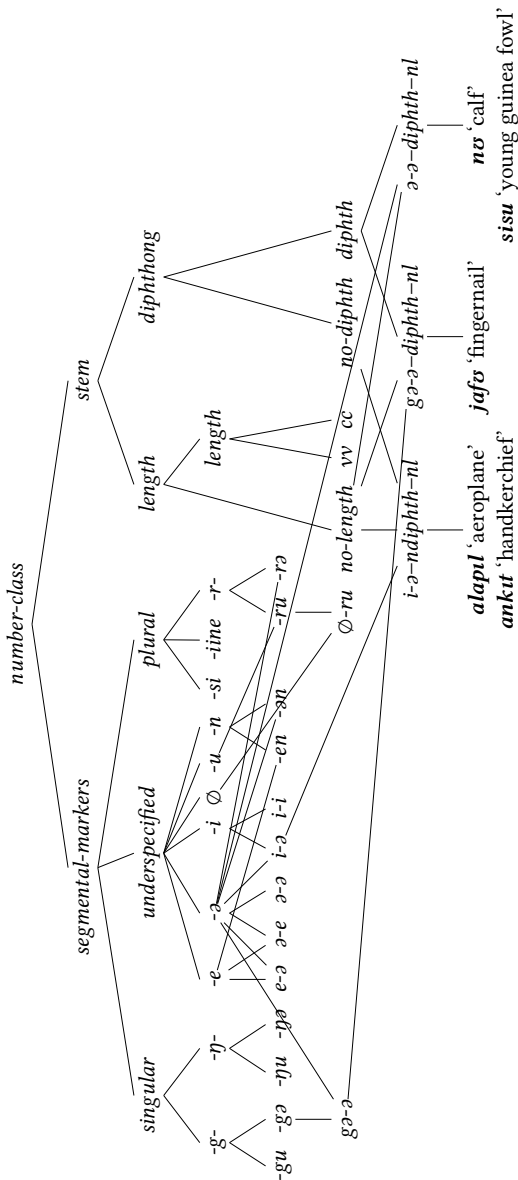


Figure 8.12: Partial inflection class hierarchy for Kasem.

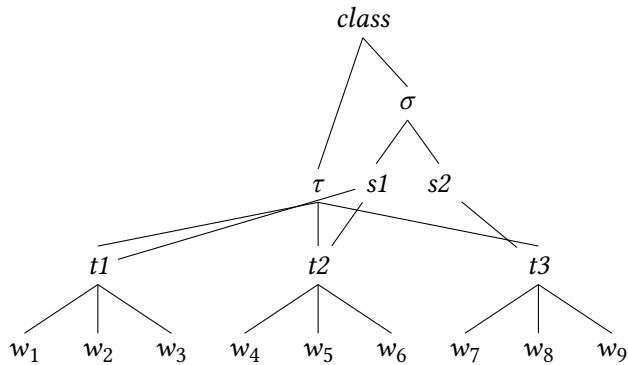


Figure 8.13: Example of cross-classifications and information.

that, because compounds usually have the same plural marker as the simplex noun, the model will be able to remember some cases. That is, the cross-validation is not completely perfect. On the other hand, not all compounds share the same plural marker as their simplex form. Additionally, it is not always clear what sort of compounds we are actually dealing with. Some seem semantically transparent like those in (55)a and (55)b, but others less so like those in (55)f and (55)g.

(55)	singular	plural	gloss
a.	baɟa	bɛn	‘bracelet, bangle, metal ring’
b.	kalum-baɟa	kalum-bɛ	‘black bracelet (for rites)’
c.	nyasaŋ-biə	nyasaŋ-biə	‘sesame seeds’
d.	zɔŋ-biə	zɔŋ-bi	‘calabash used for measuring’
e.	bɔŋɔ	bɔni, bɔm	‘goat’
f.	bɔŋɔ	bɔnnɔ	‘root’
g.	ŋwan-bɔŋɔ	ŋwan-bɔnnɔ	‘capillary’

Finally, not all words marked as compounds in the dictionary have a corresponding simplex form:

(56)	singular	plural	gloss
a.	kaləŋ-jarɔ	kaləŋ-jara	‘fisherman’
b.	wo-jaanɔ	wo-jaana	‘bird, insect’
c.	kamɔ-mɔrɔ	kamɔ-mɔra	‘potter’
d.	*jarɔ		
e.	*jaanɔ		
f.	*mɔrɔ		

There are only around 200 nouns which appear multiple times because they are present as simplex forms and compounds. One could still remove them from the dataset, considering the examples in (55) and (56), we see that compounds do not guarantee consistent plural endings, and do not guarantee a simplex forms. With this in mind, leaving the compounds in is not much too different from having items where the last three or four segments are identical. We would not remove these cases, since these are the core of what the analogical process is. Similarly, that compounds tend to belong to the same class as the simplex form, seems to also be a product of the same principles. Finally, from a more cognitive perspective, the fact that there are many lexical entries with the same stem simply means that there are more chances to memorize that form. In any case, it seems more realistic to leave the compounds in.

### 8.2.6 Results

The dataset extracted from the dictionary had 1970 nouns. Considering all these nouns, the total number of classes (disregarding lengthening and diphthongization) was 98, with 48 classes having one or two members. Although possible in theory, in practical terms it is very difficult to fit and evaluated models with this kind of distribution. On the one hand, it is impractical because there are just not enough training data for most classes, and on the other hand, errors in the very low frequency classes will unfairly penalize the model's performance. For this reason I removed all items that belong to a class with a type frequency of 8 or less. The final dataset contains a total of 1792 nouns, distributed across 33 classes. This leaves us with a system that has more classes than any of the other examples discussed in this book.

The predictors are: the last three segments of the singular stem (computed as the singular without the singular marker), the semantic annotation in the dictionary, the lengthening process (C lengthening, V lengthening, or none), the diphthongization process (none or present), the singular marker and the plural marker. As mentioned above, because ATR is a stem feature, I neutralized it for all stems. The length (in letters) of the stem and the tones of the singular form did not play any role in the models.

Because of its complexity, I will present several different models that tackle different parts of the system. The following sections describe the results for each such model. I will only look at clustering of the results for the last model predicting inflectional class. There are many more possible combinations I did not test, but the most important aspects of the system are covered.

## 8.2.6.1 Predicting diphthongization

The first case we look at is diphthongization in the plural. Since it is a binary choice, this is the simplest of the models for Kasem. The basic model (not including number markers) was: `diphthong ~ final.1 + final.2 + final.3 + meaning`<sup>27</sup>. Table 8.21 presents the results with the corresponding accuracy scores in Table 8.22.

Table 8.21: Confusion matrix for the model predicting diphthongization without segmental number markers in Kasem.

Reference		
Prediction	dp	Ndp
dp	267	66
Ndp	79	1380

Table 8.22: Accuracy scores for Table 8.21.

Overall Statistics	
Accuracy :	0.9191
95% CI :	(0.9055, 0.9313)
No Information Rate :	0.8069
Kappa :	0.7366

Table 8.22 shows that the model has a very good accuracy and kappa scores to start with. This shows that diphthongization is highly predictable. Next we test to see whether adding both number markers helps the model. We refit the analogical model with the formula: `diphthong ~ final.1 + final.2 + final.3 + lengthening + meaning + pl + sg`. The results can be seen in Table 8.23, and the corresponding accuracy values in Table 8.24.

The overall evaluation is shown in Figure 8.14. There are several important observations. First of all, lengthening and meaning do not seem to play any role in the model when the other factors are considered. The final segment of the stem was the most predictive segment, and remained relevant even after adding

<sup>27</sup>For all Kasem models the networks only included a skip layer and no hidden layers, with a decay rate of 0.01.

## 8.2 Cross-classifications between plural and singular: Kasem

Table 8.23: Confusion matrix for the model predicting diphthongization with segmental number markers in Kasem.

Reference		
Prediction	dp	Ndp
dp	303	46
Ndp	43	1400


Table 8.24: Accuracy scores for Table 8.23.

Overall Statistics	
Accuracy :	0.9503
95% CI :	(0.9392, 0.9599)
No Information Rate :	0.8069
Kappa :	0.8411

both number markers. The other two segments seem to be somewhat redundant with the number markers, even though they played a role on their own. This is to be expected if there is a strong correlation between final segments and number markers. However, the fact that the `final.1` was highly predictive even after adding the number marker, means that it is contributing to the analogical model independently of its predictive power of the segmental number markers. Finally, the singular marker was more predictive than the plural marker. This will be a recurring theme in this section: it is easier to predict plural markers (including lengthening and diphthongization) from the singular markers, than from other plural markers, and the other way around. There is no obvious explanation for this phenomenon. A possible reason is that the task of predicting a given plural marker usually follows from knowing the singular, and not from knowing other co-occurring plural markers.

### 8.2.6.2 Predicting lengthening

The second feature in degree of complexity is the lengthening (or mora insertion) in the plural. In this case we are dealing with a three way choice: no lengthening (NC), consonant lengthening (CC) and vowel lengthening (VV). The best model (not including segmental number markers) was: `lengthening ~ final.1`



./figures/kasem/p-fi-dp-sg-overall.pdf

Figure 8.14: Additive (left) and subtractive (right) accuracy and kappa scores for the model predicting diphthongization with segmental number markers in Kasem.

Table 8.25: Confusion matrix for the model predicting lengthening without segmental number markers in Kasem.

Reference			
Prediction	CC	NL	VV
CC	49	35	5
NL	58	979	137
VV	18	100	411

+ final.2 + final.3. The results of this model can be seen in Table 8.25 and the corresponding statistics in Table 8.26.

This model is, once more, already quite good. The type of lengthening a stem undergoes is highly predictable from its shape alone. In this case the semantics did not play any role. Next, we fit a model that includes all other number classes as predictors  $\text{lengthening} \sim \text{final.1} + \text{final.2} + \text{final.3} + \text{diphthong} + \text{pl} + \text{sg}$ . Results for this model can be seen in Table 8.27 and the corresponding statistics in Table 8.28.

## 8.2 Cross-classifications between plural and singular: Kasem

Table 8.26: Accuracy scores for Table 8.25.

Overall Statistics			
Accuracy : 0.803			
95% CI : (0.7838, 0.8212)			
No Information Rate : 0.6217			
Kappa : 0.6046			
Statistics by Class:			
	Class: CC	Class: NL	Class: VV
Sensitivity	0.392	0.879	0.743
Specificity	0.976	0.712	0.905
Neg Pred Value	0.955	0.782	0.888
Balanced Accuracy	0.684	0.796	0.824

Table 8.27: Confusion matrix for the model predicting lengthening without segmental number markers in Kasem.

Reference			
Prediction	CC	NL	VV
CC	103	7	11
NL	4	1076	33
VV	18	31	509

Table 8.28: Accuracy scores for Table 8.27.

Overall Statistics			
Accuracy : 0.942			
95% CI : (0.9301, 0.9523)			
No Information Rate : 0.6217			
Kappa : 0.8869			
Statistics by Class:			
	Class: CC	Class: NL	Class: VV
Sensitivity	0.824	0.966	0.920
Specificity	0.989	0.945	0.961
Neg Pred Value	0.987	0.944	0.964
Balanced Accuracy	0.907	0.956	0.940



The overall evaluation is shown in Figure 8.15. This table presents a more dramatic increase in both kappa and accuracy after adding the segmental number markers. In this case both the singular and plural segmental markers had a very similar importance. More interesting, however, is the fact that in this case we see the opposite effect in the final three segments of the stem. In the previous case of predicting diphthongization, only the final segment was independently predictive of the outcome, here the penultimate and antepenultimate segments are both independently predictive of the lengthening. This again goes to show that different subtrees in the hierarchy have their own analogical relations for their members. Finally, it is worth noting that when predicting diphthongization there was no effect from adding lengthening as a predictor, and here there is no effect from adding diphthong as a predictor. What this suggests is that the correlations described before are already being captured by the final segments. This is the first indication that there is heavy redundancy in the system. I will come back to this in the following sections.

`./figures/kasem/p-fi-length-sg-overall.pdf`

Figure 8.15: Additive (left) and subtractive (right) accuracy and kappa scores for for the model predicting diphthongization with segmental number markers in Kasem.

### 8.2.6.3 Predicting singular markers

We now turn to predicting the singular marker of a word. Because I will be discussing many different models of related phenomena it would be tedious to

present confusion matrices or heat maps for each of them. For this reason, I will only present the basic accuracy measures for model comparison. In the last section I will present the heat maps of the final models.

In the first model we are looking at the bare effects of the final segments and meaning of the stems:  $\text{singular} \sim \text{final.1} + \text{final.2} + \text{final.3} + \text{meaning}$ . This model tries to predict total of 14 different markers: *e, iə, i, u, ə, o, gu, ɲo, m, na, go, gə, ɲə, ɲu*. The accuracy scores are shown in Table 8.29.

Table 8.29: Accuracy scores for the model predicting the singular marker from the stem information only.

Overall Statistics	
Accuracy :	0.5709
95% CI :	(0.5476, 0.5939)
No Information Rate :	0.2037
Kappa :	0.5003

This model shows very good performance, especially considering the relatively large number of classes it is predicting. This works as the initial baseline of comparison. The next step is to include the plural marker as a predictor:  $\text{singular} \sim \text{final.1} + \text{final.2} + \text{final.3} + \text{meaning} + \text{pl}$ <sup>28</sup>. The accuracy scores are in Table 8.30.

Table 8.30: Accuracy scores for the model predicting the singular marker from the stem and plural marker information.

Overall Statistics	
Accuracy :	0.8186
95% CI :	(0.8, 0.8362)
No Information Rate :	0.2037
Kappa :	0.7889

The results in Table 8.30 show that there is a considerable gain from including the plural marker in the model. For comparison, using only the plural marker:  $\text{singular} \sim \text{pl}$  produces the results in Table 8.31.

<sup>28</sup>The reason for not using the plural stem in these cases is that the plural stem follows directly from knowing the singular stem plus the dimensions of diphthongization and lengthening.

## 8 Complex inflectional classes

Table 8.31: Accuracy scores for the model predicting the singular marker from the plural marker information only.

Overall Statistics	
Accuracy :	0.6077
95% CI :	(0.5847, 0.6304)
No Information Rate :	0.2037
Kappa :	0.5348

It should then be clear that although the effect of knowing the plural marker is considerable, it is even better when the model knows the shape of the singular stem. The overall results are shown in Figure 8.16, and the heat map for the model using only stem information is in Figure 8.17.

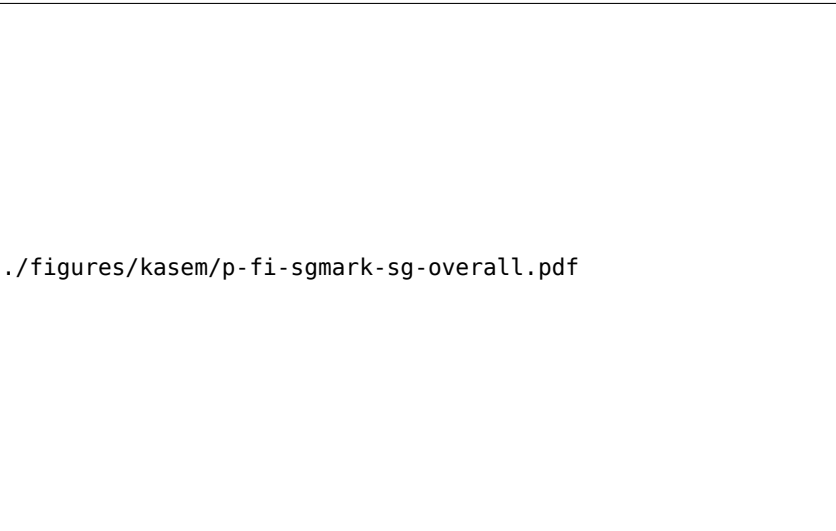


Figure 8.16: Additive (left) and subtractive (right) accuracy and kappa scores for for the model predicting singular from the singular from the stem and plural information.

### 8.2.6.4 Predicting plural markers

We now try to predict the plural marker of a noun. In this case the predicted classes are: *a, i, ru, u, iə, nu, e, nə, m, 0, si, en, iinə, in*. We first look at the basic

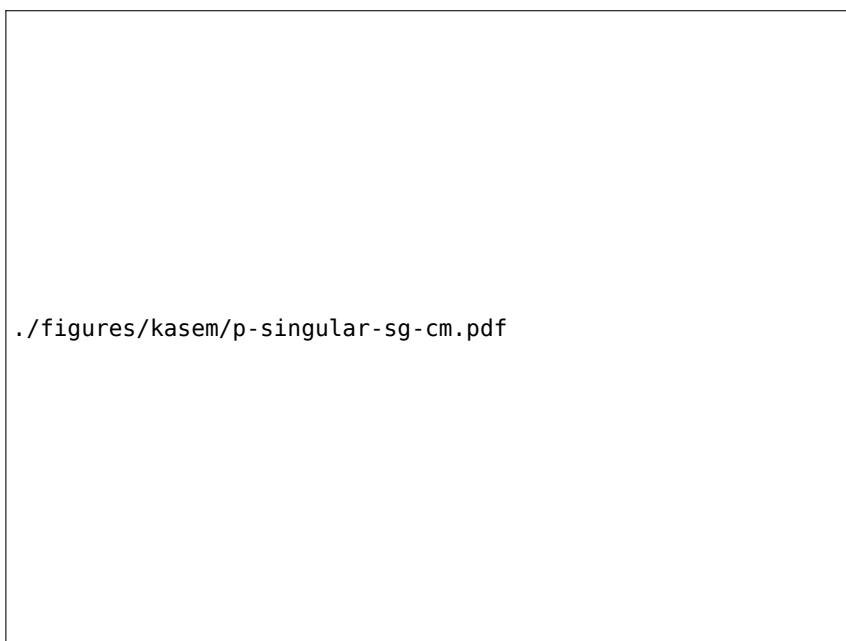


Figure 8.17: Heat map for the models predicting the singular marker from the stem information only.

Table 8.32: Accuracy scores for the model predicting the plural marker from the stem information only.

Overall Statistics	
Accuracy :	0.6345
95% CI :	(0.6117, 0.6568)
No Information Rate :	0.3265
Kappa :	0.5528

model with only the final segments and meaning of the stem:  $\text{plural} \sim \text{final.1} + \text{final.2} + \text{final.3} + \text{meaning}$ . The accuracy results are in Table 8.32.

Next, we test the effect of adding the singular marker:  $\text{plural} \sim \text{final.1} + \text{final.2} + \text{final.3} + \text{meaning} + \text{sg}$ . The results of this model are in Table 8.33.

Table 8.33 shows that the plural marker is more predictable than the singular marker. A possible simple explanation is that it is more common that one would want to predict the plural of a noun from knowing its singular form, than wanting

Table 8.33: Accuracy scores for the model predicting the plural marker from the stem and singular marker.

Overall Statistics	
Accuracy :	0.8867
95% CI :	(0.8711, 0.901)
No Information Rate :	0.3265
Kappa :	0.8615

to predict the singular form of a noun from knowing its plural. A very similar situation arises if we try to predict the plural marker from the singular marker alone: plural  $\sim$  sg. The results are in Table 8.34.

Table 8.34: Accuracy scores for the model predicting the plural marker from the singular marker information only.

Overall Statistics	
Accuracy :	0.7204
95% CI :	(0.699, 0.7411)
No Information Rate :	0.3265
Kappa :	0.6468

These results show a greater symmetry in the implicational relations. The overall results and evaluation can be seen in Figure 8.18, and the heat map for the model using only the stem is in Figure 8.19.

#### 8.2.6.5 Predicting class

Finally, we want to put these things together and predict inflectional class (defined as the combination of a singular and a plural marker). So far I did not include diphthongization and lengthening as part of the inflectional class. Doing so would result in too many labels, which the model would have a very hard time predicting. Additionally, as seen when predicting diphthongization and lengthening, both these sub-trees are fairly predictable from the same factors<sup>29</sup>. I will instead use both factors (diphthongization and lengthening) as predictors of class.

<sup>29</sup>This has the additional problem that it burdens the analogical model, since the factors will be doing multiple jobs at the same time.

./figures/kasem/p-fi-plmark-sg-overall.pdf

Figure 8.18: Additive (left) and subtractive (right) accuracy and kappa scores for for the model predicting plural from the singular stem in Kasem.

./figures/kasem/p-plural-sg-cm.pdf

Figure 8.19: Heat map for the models predicting the plural marker from the stem information only.

## 8 Complex inflectional classes

As before, there is no real limit to possible combinations of factors and classes one can test.

First we predict from the stem with a basic model that only looks at the ending and meaning of the stem: `class ~ final.1 + final.2 + final.3 + meaning`. The results are in Table 8.35 and its corresponding heat map in Figure 8.20

Table 8.35: Accuracy scores for the model predicting inflection class from the stem only.

Overall Statistics	
Accuracy :	0.5335
95% CI :	(0.5101, 0.5568)
No Information Rate :	0.1791
Kappa :	0.4928

./figures/kasem/p-class-sg-cm.pdf

Figure 8.20: Heat maps for the models predicting inflection from the stem only.

Including lengthening and diphthong as predictors with the formula: `class ~ final.1 + final.2 + final.3 + lengthening + diphthong + meaning`, produces a clear improvement. The results can be seen in Table 8.36, the corre-

sponding heat map can be seen in Figure 8.21, and the overall evaluation in Figure 8.22.

Table 8.36: Accuracy scores for the model predicting the plural marker from the singular marker information only.

Overall Statistics	
Accuracy :	0.6596
95% CI :	(0.6371, 0.6815)
No Information Rate :	0.1791
Kappa :	0.6303

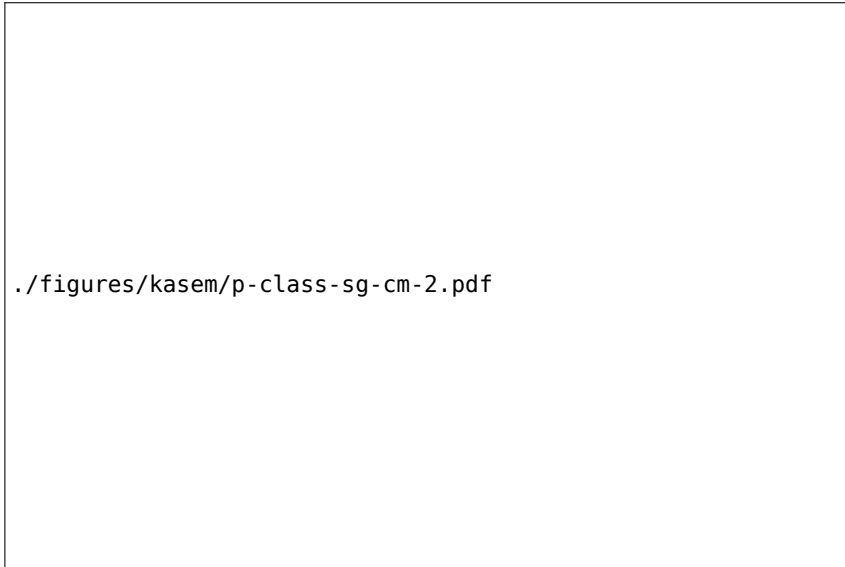


Figure 8.21: Heat maps for the models predicting inflection class from the stem, and lengthening and diphthongization information.

In this case it is also useful to look at the balanced by-class accuracy of the model. That is, we can look at how each level of the response variable (each inflectional class) increases or decreases in accuracy as we add or subtract factors. These results are shown in Figure 8.23. The interesting point here is that different classes are not equally predictable. What this means is that there is not an homogeneous increase in the class accuracy. Instead, some classes like *o-u* or *e-a*



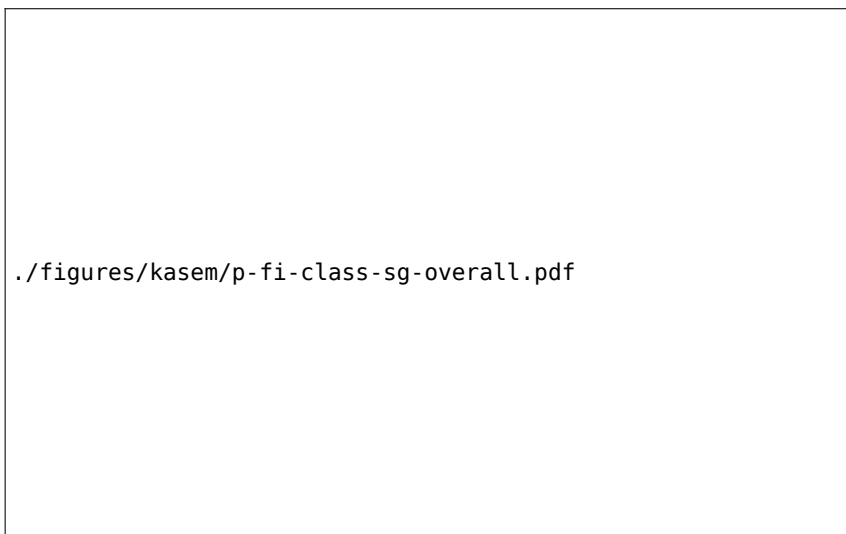


Figure 8.22: Additive (left) and subtractive (right) accuracy and kappa scores for for the model predicting inflection class.

achieve a very high balanced accuracy with the use of just one predictor, while classes like  $\partial\text{-}\partial$  and  $i\partial\text{-}e$  remain quite unpredictable all the way through. This indicates that class predictability is not symmetric, and that different classes focus on different parts of the stem.

Finally, the clustering created by this model<sup>30</sup> presents several crucial results. Like in Spanish, this is the most interesting aspect of the models. The first thing we can observe is that the larger (color coded) clusters are not homogenous with respect to the features that seem to define them. There are several important clusters to look at here. On the left top corner, in dark green, we find an inversion  $-i/\text{-}\partial - \text{-}\partial/-i$ , next to  $-u/\text{-}\partial$  which fits the general pattern of an  $\text{-}\partial$  with a high vowel. To the right, and around the  $-0.5$  X axis, we find three classes:  $\text{-}\partial/\text{-}\partial$ ,  $-i/-i$  and  $-u/-u$ . The first two are close to each other and clustered together, while the last class is clustered separate from the other two, but it is placed quite close to them on the map.

Close, and tightly grouped together, we find two clusters, one in dark blue and one in light lilac. These two clusters all share an  $-i\partial$  marker, except for one which only has a  $\text{-}\partial$  marker. In dark blue we see an inversion between  $-i\partial/-i$  and  $-i/-i\partial$ , and in light lilac a partial inversion of  $-i\partial$  marking singular and plural. The next

<sup>30</sup> As before, we fit a direct similarity model instead of relying on the errors of the analogical model.

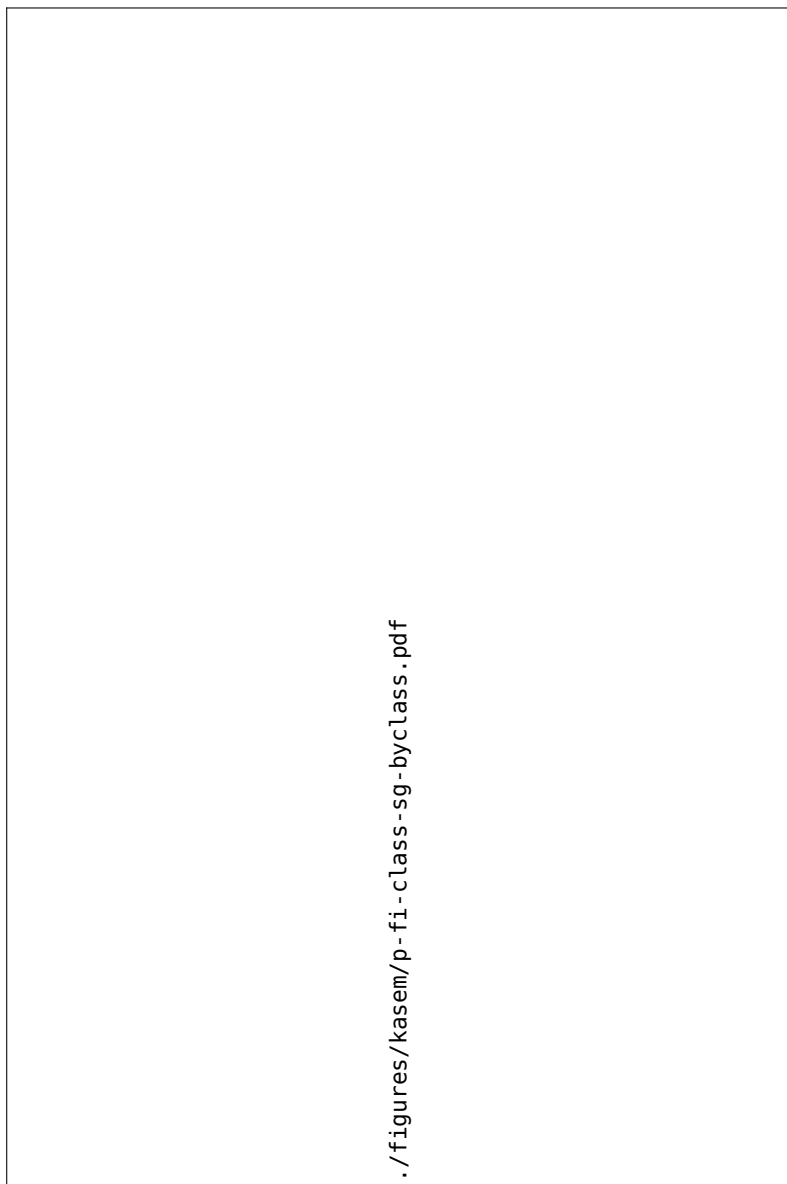


Figure 8.23: Additive balanced accuracy (by class) for the model predicting inflection class.

color clusters are less well organized from a perspective of a potential hierarchy, but from their position they make sense. On the lower right corner we see three classes that share an *-o* in the singular and *-u* in the plural, with some additional *-g-*, *-r-* and *-ŋ-*. Right at the 0.5 X and -0.25 Y we find other two classes with a *-ru* marking plural (again, close to the *-o/-ru* and *-go/-ru* classes).

Right at the center of the map we see three classes: *-u/-iinə*, *-m/-ə* and *-ə/-si*. These classes only share the *-ə* marker (or /ə/ segment in the case of *-innə*), but they have in common that they have one marker not shared by any other class. At the same X coordinate, but at around 0.5 Y, we have two close classes having a *-[+velar]ə* marker for the singular and *-i* in the plural, and not too far off we have the very similar *-ŋə/-in* class (arguably the class *-gə/-Ø* is also related to these three classes). A class that seems somewhat out of place is the *-gu/-ru* class, also in dark orange. Finally, in the top right corner we have two groups. In light blue we have classes with *-ə/-e* plus additional markers, and in dark lilac we have the inversion *-nə/-m* – *-m/-nə*.

A second important result that can be observe in this clustering is that the presence or absence of *ŋ*, *n*, *r*, *s* and *m* markers is not random on the map. All these markers only appear with positive values on the X axis. Similarly, most velar markers are in the upper right quadrant. What this indicates is that these markers cluster independently of the vocalic markers, lending some evidence to the hypothesis that each subtree in the hierarchy has its own analogical function.

Important for the sketch of the system presented above is that for most classes their position on the plane depends more on the vowel presence or combinations, than on what they mark. That is, *-x/-y* classes are close to other classes with either *-x* or *-y* present, independently of whether *-x* and *-y* are marking the same number. This is exactly what the hierarchy suggested would predict.

Finally, because of the complexity of the system, we can test whether there are extra similarity dimensions we are missing in this MDS plot. To do this, we extract three main components of the similarity matrix instead of two, and plot them side by side. This is similar to looking at a cube from three of its faces. In the plots in Figure 8.25, X is the first component, Y the second and Z the third.

The XY plot shows the same map as before for comparison. The most interesting effect is found in the ZY plot. Here a strong grouping of the classes across vocalic lines appears. Classes with /o/ and /u/ are mostly on the lower quadrants, and classes with /ə/ and /i/ tend to be higher. Particularly interesting is the repositioning of *-ə/-i* to the right quadrant, closer to other classes with the same sequence of vocalic markers. The XZ plot is less interesting, but it shows a much stronger separation of the purely vocalic class from classes with multi-

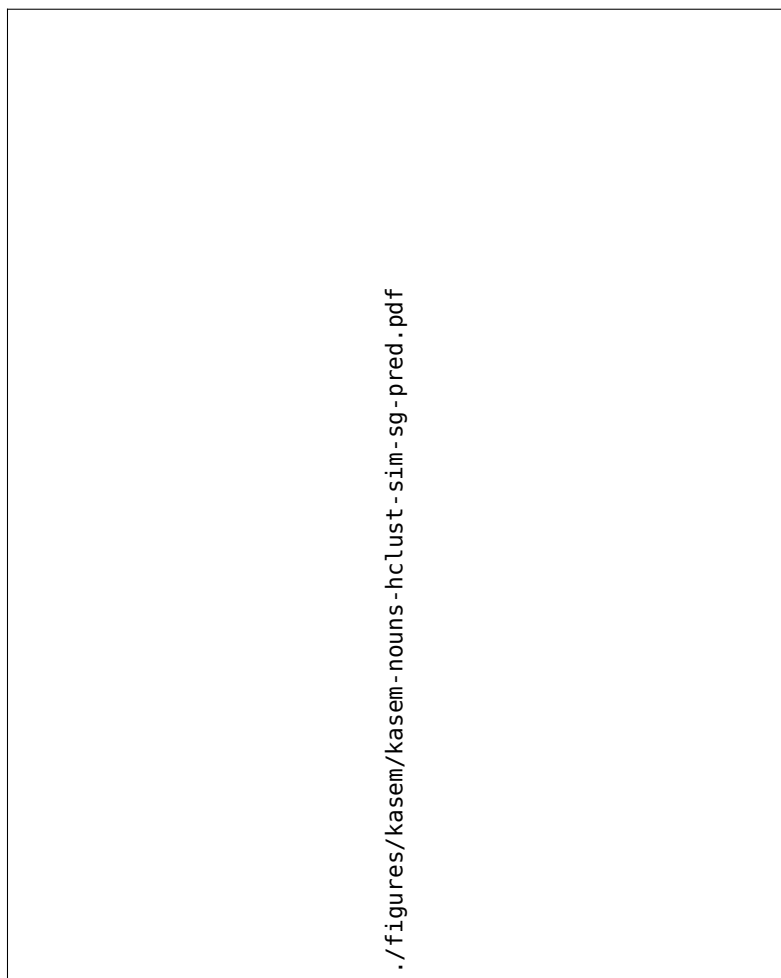


Figure 8.24: Clustering of inflection class in Kasem based on the singular stem, lengthening and diphthongization.

ple exponents. Although the evidence is somewhat weaker, we see that different similarity dimensions capture what seems to be different aspects of the hierarchy.

What this decomposition shows is that the grouping effects between the classes go beyond two dimensions. That is, our two dimensional representation of class similarity can only capture a portion of the relevant information. This makes sense from a cross-classification perspective. Two classes might be similar to each other along some dimension, but different from each other along some other dimension. The MDS diagrams are only approximations of the actual similarity effects between classes.

### 8.3 Interim Conclusion

In this chapter I looked at two complex inflectional systems: Spanish verb inflection and Kasem singular-plural classes. In Spanish, verbs are divided into three main inflection classes: *-ar*, *-er* and *-ir* verbs. Additionally, a set of verbs show different kinds of vocalic and consonant stem alternations in the present tense and past participle. Analogical models trained on the phonological shape of the stems could predict with high accuracy the main inflection class of verbs, and the stem alternation that verbs exhibit. The clustering based on stem similarity showed that verbs that undergo the same stem alternation have similar stems, even if they belong to different main inflection classes.

I propose that these facts taken together constitute very strong evidence that the analogical relations do not only choose one of the trees in the hierarchy, but go up all of them. Naturally, this does not mean that we should always see perfect correlations, but rather that the correlations between the analogical relations and the grammatical hierarchy will be present.

In Kasem, nouns can take a variety of different singular and plural markers. A key feature of this system is that individual markers can denote singular and plural in different nouns. In addition to this, nouns can undergo diphthongization and vowel lengthening in the plural. These three dimensions (markers, lengthening and diphthongization) produce the inflection class of nouns. The analogical models, trained on the phonological shape and meaning of the stems, could correctly distinguish these three dimensions, and predict with a high degree of accuracy the inflection class of nouns. The models showed that inflection class is almost equally predictable from the stem as it is from the singular or plural marker alone.

The clustering analysis in Kasem showed that inflection classes that shared

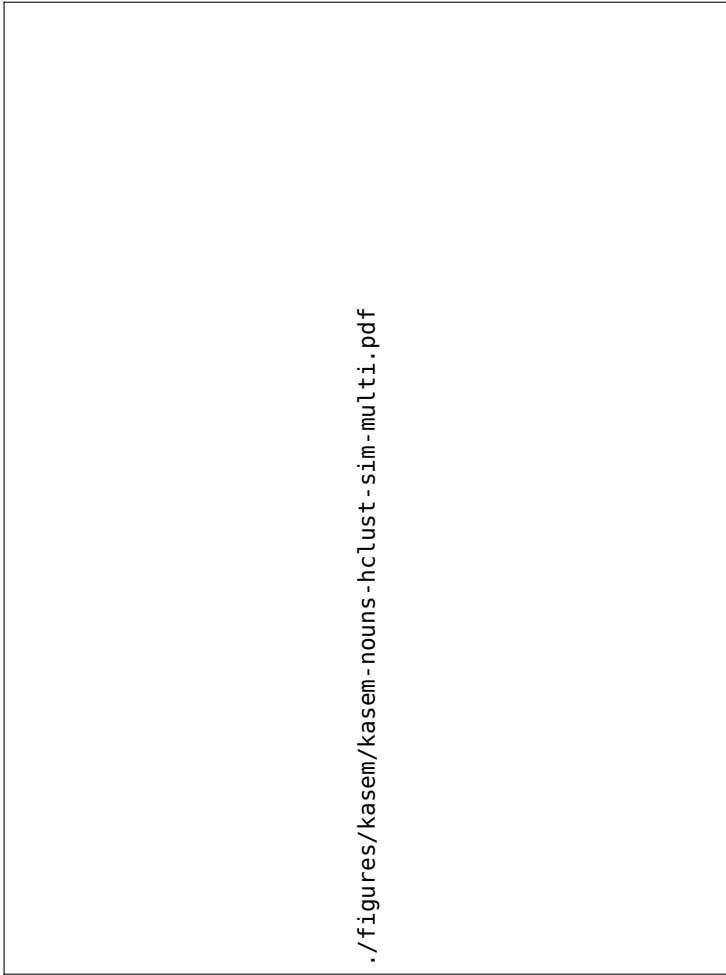


Figure 8.25: Clustering of inflection class in Kasem based on the singular stem, lengthening and diphthongization across multiple dimensions.

the same markers clustered together, even if markers were flipped, i.e. marking singular in one class and plural in the other class. This means that the analogical relations must also hold at a more abstract level, and not just on the leaves of the hierarchy. This is because if nouns of classes  $\mathfrak{a}-i$  and  $u-\mathfrak{a}$  (as many other cases discussed above) are similar to each other, it means that at some level both classes must share a general type  $\mathfrak{a}$  underspecified for number.

Overall, this chapter shows that the kind of analogical classifiers proposed in this book can model very complex systems with many classes. It also shows analogical relations still reveal aspects of the hierarchy, even if said hierarchy includes very complex interactions of multiple dimensions.

# References

- Ackerman, Farrell & Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language* 89(3). 429–464.
- Ackerman, Farrell & Robert Malouf. 2016. Word and pattern morphology: An information-theoretic approach. *Word Structure* 9(2). 125–131.
- Afonso, Olivia, Alberto Domínguez, Carlos J. Álvarez & David Morales. 2014. Sublexical and lexico-syntactic factors in gender access in Spanish. *Journal of Psycholinguistic Research* 43(1). 13–25.
- Aguirre, Carmen & Wolfgang U. Dressler. 2008. On Spanish verb inflection. *Folia Linguistica* 40. 75–96.
- Alber, Birgit. 2009. Past participles in Mocheno: Allomorphy and alignment. In Michael T. Putnam (ed.), *Studies on German-language islands*, 33–64. Amsterdam, Philadelphia: Benjamins.
- Albright, Adam. 2008a. Explaining universal tendencies and language particulars in analogical change. In Jeff Good (ed.), *Linguistic universals and language change*, 144–184. Oxford: Oxford University Press.
- Albright, Adam. 2008b. How many grammars am I holding up? Discovering phonological differences between word classes. In *Proceedings of the 26th West Coast Conference on Formal Linguistics*, 1–20. Somerville, MA: Cascadia Proceedings Project.
- Albright, Adam. 2009. Modeling analogy as probabilistic grammar. In James P. Blevins & Juliette Blevins (eds.), *Analogy in grammar*, 200–228. Oxford, New York: Oxford University Press.
- Albright, Adam, Argelia Andrade & Bruce Hayes. 2001. Segmental environments of Spanish diphthongization. *UCLA Working Papers in Linguistics* 7. 117–151.
- Albright, Adam & Bruce Hayes. 1999. *An automated learner for phonology and morphology*. [https://www.researchgate.net/profile/Bruce\\_Hayes/publication/2876878\\_An\\_Automated\\_Learner\\_for\\_Phonology\\_and\\_Morphology/links/02bfe51194b1fab53a000000.pdf](https://www.researchgate.net/profile/Bruce_Hayes/publication/2876878_An_Automated_Learner_for_Phonology_and_Morphology/links/02bfe51194b1fab53a000000.pdf), accessed 2016-9-10.
- Albright, Adam & Bruce Hayes. 2002. Modeling English past tense intuitions with minimal generalization. In *Proceedings of the ACL-02 workshop on morpho-*



## References

- logical and phonological learning*, vol. 6, 58–69. Association for Computational Linguistics.
- Albright, Adam & Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90(2). 119–161.
- Alexander, Ronelle. 2006. *Bosnian, Croatian, Serbian, a grammar: With sociolinguistic commentary*. Madison: University of Wisconsin Press.
- Anderson, Stephen R. 2008. Phonologically conditioned allomorphy in the morphology of Surmiran (Rumantsch). *Word Structure* 1(2). 109–134.
- Anderson, Stephen R. 2015. Morphological change. In Richard D. Janda & Brian D. Joseph (eds.), *The Routledge handbook of historical linguistics*, 264–285. Oxford: Blackwell Publishing.
- Anttila, Raimo. 1977. *Analogy*. Berlin: Mouton de Gruyter.
- Anttila, Raimo. 2003. Analogy: The warp and woof of cognition. In Richard D. Janda & Brian D. Joseph (eds.), *The Routledge handbook of historical linguistics*, 425–440. Malden, Mass.: Blackwell Publishing.
- Arndt-Lappe, Sabine. 2011. Towards an exemplar-based model of stress in English noun–noun compounds. *Journal of Linguistics* 47. 549–585.
- Arndt-Lappe, Sabine. 2014. Analogy in suffix rivalry: The case of English *-ity* and *-ness*. *English Language and Linguistics* 18(3). 497–548.
- Aronoff, Mark. 1994. *Morphology by itself: Stems and inflectional classes*. Massachusetts: MIT Press.
- Arppe, Antti, Peter Hendrix, Petar Milin, R. Harald Baayen & Cyrus Shaoul. 2014. Ndl: Naive discriminative learning. *R package versions 0.1*.
- Arregi, Karlos. 2000. How the Spanish verb works. In *30th linguistic symposium on romance languages*. Gainesville: University of Florida. <http://home.uchicago.edu/karlos/Arregi-2000-how.pdf>.
- Awedoba, Albert K. 1980. Borrowed nouns in Kasem nominal classes. *Anthropological Linguistics* 22(6). 248–263.
- Awedoba, Albert K. 1996. Kasem nominal genders and names. *Research Review* 12(2). 8–24.
- Awedoba, Albert K. 2003. Criteria for noun classification in Kasem. In Manfred von Roncador, Kerstin Winkelmann & Ulrich Kleinewillinghöfer (eds.), *Cahiers Voltaïques / Gur Papers*, vol. 6, 3–15.
- Baayen, R. Harald. 2007. Storage and computation in the mental lexicon. In Gonia Jarema & Gary Libben (eds.), *Mental lexicon: Core perspectives*, 81–104. Amsterdam: Elsevier.
- Baayen, R. Harald. 2010. Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon* 5(3). 436–461.

- Baayen, R. Harald. 2011. Corpus linguistics and naive discriminative learning. *Revista Brasileira de Linguística Aplicada* 12(2). 295–328.
- Baayen, R. Harald & Peter Hendrix. 2011. Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning. In *Empirically examining parsimony and redundancy in usage-based models, lsa workshop*.
- Baayen, R. Harald, Petar Milin, Dusica Filipović Đurđević, Peter Hendrix & Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118(3). 438–81.
- Baerman, Matthew. 2007. Morphological reversals. *Journal of Linguistics* 43. 33–61.
- Baptista, Barbara O. & Jair L.A. Silva Filho. 2006. The influence of voicing and sonority relationships on the production of English final consonants. In Barbara O. Baptista & Michal Alan Watkins (eds.), *English with a Latin beat: Studies in Portuguese/Spanish-English interphonology*, 73–90. Amsterdam: John Benjamins.
- Bargery, George Percy & Diedrich Westermann. 1951. *A Hausa-English dictionary and English-Hausa vocabulary*. Oxford: Oxford University Press.
- Bateman, Nicoleta & Maria Polinsky. 2010. Romanian as a two-gender language. In Donna B. Gerdts, John C. Moore & Maria Polinsky (eds.), *Hypothesis A/hypothesis B: Linguistic explorations in honor of David M. Perlmutter* (Current Studies in Linguistics), 41–77. Cambridge, Massachusetts: MIT Press.
- Bauer, Laurie. 2003. *Introducing linguistic morphology*. Washington, D. C.: Georgetown University Press.
- Bechtel, William & Adele Abrahamsen. 2002. *Connectionism and the mind: Parallel processing, dynamics, and evolution in networks*. Malden, Mass.: Blackwell Publishing.
- Becker, Thomas. 1990. *Analogie und morphologische Theorie* (Studien zur Theoretischen Linguistik 11). München: Fink.
- Becker, Thomas. 1993. Back-formation, cross-formation, and bracketing paradoxes in paradigmatic morphology. In Geert E. Booij & Jaap van Marle (eds.), *Yearbook of Morphology 1993*, 1–26. Dordrecht: Springer.
- Bellido, Paloma García. 1986. *Lexical diphthongization and high-mid alternations in Spanish: An autosegmental account*. Seattle, WA: Linguistic Analysis.
- Beniamine, Sacha. 2017. A computational approach to the abstraction of morphophonological alternations. In *Typologie et modélisation des systèmes morphologiques*. Paris. <http://www.llf.cnrs.fr/fr/node/5611>.

## References

- Beniamine, Sacha & Olivier Bonami. 2016. Generalizing patterns in Instrumented Item-and-Pattern Morphology. In *Structural complexity in natural language(s) (SCNL)*. Paris.
- Bergen, Benjamin & Nancy Chang. 2005. Embodied construction grammar in simulation-based language understanding. In Jan-Ola Östman & Mirjam Fried (eds.), *Construction grammars: Cognitive grounding and theoretical extensions*, 147–190. Amsterdam, Philadelphia: John Benjamins.
- Bermúdez-Otero, Ricardo. 2013. The Spanish lexicon stores stems with theme vowels, not roots with inflectional class features. *International Journal of Latin and Romance Linguistics* 25(1). 3–103.
- Blevins, James P. 2006. Word-based morphology. *Journal of Linguistics* 42(3). 531–573.
- Blevins, James P. 2008. Declension classes in Estonian. *Linguistica Uralica* 4. 241–267.
- Blevins, James P. 2013. The information-theoretic turn. *Psihologija* 46(3). 355–375.
- Blevins, James P. 2016. *Word and paradigm morphology*. Oxford, New York: Oxford University Press.
- Blevins, James P., Petar Milin & Michael Ramscar. 2016. The Zipfian paradigm cell filling problem. In Ferenc Kiefer, James Blevins & Huba Bartos (eds.), *Perspectives on morphological organization: Data and analyses*, 141–158.
- Bloomfield, Leonard. 1933. *Language*. New York: Holt, Reinhart & Winston.
- Bodomo, Adams. 1994. The noun class system of Dagaare: A phonology-morphology interface. In *Working Papers in Linguistics, Norwegian University for Science and Technology*.
- Bodomo, Adams. 1997. *The structure of Dagaare*. Stanford: CSLI publications.
- Boersma, Paul P. G. 1997. How we learn variation, optionality, and probability. In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, 43–58.
- Boersma, Paul P. G. 1998. *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. Den Haag: Holland Academic Graphics/IFOTT.
- Boersma, Paul P. G. & Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32(1). 45–86.
- Boloh, Yves & Laure Ibernou. 2010. Gender attribution and gender agreement in 4- to 10-year-old French children. *Cognitive Development* 25(1). 1–25.
- Bonami, Olivier & Sacha Beniamine. 2016. Joint predictiveness in inflectional paradigms. *Word Structure* 9(2). 156–182.

- Bonami, Olivier & Gilles Boyé. 2003. Supplétion et classes flexionnelles. *Langages* 37(152). 102–126.
- Bonami, Olivier & Gilles Boyé. 2006. Deriving inflectional irregularity. In *Proceedings of the 13th International Conference on HPSG*, 39–59. Varna: CSLI Publications.
- Booij, Geert E. 1998. Phonological output constraints in morphology. In Wolfgang Kehrein & Richard Wiese (eds.), *Phonology and morphology of the Germanic languages*, 143–163. Tübingen: Niemeyer.
- Booij, Geert E. 2010. *Construction morphology*. Oxford, New York: Oxford University Press.
- Borg, Ingwer & Patrick J. F. Groenen. 2005. *Modern multidimensional scaling: Theory and applications*. New York: Springer.
- Boyé, Gilles & Patricia Cabredo Hofherr. 2010. The distribution of prethematic vowels in Spanish verbs. <http://w3.erss.univ-tlse2.fr:8080/index.jsp?perso=boye&subURL=BoCa-Probus-PrethematicVowelsSpanish.pdf>, accessed 2010-5-10.
- Boyé, Gilles & Patricia Cabredo Hofherr. 2004. Étude de la distribution des suffixes-er/-ir dans les infinitifs de l'espagnol à partir d'un corpus exhaustif. *Corpus* (3). 237–260.
- Boyé, Gilles & Patricia Cabredo Hofherr. 2006. The structure of allomorphy in Spanish verbal inflection. *Cuadernos de Lingüística del Instituto Universitario Ortega y Gasset* 13. 9–24.
- Brame, Michael K. & Ivonne Bordelois. 1973. Vocalic alternations in Spanish. *Linguistic Inquiry* 4(2). 111–168.
- Braune, Wilhelm. 1895. *Gotische Grammatik: Mit einigen Lesestücken und Wortverzeichnis*. Halle (Saale): Niemeyer.
- Breiman, Leo. 2001. Random forests. *Machine Learning* 45(1). 5–32.
- Bresnan, Joan, Ash Asudeh, Ida Toivonen & Stephen Wechsler. 2016. *Lexical-functional syntax*. Chichester: Wiley-Blackwell.
- Bresnan, Joan, Anna Cueni, Tatiana Nikitina & R. Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Krämer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94.
- Bresnan, Joan & Jennifer Hay. 2008. Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua* 118. 245–259.
- Brindle, Jonathan Allen. 2009. On the identification of noun class and gender systems in Chakali. In *Proceedings of the 38th annual conference on african linguistics*, 84–94. Somerville, MA: Cascadilla Proceedings Project.

## References

- Brovetto, Claudia & Michael T. Ullman. 2005. The mental representation and processing of Spanish verbal morphology. In *Selected Proceedings of the 7th Hispanic Linguistics Symposium*, 98–105. Somerville, MA: Cascadilla Proceedings Project.
- Brown, Dunstan & Andrew Hippisley. 2012. *Network morphology: A defaults-based theory of word structure*. Cambridge: Cambridge University Press.
- Brown, Wayles. 1993. Serbo-Croat. In Bernard Comrie & Greville G. Corbett (eds.), *The Slavonic Languages*, 306–387. London, New York: Routledge.
- Butterworth, Brian. 1983. Lexical representation. In Brian Butterworth (ed.), *Language production: development, writing, and other language processes*, 257–294. London: Academic Press.
- Bybee, Joan. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10(5). 425–455.
- Bybee, Joan L. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.
- Bybee, Joan L. & Clayton Beckner. 2015. Language use, cognitive processes and linguistic change. In Claire Bower & Bethwyn Evans (eds.), *The Routledge handbook of historical linguistics*, 503–518. London: Routledge.
- Bybee, Joan L. & Dan I. Slobin. 1982. Rules and schemas in the development and use of the English past tense. *Language* 58(2). 265–289.
- Caffarra, Sendy & Horacio A. Barber. 2015. Does the ending matter? The role of gender-to-ending consistency in sentence reading. *Brain Research* 1605. 83–92.
- Caffarra, Sendy, Anna Siyanova-Chanturia, Francesca Pesciarelli, Francesco Vespignani & Cristina Cacciari. 2015. Is the noun ending a cue to grammatical gender processing? An ERP study on sentences in Italian. *Psychophysiology* 52(8). 1019–1030.
- Callow, John C. 1965. Kasem nominals: A study in analyses. *Journal of West African Languages* 2(1). 29–36.
- Carreira, Maria. 1991. The alternating diphthongs of Spanish: A paradox revisited. In *Current studies in Spanish linguistics*, 407–445. Washington, D. C.: Georgetown University Press.
- Carstairs, Andrew. 1990. Phonologically conditioned suppletion. *Contemporary Morphology*. 17–23.
- Carstairs, Andrew. 1998. Some implications of phonologically conditioned suppletion. In Geert E. Booij & Jaap van Marle (eds.), *Yearbook of Morphology 1998*, 67–94. Dordrecht: Springer.
- Casali, Roderic F. 2008. ATR harmony in African languages. *Language and Linguistics Compass* 2(3). 496–549.

- Chomsky, Noam & Morris Halle. 1968. *The sound pattern of English*. New York: Harper and Row.
- Churchland, Paul M. 1989. *A neurocomputational perspective: The nature of mind and the structure of science*. Massachusetts: MIT press.
- Clahsen, Harald, Fraibet Aveledo & Iggy Roca. 2002. The development of regular and irregular verb inflection in Spanish child language. *Journal of Child Language* 29. 591–622.
- Clopper, C. J. & Egon S. Pearson. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26(4). 404–413.
- Cojocaru, Dana. 2003. *Romanian grammar*. Durham: The Slavic and East European Language Resource Center.
- Contini-Morava, Ellen. 1994. *Noun classification in Swahili*. <http://www2.iath.virginia.edu/swahili/swahili.html>, accessed 2016-11-9.
- Corbett, Greville G. 1991. *Gender*. Cambridge: Cambridge University Press.
- Corbett, Greville G. & Norman M. Fraser. 1993. Network Morphology: A DATR account of Russian nominal inflection. *Journal of Linguistics* 29. 113–142.
- Costanzo, Angelo Roth. 2011. *Romance conjugational classes: Learning from the peripheries*. Columbus, OH: The Ohio State University dissertation.
- Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.
- Croft, William & Alan D. Cruse. 2004. *Cognitive linguistics*. Cambridge, MA: Cambridge University Press.
- Cucerzan, Silviu & David Yarowsky. 2003. Minimally supervised induction of grammatical gender. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 40–47. Association for Computational Linguistics.
- Cuervo, Rufino José & Ignacio Ahumada. N.d. *Notas a la Gramática de la lengua castellana de don Andrés Bello*. Bogotá: Instituto Caro y Cuervo.
- Cysouw, Michael. 2007. New approaches to cluster analysis of typological indices. In Peter Grzybek & Reinhard Köhler (eds.), *Exact methods in the study of language and text*, 61–76. Berlin, Boston: De Gruyter Mouton.
- Czaplicki, Bartłomiej. 2013. Arbitrariness in grammar: Palatalization effects in Polish. *Lingua* 123. 31–57.
- Dakubu, Mary Esther Kropp. 1997. Oti-Volta vowel harmony and Dagbani. *Gur Papers* 2. 81–88.
- De Smet, Hendrik & Olga Fischer. 2017. The role of analogy in language change: Supporting constructions. In Marianne Hundt, Sandra Mollin & Simone E.

## References

- Pfenninger (eds.), *The changing English language*. Cambridge, New York: Cambridge University Press.
- De Vaan, Laura, Robert Schreuder & R. Harald Baayen. 2007. Regular morphologically complex neologisms leave detectable traces in the mental lexicon. *The Mental Lexicon* 2(1). 1–24.
- de Haas, Wim G. 1987. An autosegmental approach to vowel coalescence. *Lingua* 73(3). 167–199.
- de Haas, Wim G. 1988. *A formal theory of vowel coalescence: A case study of Ancient Greek*. Berlin: Walter de Gruyter.
- DeMello, George. 1993. -Ra vs.-se subjunctive: A new look at an old topic. *Hispania* 76(2). 235–244.
- Derwing, Bruce L. & Royal Skousen. 1994. Productivity and the English past tense. In Susan D. Lima, Roberta Corrigan & Gregory K. Iverson (eds.), *The reality of linguistic rules*, vol. 26, 193–218. Amsterdam, Philadelphia: John Benjamins.
- Di Sciullo, Anna-Maria & Edwin Williams. 1987. *On the definition of word*. Cambridge, Massachusetts: Springer.
- Dinu, Liviu P., Vlad Niculae & Octavia-Maria Sulea. 2012. Dealing with the grey sheep of the Romanian gender system, the neuter. In *COLING (Demos)*, 119–124.
- Echegoyen, Artemisa & Katherine Voigtlander. 1979. *Luces contemporáneas del otomí: Gramática del otomí de la sierra*. Mexico, D. F.: Instituto Lingüístico de Verano. <https://www.sil.org/resources/archives/2018>, accessed 2017-2-21.
- Echegoyen, Artemisa & Katherine Voigtlander. 2007. *Diccionario yuhú: Otomí de la Sierra Madre Oriental: Estados de Hidalgo, Puebla y Veracruz, México*. Estados de Hidalgo, Puebla y Veracruz, México: Instituto Lingüístico de Verano.
- Eddington, David. 1996. Diphthongization in Spanish derivational morphology: An empirical investigation. *Hispanic Linguistics* 8(1). 1–13.
- Eddington, David. 2000. Analogy and the dual-route model of morphology. *Lingua* 110(4). 281–298.
- Eddington, David. 2002. Spanish gender assignment in an analogical framework. *Journal of Quantitative Linguistics* 9(1). 49–75.
- Eddington, David. 2004. Issues in modeling language processing analogically. *Lingua* 114(7). 849–871.
- Eddington, David. 2009. Linguistic processing is exemplar-based. *Studies in Hispanic and Lusophone Linguistics* 2(2).
- Eddington, David & Jordan Lachler. 2006. A computational analysis of Navajo verb stems. In Sally Rice & John Newman (eds.), *Empirical and experimental methods in cognitive/functional research*, 143–161. CSLI Publications.

- Erelt, Mati, Tiiu Erelt & Kristiina Ross. 1997. *Eesti keele käsiraamat*. Tallinn: Eesti keele sihtasutus.
- Erelt, Mati, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael & Silvi Vare. 1995. *Eesti keele grammatika: Morfoloogia*. Tallinn: Eesti Teaduste Akadeemia Eesti Keele Instituut.
- Erelt, Tiiu, Tiina Leemets, Sirje Mäearu & Maire Raadik. 2001. *Eesti keele sõnaraamat*: ÕS. Tallinn: Eesti Keele Sihtasutus.
- Farkas, Donka F. 1990. Two cases of underspecification in morphology. *Linguistic Inquiry* 21(4). 539–550.
- Farkas, Donka F. & Draga Zec. 1995. Agreement and pronominal reference. In Guglielmo Cinque & Giuliana Giusti (eds.), *Advances in Romanian linguistics*, vol. 10, 83–101. Amsterdam, Philadelphia: John Benjamins.
- Federici, Stefano, Vito Pirrelli & François Yvon. 1995. A dynamic approach to paradigm-driven analogy. In *International Joint Conference on Artificial Intelligence*, 385–398.
- Feist, Timothy & Enrique L. Palancar. 2015. *Oto-Manguean inflectional class database*.
- Fertig, David L. 2013. *Analogy and morphological change*. Edinburgh: Edinburgh University Press.
- Fillmore, Charles J. & Paul Kay. 1995. *A Construction Grammar coursebook*. Berkeley: Unpublished ms, University of California.
- Foley, James Addison. 1965. *Spanish morphology*. Massachusetts: Massachusetts Institute of Technology dissertation.
- Fondow, Steven Richard. 2010. *Spanish velar insertion and analogy: A usage-based diachronic analysis*. Columbus, Ohio: Columbus, OH dissertation. [https://etd.ohiolink.edu/rws\\_etd/document/get/osu1290438177/inline](https://etd.ohiolink.edu/rws_etd/document/get/osu1290438177/inline), accessed 2017-6-3.
- Francis, Elaine J. & Laura A. Michaelis. 2014. Why move? How weight and discourse factors combine to predict relative clause extraposition in English. In Brian MacWhinney, Edith A. Moravcsik & Andrej L. Malchukov (eds.), *Competing motivations*, 70–87. Oxford, New York: Oxford University Press.
- Galván Torres, Adriana Rosalina. 2007. *Die Entwicklung der spanischen Diphthongierung anhand der Natürlichkeitstheorie*. Norderstedt: GRIN Verlag.
- Gerdts, Donna B., John C. Moore & Maria Polinsky (eds.). 2010. *Hypothesis A/hypothesis B: Linguistic explorations in honor of David M. Perlmutter* (Current Studies in Linguistics). Cambridge, Massachusetts: MIT Press.
- Ginzburg, Jonathan & Ivan A. Sag. 2000. *Interrogative investigations*. Stanford: CSLI publications.



## References

- Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: Univ. of Chicago Press.
- Goldberg, Adele E. 2006. *Constructions at Work*. Oxford: Oxford University Press.
- Goldsmith, John A. 2009. Morphological analogy: Only a beginning. In James P. Blevins & Juliette Blevins (eds.), *Analogy in grammar*, 138–164. Oxford, New York: Oxford University Press.
- Goldsmith, John A., Jason Riggle & C. L. Yu Alan. 2011. *The handbook of phonological theory*. Chichester: John Wiley & Sons.
- Gönczöl, Ramona. 2007. *Romanian: An essential grammar*. New York: Routledge.
- Gouskova, Maria, Luiza Newlin-Łukowicz & Sofya Kasyanenko. 2015. Selectional restrictions as phonotactics over sublexicons. *Lingua* 167. 41–81.
- Guzmán Naranjo, Matías & Olivier Bonami. 2016. *Overabundance as hybrid inflection: Quantitative evidence from Czech*. Grammar and Corpora. IDS. Mannheim.
- Guzmán Naranjo, Matías & Elena Pyatigorskaya. 2016. *Comparing naive discriminative learning, sublexicon phonotactics, and analogical learning*. Olinco. Olomouc.
- Hahn, Ulrike & Nick Chater. 1998. Similarity and rules: Distinct? Exhaustive? Empirically distinguishable? *Cognition* 65(2). 197–230.
- Hahn, Ulrike & Ramin Charles Nakisa. 2000. German inflection: Single route or dual route? *Cognitive Psychology* 41(4). 313–360.
- Hall, Robert A. 1965. The “neuter” in Romance: A pseudo-problem. *Word* 21(3). 421–427.
- Halle, Morris. 1978. Further thoughts on Kasem nominals. *Linguistic Analysis Seattle* 4(2). 167–185.
- Halle, Morris & Alec Marantz. 1993. Distributed morphology and the pieces of inflection. In Kenneth Locke Hale & Samuel Jay Keyser (eds.), *The view from building 20: Essays in linguistics in honor of Sylvain Bromberger*, 111–176. Cambridge, MA: MIT Press.
- Hammond, Lila. 2005. *Serbian: An essential grammar*. New York: Routledge.
- Harris, James W. 1969. *Spanish phonology*. Cambridge, MA: MIT Press.
- Harris, James W. 1978. Two theories of non-automatic morphophonological alternations: Evidence from Spanish. *Language* 54(1). 41–60.
- Harris, James W. 1985. Spanish diphthongisation and stress: A paradox resolved. *Phonology* 2. 31–45.
- Harris, James W. 1987. Disagreement rules, referral rules, and the Spanish feminine article *el*. *Journal of Linguistics* 23. 177–183.

- Harris, James W. 1991. The exponence of gender in Spanish. *Linguistic Inquiry* 22(1). 27–62.
- Hay, Jennifer & Joan Bresnan. 2006. Spoken syntax: The phonetics of *giving a hand* in New Zealand English. *The Linguistic Review* 23(3). 321–349.
- Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry* 39(3). 379–440.
- Hock, Hans Henrich. 1991. *Principles of historical linguistics*. Amsterdam, Philadelphia: Walter de Gruyter.
- Hock, Hans Henrich. 2003. Analogical change. In Richard D. Janda & Brian D. Joseph (eds.), *The handbook of historical linguistics*, 441–460. Malden, Mass.: Blackwell Publishing.
- Holmes, Virginia M. & B. Dejean de la Bâtie. 1999. Assignment of grammatical gender by native speakers and foreign learners of French. *Applied Psycholinguistics* 20. 479–506.
- Holmes, Virginia M. & Juan Segui. 2004. Sublexical and lexical influences on gender assignment in French. *Journal of Psycholinguistic Research* 33(6). 425–457.
- Hooper, Joan B. 1976. *An introduction to natural generative phonology*. New York: Academic Press.
- Itkonen, Esa. 2005. *Analogy as structure and process*. Amsterdam, Philadelphia: John Benjamins.
- Kapatsinski, Vsevolod. 2010. What is it I am writing? Lexical frequency effects in spelling Russian prefixes: Uncertainty and competition in an apparently regular system. *Corpus Linguistics and Linguistic Theory* 6(2). 157–215.
- Kapatsinski, Vsevolod. 2012. What statistics do learners track? Rules, constraints and schemas in (artificial) grammar learning. In Stefan Th. Gries & Dagmar Divjak (eds.), *Frequency effects in language learning and processing*, 53–82. Berlin, Boston: De Gruyter Mouton.
- Kapatsinski, Vsevolod. 2014. What is grammar like? A usage-based constructionist perspective. *Linguistic Issues in Language Technology* 11(1). 1–41.
- Kaplan, Ronald M. & Joan Bresnan. 1982. Lexical-functional grammar: A formal system for grammatical representation. In Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell & Annie Zaenen (eds.), *Formal issues in lexical-functional grammar*, 29–130. Stanford: CSLI Publications.
- Kempas, Ilpo. 2011. Sobre la variación en el marco de la libre elección entre cantar y cantase en el español peninsular. *Moenia* (17). 243–264.
- Kempe, Vera & Patricia J. Brooks. 2001. The role of diminutives in the acquisition of Russian gender: Can elements of child-directed speech aid in learning morphology? *Language Learning* 51(2). 221–256.

## References

- Kempe, Vera, Patricia J. Brooks & Anatoliy Kharkhurin. 2010. Cognitive predictors of generalization of Russian grammatical gender categories. *Language Learning* 60(1). 127–153.
- Kempe, Vera, Patricia J. Brooks, Natalija Mironova & Olga Fedorova. 2003. Diminutivization supports gender acquisition in Russian children. *Journal of Child Language* 30. 471–485.
- Kikuchi, Seiichiro. 1997. A correspondence-theoretic approach to alternating diphthongs in Spanish. *Journal of Linguistic Science Tohoku University* 1. 39–50.
- Kilani-Schoch, Marianne & Wolfgang U. Dressler. 2005. *Morphologie naturelle et flexion du verbe français*. Tübingen: Gunter Narr Verlag.
- Koenig, Jean-Pierre. 1999. *Lexical relations*. Stanford: CSLI publications.
- Kohavi, Ron. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on artificial intelligence*, 1137–1143. Morgan Kaufmann.
- Köpcke, Klaus-Michael. 1988. Schemas in German plural formation. *Lingua* 74(4). 303–335.
- Köpcke, Klaus-Michael. 1998a. Prototypisch starke und schwache Verben der deutschen Gegenwartssprache. *Germanistische Linguistik* 141(142). 45–60.
- Köpcke, Klaus-Michael. 1998b. The acquisition of plural marking in English and German revisited: Schemata versus rules. *Journal of Child Language* 25. 293–319.
- Köpcke, Klaus-Michael, Klaus-Uwe Panther & David A. Zubin. 2010. Motivating grammatical and conceptual gender agreement in German. In Hans-Jörg Schmid & Susanne Handl (eds.), *Cognitive foundations of linguistic usage patterns*, 171–194. Berlin, Boston: De Gruyter Mouton.
- Köpcke, Klaus-Michael & David A. Zubin. 1984. Sechs Prinzipien für die Genuszuweisung im Deutschen: Ein Beitrag zur natürlichen Klassifikation. *Linguistische Berichte*. 26–50.
- Kordić, Snježana. 1997. *Serbo-Croatian*. München: Lincom Europa.
- Kramer, Ruth. 2015. Impoverishment, gender and number: Predicting the patterns of syncretism. In *Roots IV*. Georgetown University.
- Krott, Andrea, R. Harald Baayen & Robert Schreuder. 2001. Analogy in morphology: Modeling the choice of linking morphemes in Dutch. *Linguistics* 39. 51–94.
- Kuryłowicz, Jerzy. 1945. La nature des procès dits «analogiques». *Acta Linguistica* 5(1). 15–37.

- Lečić, Dario. 2015. Morphological doublets in Croatian: The case of the instrumental singular. *Russian Linguistics* 39(3). 375–393.
- Lee, Hansol H. B. 1989. *Korean grammar*. Oxford, New York: Oxford University Press.
- Lepage, Yves. 1998. Solving analogies on words: An algorithm. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, 728–734.
- Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10(8). 707–710.
- Ljubešić, Nikola & Filip Klubička. 2014. {bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, 29–35. Gothenburg, Sweden: Association for Computational Linguistics.
- Lyster, Roy. 2006. Predictability in French gender attribution: A corpus analysis. *Journal of French Language Studies* 16. 69–92.
- Maiden, Martin. 2001. A strange affinity: ‘Perfecto y tiempos afines’. *Bulletin of Hispanic Studies* 78(4). 441–464.
- Maiden, Martin. 2005. Morphological autonomy and diachrony. In Geert E. Booij & Jaap van Marle (eds.), *Yearbook of Morphology 2004*, 137–175. The Netherlands: Springer.
- Malkiel, Yakov. 1966. Diphthongization, monophthongization, metaphony: Studies in their interaction in the paradigm of the Old Spanish-ir verbs. *Language* 42(2). 430–472.
- Malkiel, Yakov. 1988. A Cluster of (Old) Portuguese derivational suffixes:–” ece,–ice, ez (a)”, viewed in relation to their Spanish counterparts. *Bulletin of Hispanic studies* 65(1). 1–19.
- Marchal, Harmony, Maryse Bianco, Philippe Dessus & Benoît Lemaire. 2007. The development of lexical knowledge: Toward a model of the acquisition of lexical gender in French. In *Proceedings of the european cognitive science conference 2007*, 268–273. Taylor and Francis.
- Mateo, Francis & Antonio J. Rojo Sastre. 1995. *El arte de conjugar en español: Diccionario de 12000 verbos*. Paris: Hatier.
- Matthews, Clive A. 2005. French gender attribution on the basis of similarity: A comparison between AM and connectionist models. *Journal of Quantitative Linguistics* 12. 262–296.
- Matthews, Clive A. 2010. On the nature of phonological cues in the acquisition of French gender categories: Evidence from instance-based learning models. *Lingua* 120(4). 879–900.

## References

- McClelland, James L. & David E. Rumelhart. 1986. A distributed model of human learning and memory. In James L. McClelland & David E. Rumelhart (eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Psychological and biological models*, 170–2015. Cambridge: MIT Press.
- McDonough, Joyce M. 2013. The Dene verb: How phonetics supports morphology. In *Proceedings of 18th Workshop on Structure and Constituency in the Languages of the Americas*. University of California, Berkeley.
- Meyniel, Jean-Philippe, Paul H. Cottu, Charles Decraene, Marc-Henri Stern, Jérôme Couturier, Ingrid Lebigot, André Nicolas, Nina Weber, Virginie Fourchotte & Séverine Alran. 2010. A genomic and transcriptomic approach for a differential diagnosis between primary and secondary ovarian carcinomas in patients with a previous history of breast cancer. *BMC Cancer* 10(1). 1–10.
- Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig & Jon Orwant. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014). 176–182.
- Migeod, Frederick William Hugh. 1914. *A grammar of the Hausa language*. London: K. Paul, Trench, Trübner & co., Ltd.
- Mladenović, A. 1977. Neka pitanja varijantnosti norme u savremenom srpskohrvatskom književnom jeziku. In Stanisław Urbáńczyk (ed.), *Wariancja normy we współczesnych słowiańskich językach literackich*, vol. 38 (Prace Komisji Słowianoznawstwa), 51–56. Kraków.
- Morin, Regina. 2006. Spanish gender assignment in computer and Internet related loanwords. *Rivista di Linguistica* 18. 325–54.
- Moscoso del Prado Martín, Fermín, Aleksandar Kostić & R. Harald Baayen. 2004. Putting the bits together: An information theoretical perspective on morphological processing. *Cognition* 94. 1–18.
- Motsch, Wolfgang. 1977. Ein Plädoyer für die Beschreibung von Wortbildungen auf der Grundlage des Lexikons. In Herbert Ernst Brekle & Kastovsky Dieter (eds.), *Perspektiven der Wortbildungsforschung*, 180–202.
- Müller, Stefan & Stephen Wechsler. 2014. Lexical approaches to argument structure. *Theoretical Linguistics* 40. 1–76.
- Mürk, Harri William. 1997. *A handbook of Estonian: Nouns, adjectives and verbs*. Bloomington: Indiana University, Research Institute for Inner Asia Studies.
- Murtagh, Fionn & Pierre Legendre. 2014. Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of Classification* 31(3). 274–295.

- Mwalonya, Joseph, Alison Nicolle, Steve Nicolle & Juma Zimbu. 2004. *Mgombato: Digo-English-Swahili Dictionary*. Kwale: Digo Language and Literacy Project.
- Naden, Tony. 1988. The Gur languages. *The languages of Ghana* 2. 12–49.
- Naden, Tony. 1989. Gur. In John Bendor-Samuel (ed.), *The Niger-Congo languages*, 141–168. New York: University Press of America.
- Nastase, Vivi & Marius Popescu. 2009. What’s in a name? In some languages, grammatical gender. In *Conference on empirical methods in natural language processing*, 1368–1377. Singapore: ACL and AFNLP.
- Năvălici, Cristian. 2013. PyDEX. <https://github.com/cristianav/PyDEX>, accessed 2016-6-15.
- Neuvel, Sylvain. 2001. Pattern analogy vs. word-internal syntactic structure in West-Greenlandic: Towards a functional definition of morphology. In Geert E. Booij & Jaap van Marle (eds.), *Yearbook of Morphology 2000*, 253–278. Amsterdam: Springer.
- Nevins, Andrew. 2011. Phonologically conditioned allomorph selection. In Colin Ewen, Elizabeth Hume, Marc Van Oostendorp & Keren Rice (eds.), *The companion to phonology*, 2357–2382. London & New York: Continuum.
- Newcombe, Robert G. 1998. Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Stat. Med. Statistics in Medicine*. 857–72.
- Newman, Paul. 2000. *The Hausa language: An encyclopedic reference grammar*. New Haven: Yale University Press.
- Niggli, Idda & Urs Niggli. 2007. *Dictionnaire bilingue Kasim-Français Français-Kassem*. SIL International. [kassem-bf.webonary.org/](http://kassem-bf.webonary.org/), accessed 2016-10-11.
- Nosofsky, Robert M. 1990. Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology* 34(4). 393–418.
- Nosofsky, Robert M., Steven E. Clark & Hyun Jung Shin. 1989. Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15(2). 282–304.
- Nurse, Derek & Thomas J. Hinnebusch. 1993. *Swahili and Sabaki: A linguistic history*. Berkeley: University of California Press.
- O’Bryan, Margie. 1974. The role of analogy in non-derived formations in Zulu. *Studies in the Linguistic Sciences* 4. 144–178.
- Paul, Hermann. 1880. *Prinzipien der Sprachgeschichte*. Tübingen: Walter de Gruyter.
- Phelps, Elaine. 1975. Simplicity criteria in generative phonology - Kasem nominals. *Linguistic Analysis* 1(4). 297–332.

## References

- Phelps, Elaine. 1979. Abstractness and rule ordering in Kasem: A refutation of Halle's maximizing principle. *Linguistic Analysis* 5(1). 29–69.
- Pinker, Steven & Michael T. Ullman. 2002. The past and future of the past tense. *Trends in Cognitive Sciences* 6(11). 456–463.
- Pirrelli, Vito & Stefano Federici. 1994a. Derivational paradigms in morphonology. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, 234–240.
- Pirrelli, Vito & Stefano Federici. 1994b. On the pronunciation of unknown words by analogy in text-to-speech systems. In *Proceedings of the Second Onomastica Research Colloquium*.
- Pollard, Carl & Ivan A. Sag. 1994. *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.
- Port, Robert. 2010a. Forget about phonemes: Language processing with rich memory. In *Proceedings of the interspeech 2010*.
- Port, Robert F. 2010b. Rich memory and distributed phonology. *Language Sciences* 32(1). 43–55.
- Pothos, Emmanuel M. 2005. The rules versus similarity distinction. *Behavioral and Brain Sciences* 28. 1–14.
- Pountain, Christopher J. 2006. Gender and Spanish agentive suffixes: Where the motivated meets the arbitrary. *Bulletin of Spanish Studies* 83(1). 19–42.
- Protassova, Ekaterina & Maria D. Voeikova. 2007. Diminutives in Russian at the early stages of acquisition. *Language Acquisition and Language Disorders* 43. 43–72.
- R Development Core Team. 2008. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rainer, Franz. 1993. *Spanische Wortbildungslehre*. Berlin, New York: Walter de Gruyter.
- Rainer, Franz. 2013. Formación de palabras y analogías: Aspectos diacrónicos. In Isabel Pujol Payet (ed.), *Formación de palabras y diacronía* (Anexos Revista de Lexicografía 19), 141–172. A Coruña: Servicio de Publicaciones.
- Roca, Iggy. 2010. Theme vowel allomorphy in Spanish verb inflection: An autosegmental optimality account. *Lingua. Verb First, Verb Second* 120(2). 408–434.
- Roelofs, Ardi & R. Harald Baayen. 2002. Morphology by itself in planning the production of spoken words. *Psychonomic Bulletin & Review* 9(1). 132–138.
- Rojo, Guillermo. 2008. De nuevo sobre la frecuencia de las formas llegara y llegase. In Jörn Albrecht & Frank Harslem (eds.), *Heidelberger Spätlese. Aus-*

- gewählte Tropfen aus verschiedenen Lagen der spanischen Sprach-und Übersetzungswissenschaft Festschrift, vol. 70, 161–182.
- Rokach, Lior & Oded Maimon. 2005. Clustering methods. In *Data mining and knowledge discovery handbook*, 321–352. Springer.
- Rubach, Jerzy. 2007. Feature geometry from the perspective of Polish, Russian, and Ukrainian. *Linguistic Inquiry* 38(1). 85–138.
- Rubach, Jerzy & Geert E. Booij. 2001. Allomorphy in optimality theory: Polish iotation. *Language*. 26–60.
- Rumelhart, David E. & James L. McClelland. 1986a. On learning the past tenses of English verbs. In *Parallel distributed processing: Explorations in the microstructure of cognition: Psychological and biological models*. Cambridge: MIT Press.
- Rumelhart, David E. & James L. McClelland (eds.). 1986b. *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 2 Psychological and biological models*. Cambridge: MIT Press.
- Russell, Donald Andrew & Michael Winterbottom (eds.). 1989. *Classical literary criticism*. Oxford, New York: Oxford University Press.
- Sadler, Louisa. 2006. Gender resolution in Rumanian. In Miriam Butt, Mary Dalrymple & Tracy Holloway King (eds.), *Intelligent linguistic architectures: Variations on themes by Ronald M. Kaplan*. Stanford, CA: CSLI Publications.
- Sag, Ivan A., Hans C. Boas & Paul Kay. 2012. Introducing sign-based construction grammar. In Ivan A. Sag Hans C. Boas (ed.), *Sign-based construction grammar*, 69–202. Wiley.
- Saldanya, Manuel Pérez & Teresa Vallès. 2005. Catalan morphology and low-level patterns in a network model. *Catalan Journal of Linguistics* 4. 199–223.
- Salim, Bello Ahmad. 1981. *Linguistic borrowing as external evidence in phonology: The assimilation of English loanwords in Hausa*. York: University of York dissertation.
- Salmons, Joseph C. 1993. The structure of the lexicon: Evidence from German gender assignment. *Studies in Language* 17(2). 411–435.
- Sánchez, María F. 1995. *Clasificación y análisis de préstamos del inglés en la prensa de España y México*. Lewiston: Edwin Mellen Press.
- Schlücker, Barbara & Ingo Plag. 2011. Compound or phrase? Analogy in naming. *Lingua* 121(9). 1539–1551.
- Schmid, Helmut. 1995. Treetagger: A language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart* 43. 28.
- Scholkopf, Bernhard & Alexander J. Smola. 2001. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA.: MIT Press.



## References

- Schön, James Frederick. 1862. *Grammar of the Hausa language*. London: Church missionary house.
- Schwenter, Scott. 2013. Strength of priming and the maintenance of variation in the Spanish past subjunctive. [https://www.academia.edu/4857119/\\_Strength\\_of\\_Priming\\_and\\_the\\_-Maintenance\\_of\\_Variation\\_in\\_the\\_Spanish\\_Past\\_Subjunctive-NWAV\\_42\\_2013](https://www.academia.edu/4857119/_Strength_of_Priming_and_the_-Maintenance_of_Variation_in_the_Spanish_Past_Subjunctive-NWAV_42_2013), accessed 2015-10-10.
- Schwichtenberg, Beate & Niels O. Schiller. 2004. Semantic gender assignment regularities in German. *Brain and Language* 90(1). 326–337.
- Seigneuric, Alix, Daniel Zagar, Fanny Meunier & Elsa Spinelli. 2007. The relation between language and cognition in 3- to 9-year-olds: The acquisition of grammatical gender in French. *Journal of Experimental Child Psychology* 96(3). 229–246.
- Singh, Rajendra & Alan Ford. 2003. In praise of Śakaṭāyana: Some remarks on whole word morphology. In Rajendra Singh, Stanley Starosta & Sylvain Neuvel (eds.), *Explorations in seamless morphology*, 66–76. New Delhi, Thousand Oaks, London: Sage.
- Singh, Rajendra, Stanley Starosta & Sylvain Neuvel (eds.). 2003. *Explorations in seamless morphology*. New Delhi, Thousand Oaks, London: Sage.
- Skousen, Royal. 1989. *Analogical modeling of language*. Dordrecht: Springer Science & Business Media.
- Skousen, Royal. 1992. *Analogy and structure*. Dordrecht: Springer.
- Skousen, Royal, Deryle Lonsdale & Dilworth B. Parkinson. 2002. *Analogical modeling: An exemplar-based approach to language*. Amsterdam: John Benjamins.
- Smead, Robert N. 2000. On the assignment of gender to Chicano anglicisms: Processes and results. *Bilingual Review/La Revista Bilingüe*. 277–297.
- Smola, Alex J. & Bernhard Schölkopf. 1998. *Learning with kernels*. Sankt Augustin: GMD Forschungszentrum Informationstechnik.
- Song, Jae Jung. 2006. *The Korean language: Structure, use and context*. London & New York: Routledge.
- Steels, Luc. 2011. *Design patterns in fluid construction grammar*. Amsterdam, Philadelphia: John Benjamins Publishing.
- Steriade, Donca. 2008. A pseudo-cyclic effect in Romanian morphophonology. *Inflectional Identity* 18.
- Strauss, Trudie & Michael Johan von Maltitz. 2017. Generalising Ward’s method for use with Manhattan distances. *PLoS ONE* 12(1).
- Stump, Gregory. 2016. *Inflectional paradigms: Content and form at the syntax-morphology interface*. Cambridge, New York: Cambridge University Press.

- Suh, Yunju. 2008. Korean suffix allomorphy in OT. In *The Proceedings of SUNY/CUNY/NYU Mini-Conference: Linguistics in the Big Apple*.
- Taylor, John R. 2012. *The mental corpus: How language is represented in the mind*. Oxford: Oxford University Press.
- Thornton, Anna M. 2010a. Diachronic paths to reduction and maintenance of overabundance in Italian verb paradigms'. In *14th IMM, Budapest*. Budapest.
- Thornton, Anna M. 2010b. Towards a typology of overabundance. In *Décembrettes 7: International Conference on Morphology, University of Toulouse*, 2–3. Toulouse.
- Thornton, Anna M. 2011. Overabundance (multiple forms realizing the same cell): A non-canonical phenomenon in Italian verb morphology. In Martin Maiden, John Charles Smith, Maria Goldbach & Marc-Olivier Hinzelin (eds.), *Morphological autonomy: Perspectives from romance inflectional morphology*. Oxford: Oxford University Press.
- Trask, Robert Lawrence. 1996. *Historical linguistics*. Oxford, New York: Oxford University Press.
- Tucker, G. Richard, Wallace E. Lambert & André Rigault. 1977. *The French speaker's skill with grammatical gender: An example of rule-governed behavior*. The Hague: De Gruyter.
- Tucker, G. Richard, Wallace E. Lambert, André Rigault & Norman Segalowitz. 1968. A psychological investigation of French speakers' skill with grammatical gender. *Journal of Verbal Learning and Verbal Behavior* 7(2). 312–316.
- Ullman, Michael T. 2001. The declarative/procedural model of lexicon and grammar. *Journal of Psycholinguistic Research* 30(1). 37–69.
- Ullman, Michael T. 2004. Contributions of memory circuits to language: The declarative/procedural model. *Cognition* 92(1). 231–270.
- Vallès, Teresa. 2004. *La creativitat lèxica en un model basat en l'ús: Una aproximació cognitiva a la neologia i la productivitat*. Barcelona: L'Abadia de Montserrat.
- van Marle, Jaap. 1985. *On the paradigmatic dimension of morphological creativity*. Dordrecht: Foris.
- Venables, William N. & Brian D. Ripley. 2002. *Modern applied statistics with S*. Fourth. New York: Springer.
- Viks, Ülle. 1992. *A concise morphological dictionary of Estonian: Introduction & grammar*. Vol. 1. Tallinn: Estonian Academy of sciences, Institute of language and literature.

## References

- Viks, Ülle. 1994. A morphological analyzer for the Estonian language: The possibilities and impossibilities of automatic analysis. *Automatic Morphology of Estonian* 1. 7–28.
- Viks, Ülle. 1995. *Rules for recognition of inflection types*. <http://www.eki.ee/teemad/morfoloogia/viks2.html>, accessed 2016-6-10.
- Voeykova, Maria D. 1998. Acquisition of diminutives by a Russian child: Preliminary observations in connection with the early adjectives. *Studies in the acquisition of number and diminutive marking*. 97–113.
- Voigtlander, Katherine & Artemisa Echegoyen. 2007. *Gramática del yuhú: Otomí de la Sierra Madre Oriental*. Mexico, D. F.: Instituto Lingüístico de Verano. <https://www.sil.org/resources/archives/2018>, accessed 2017-2-21.
- Vrabie, Emil. 1989. On the distribution of the neuter plural endings in Modern Standard Romanian (MSR). *The Slavic and East European Journal* 33(3). 400–410.
- Vrabie, Emil. 2000. Feminine noun plurals in Standard Romanian. *The Slavic and East European Journal* 44(4). 537–552.
- Wanner, Dieter. 2006. An analogical solution for Spanish Soy, Doy, Voy, and Estoy. *Probus* 18(2). 267–308.
- Wechsler, Stephen. 2008. Elsewhere in gender resolution. In Kristin Hanson & Sharon Inkelas (eds.), *The nature of the word: Essays in honor of Paul Kiparsky*, 567–586. MIT Press.
- Welmers, William E. 1973. *African language structures*. Berkeley: University of California Press.
- Whitaker, William. 2016. *William Whitaker's words*. <http://mk270.github.io/whitakers-words/>, accessed 2016-6-6.
- Whitney, William Dwight. 1986. *Sanskrit grammar: Including both, the classical language and the older dialects of Veda and Brāhmaṇa*. London: Kegan Paul, Trench, Trübner & Co.
- Wikimedia Foundation. 2016. *Wiktionary*. [https://en.wiktionary.org/wiki/Appendix:Swahili\\_noun\\_classes](https://en.wiktionary.org/wiki/Appendix:Swahili_noun_classes), accessed 2016-9-9.
- Wilkinson, Hugh E. 1971. Vowel alternation in the Spanish -ir verbs. *Ronshu* 12. 1–21.
- Wills, Andy J. & Emmanuel M. Pothos. 2012. On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin* 138(1). 102–125.
- Yaden, Bridet. 2003. Mental representations of Spanish morphology: Rules or analogy? In Paula Kempchinsky & Carlos-Eduardo Piñeros (eds.), *Theory, practice, and acquisition*, 299–312. Somerville, MA: Cascadilla Press.

- Yvon, François. 1997. Paradigmatic cascades: A linguistically sound model of pronunciation by analogy. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 428–435.
- Zaleska, Joanna. 2017. *Coalescence without Coalescence*. Leipzig: Universität Leipzig dissertation.
- Zubin, David A. & Klaus-Michael Köpcke. 1984. Affect classification in the German gender system. *Lingua* 63(1). 41–96.
- Zubin, David A. & Klaus-Michael Köpcke. 1985. Natural classification in language. A study of the German gender system. *Buffalo Cognitive Science Report* 2.
- Zubin, David A. & Klaus-Michael Köpcke. 1986. Gender and folk taxonomy: The indexical relation between grammatical and lexical categorization. In Colette G. Craig (ed.), *Noun classes and categorization*, 139–180. Amsterdam: John Benjamins.
- Zwicky, Arnold M. 1986. The general case: Basic form versus default form. In *Annual Meeting of the Berkeley Linguistics Society*, vol. 12, 305–314.



# Language index

Bosnian, Croatian and Serbian, 105–  
111

Derivation, 111–117  
Dutch, 5

English, 6, 24  
Estonian, 22

German, 6, 7, 11  
Gothic, 13–15

Hausa, 130–135  
Highland Otomi, 127–129

Kasem, 164–207

Latin, 7, 77–81

Romanian, 81–103  
Russian, 111–117

Sanskrit, 48  
some language, *see* some other lan-  
guage  
*see also* some other lect also  
of interest

Spanish, 6, 8, 10, 139–163  
Swahili, 119–126



# Subject index

- Accuracy, 64
- Analogy and rules, 28–30
- ATC, 42–53
- Class distance, 69
- Clustering, 69
- Expected Accuracy, 64
- False Negative, 64
- False Positive, 64
- Feature structures, 37–38
- Inflection, 139–207
- Kappa score, 64
- Multiple rule systems, 22–26
- Neural networks, 26
- Phonologically conditioned allomorphy, 20
- Prefixation, 119–126
- Proportional analogy, 8–11
- Rule systems, 17–20
- Schemata, 20
- Softmax, 60
- some term, *see* some other term
  - see also* some other term
  - also of interest
- True Negative, 64
- True Positive, 64
- Type hierarchies, 38–42
- Variable importance, 69







# Analogical classification in formal grammar

The organization of the lexicon, and especially the relations between groups of lexemes is a strongly debated topic in linguistics. Some authors have insisted on the lack of any structure of the lexicon. In this vein, Di Sciullo Williams (1987: 3) claim that “[t]he lexicon is like a prison – it contains only the lawless, and the only thing that its inmates have in common is lawlessness”. In the alternative view, the lexicon is assumed to have a rich structure that captures all regularities and partial regularities that exist between lexical entries. Two very different schools of linguistics have insisted on the organization of the lexicon.

On the one hand, for theories like HPSG (Pollard Sag 1994), but also some versions of construction grammar (Fillmore Kay 1995), the lexicon is assumed to have a very rich structure which captures common grammatical properties between its members. In this approach, a type hierarchy organizes the lexicon according to common properties between items. For example, Koenig (1999: 4, among others), working from an HPSG perspective, claims that the lexicon “provides a unified model for partial regularities, medium-size generalizations, and truly productive processes”.

On the other hand, from the perspective of usage-based linguistics, several authors have drawn attention to the fact that lexemes which share morphological or syntactic properties, tend to be organized in clusters of surface (phonological or semantic) similarity (Bybee Slobin 1982; Skousen 1989; Eddington 1996). This approach, often called analogical, has developed highly accurate computational and non-computational models that can predict the classes to which lexemes belong. Like the organization of lexemes in type hierarchies, analogical relations between items help speakers to make sense of intricate systems, and reduce apparent complexity (Köpcke Zubin 1984).

Despite this core commonality, and despite the fact that most linguists seem to agree that analogy plays an important role in language, there has been remarkably little work on bringing together these two approaches. Formal grammar traditions have been very successful in capturing grammatical behaviour, but, in the process, have downplayed the role analogy plays in linguistics (Anderson 2015). In this work, I aim to change this state of affairs. First, by providing an explicit formalization of how analogy interacts with grammar, and second, by showing that analogical effects and relations closely mirror the structures in the lexicon. I will show that both formal grammar approaches, and usage-based analogical models, capture mutually compatible relations in the lexicon.

ISBN 978-3-96110-186-3

