One Formal Approach Leads to Another

Marjorie McShane
New Mexico State University

1. Introduction

This paper takes a fresh look at nominal declension in Polish and argues that the notion PARADIGM must be defined according to the application at hand. It shows how the complexities of Polish inflection were handled in one computational application (building a morphological analyzer using machine learning) and suggests that the highly explicit approach required by that application could be exploited in other realms as well—e.g., to comprehensively describe Polish inflection and to lighten the cognitive load for humans attempting to master it.

The question of paradigm delineation in Polish arose in connection with the following task: testing the morphological learner in a knowledge elicitation system intended to cover any natural language.[1] The morphological learner takes as input paradigms supplied by a language informant. On the basis of these paradigms—and only these paradigms—the learner creates rules of inflection that are iteratively tested and refined in a test-debug loop. The process works as follows: the informant provides all the forms for the so-called Primary Example for Paradigm 1; the morphological learner generates rules to cover these word forms; the informant lists a few more words that he believes belong to the paradigm; the learner produces the inflectional forms of those words based on its first round of rules; the informant corrects any mistakes; the learner relearns its rules based on those corrections; the informant tests

[1] The project in question, Expedition, is being carried out at the Computing Research Laboratory of New Mexico State University (http://crl.nmsu.edu/expedition). The goal of Expedition is to create the tools to quickly ramp up translation systems from lesser-studied languages into English (Nirenburg and Raskin 1998). The morphological learner is part of the knowledge acquisition module, called Boas (Nirenburg 1998; Oflazer and Nirenburg 1999; Oflazer, Nirenburg and McShane, forthcoming).

Steven Franks, Tracy Holloway King, and Michael Yadroff, eds. *Annual Workshop on Formal Approaches to Slavic Linguistics: The Bloomington Meeting, 2000*. Ann Arbor: Michigan Slavic Publications, 2001, 191–207.

the new rules with more words he attributes to the paradigm; and so on, until the informant believes the rules are robust enough to cover all the slight inflectional variations present in Paradigm 1 as he envisions it. Then he proceeds to Paradigm 2. The informant may create as many or as few paradigms as he deems necessary to cover all regular patterns of inflection in his language. It is irrelevant for the learning program whether the informant splits words into many narrowly specified paradigms or bunches them into more broadly defined paradigms. However, there is one absolute constraint: all inflectional forms of all words in a paradigm must be unambiguously predictable based solely on the spelling of the citation form.

Delineating machine tractable paradigms can be a formidable challenge, especially if: (i) the language has complex inflectional patterns; (ii) there are no grammars of the language or the paradigms described in grammars are computationally unsuitable; (iii) the language informant is not a linguist (a likely scenario for the Expedition project). Polish is a particularly difficult case in point because nominal inflection is affected by a dizzying array of phonological, spelling, and semantic rules, not to mention significant irregularity and unpredictability. Furthermore, the small number of paradigms posited in traditional treatments are unsuitable for machine learning because they rely on knowledge beyond that conveyed by the spelling of the citation form. Therefore, the current task required dissecting traditional paradigms and creating new ones based on more exacting common denominators.

Section 2 discusses the complexities of nominal inflection in Polish, the types of knowledge presupposed in traditional treatments, and what types of phenomena can and cannot be conflated in a paradigm for machine processing. Section 3 presents a subset of the masculine paradigms used to test the morphological learner. Section 4 concludes the paper.

2. Polish Nominal Declension: The Basics

Polish nouns inflect for seven cases (Nom., Acc., Gen., Dat., Loc., Instr., Voc.) and two numbers (Sg., Pl.). A survey of Polish grammars revealed that there is no widely accepted inventory of paradigms of the type

found, for example, in Russian grammars.[2] However, the following list approximates the intersection of what Polish grammars consider the basic inventory of paradigms (numerous small groups are omitted).

Table 1. Traditional Paradigm Delineation

Paradigm Description	Example	Gloss
Masc./Neut. alternating	*herb*	coat of arms
Masc./Neut. non-alternating	*kraj*	country
Masc. mixed	*kolega*	colleague
Fem. alternating	*gazeta*	newspaper
Fem. non-alternating	*koszula*	shirt
Fem. in a consonant	*rzecz*	thing

Positing broadly defined paradigms like these presupposes rules from various components of the language system, including the following.

2.1. Semantic Rules

Some inflectional rules derive from the inherent semantic features of nouns. For example:

- VIRILE NOUNS: For Masc. nouns denoting men, the Acc. Sg./Pl. coincides with the Gen. Sg./Pl. In addition, more than one Nom. Pl. ending is often possible for a given lexical item.
- ANIMAL, ETC. NOUNS: For masculine nouns denoting animals and a rather obscure conglomeration of other semantic classes (cigarettes, dances, games, units of currency, vehicles, fruits, vegetables), the Acc. coincides with the Gen. in the Sg. but not in the Pl.[3]

2.2. Phonological and Spelling Rules

Polish consonants are traditionally divided into hard (*b, p, f, w, m, n, s, z, t, d, r, ł, sł, st, zm*), soft (*ć/ci, dź/dzi, ś/si, ź/zi, ń/ni, ki, gi, chi, li, bi, fi, wi mi*), and functionally soft, which are phonologically hard but take soft

[2] The traditional delineation of paradigms in Russian is: 1st declension (Masc./Neut. with a hard or soft stem); 2nd declension (Fem. with a hard or soft stem); 3rd declension (Fem. ending in a soft sign).

[3] See Westfal 1956 for discussion of the masculine Gen. Sg. in Polish.

endings (*c, dz, sz, ż, c, sz, l, rz, cz, dż, cz, dż*).[4] Rules of inflection are generally written using these categories as a point of reference. Alternating consonants soften in some inflectional forms. In contrast, non-alternating consonants are not subject to softening: they either originate soft and remain soft or originate hard and remain hard. For computational purposes, the alternating/non-alternating dichotomy is in part useful but in part misleading. It is useful in that alternating and non-alternating consonants take different endings in certain forms and therefore should be split into different paradigms. It is misleading in that the grouping is based on phonetics, not spelling, so some alternating consonants are spelled the same in all inflectional forms, while some non-alternating consonants are spelled differently in different inflectional forms.

Consider, for example, the Loc. Sg. of Masc. nouns—an inflectional form in which alternating consonants show their alternations. Whereas *b* and *p* alternate but show no spelling mutations (*herb ~ herbie* 'coat of arms', *sklep ~ sklepie* 'store'), *ń* and *ść* do not alternate but do show spelling mutations (*koń ~ koniu* 'horse', *liść ~ liściu* 'leaf').[5] The two traditional rules of Loc. Sg. formation that cover these four words correspond to three machine rules of a completely different type (the machine rules are conveyed approximately):

- Traditional rule for *herb* and *sklep*: **add *-e*** [the canonical ending for alternating consonants] **and soften consonants as necessary** [so *b bi* and *p pi*].
- Traditional rule for *koń* and *liść*: **add *-u*** [the canonical ending for non-alternating consonants] **and incorporate spelling rules** [so *ń + u niu* and *ć + u ciu*].
- ➢ Machine rule for *herb* and *sklep*: **add *-ie*.**
- ➢ Machine rule for *koń*: ***ń n* and add *-iu*.**
- ➢ Machine rule for *liść*: ***ć c* and add *-iu*.**

[4] The vowel *i* is used to show the softening of certain consonants when they precede a vowel; e.g., soft *b* preceding *e* is spelled *bie*. In linguistic sources, the soft *b* itself is often conveyed as *bi* (or, alternatively, *b'*).

[5] I define "mutation" as one character changing into another character; so *ń n* is a mutation.

In short, while the traditional classification of alternating/non-alternating is linguistically appropriate and in part useful for machine processing, it masks a spelling problem that must be handled explicitly in the computational system at hand.

1.3. Lexical Idiosyncrasy

A number of inflectional properties of Polish nouns are unpredictable and must be listed explicitly in the lexicon. For example, the Gen. Sg. ending for alternating inanimate nouns can be *-u* or *-a*; the Nom. Pl. ending for alternating virile nouns can be *-owie* and/or *-y/-i*; the Gen. Pl. ending for non-alternating Masc. nouns is largely unpredictable, and often more than one form is possible for a given lexical item.

Another example of lexical idiosyncrasy involves what I will generically call vowel shifts: the insertion, deletion, or changing of a vowel in the final syllable of an inflectional form. As shown in Table 2, it is often not predictable whether a word will or will not undergo a vowel shift—that is, without resorting to incomplete and computationally intractable generalizations concerning consonant clusters and the like.

Table 2. Examples of Lexical Idiosyncrasies

Alt.	Used?	Example	Gloss
ó → o	yes	gróbNOM.SG grobuGEN.SG	grave
	no	mózgNOM.SG mózguGEN.SG	brain
ą → ę	yes	żołądźNOM.SG żołędziuLOC.SG	acorn
	no	pociągNOM.SG pociąguLOC.SG	train
ę → ą	yes	rękaNOM.SG rąkGEN.PL	hand
	no	potęgaNOM.SG potęgGEN.PL	might
Ø → e	yes	perłaNOM.SG perełGEN.PL	pearl
	no	liczbaNOM.SG liczbGEN.PL	number

Because [± vowel shift] is lexically stipulated, words with and without vowel shifts must be assigned to different paradigms: the rules for one paradigm will always include a vowel shift, while the rules for the other paradigm will not. Different types of vowel shifts can be conflated into a single paradigm as long as each vowel undergoes a predictable shift and each shift is explicitly taught to the morphological learner using an example.

The classes of factors discussed above—alternating/non-alternating stems, phonological rules, spelling rules, and lexical idiosyncrasies—must be explicitly accounted for by the paradigms the language informant feeds to the morphological learner.

Now that the complexity of Polish nominal inflection is clear, let us proceed to the concrete matter of building a morphological analyzer to deal with it. Although only a small subset of Masc. paradigms will be discussed, they represent all the problems presented by Masc., Fem. and Neut. nouns in Polish.[6]

2. Computationally Tractable Paradigms

Table 3 (opposite) lists most of the productive patterns of inflection for Masc. nouns in Polish, grouped in a computationally tractable manner.[7] The primary diagnostics are listed in the middle three columns: whether the stem ends in an alternating or a non-alternating consonant; what the Gen. Sg. ending is; whether or not there are vowel shifts.

One could, of course, split the paradigms much more finely, creating a paradigm for every slight inflectional variation and ending up with a hundred or more nominal paradigms. This solution would be perfectly acceptable for the morphological learner, but might be unwieldy for human use. The idea of this experiment was to create a relatively small number of robust paradigms that could be manipulated by both humans and machines. Since the morphological learner first generalizes on the basis of consonants and vowels, then produces consonant- and vowel-specific rules if conflicts occur, letter-specific behavior can be collapsed into a single paradigm as long as examples are provided for each letter in question.

The subsections below focus on the data and conceptual issues related to the four shaded paradigms from Table 3, which cover most of the tricky issues raised by Polish nominal inflection. Vocative forms

[6] McShane, forthcoming, develops this approach to paradigm delineation for all open-class parts of speech in Polish.

[7] These paradigms were among those used to test the morphological learning program. The architecture of the learner, the testing procedure, and a detailed account of the results for a subset of Polish Masc. nouns are described in Oflazer, Nirenburg, and McShane, forthcoming.

Table 3. Computationally Tractable Masc. Paradigms

Animacy	Alt.?	Gen. Sg.	Vowel Shifts?	Example	Gloss
inanimate	+	-u	-	telefon	telephone
inanimate	+	-u	+	grób	grave
inanimate	+	-a	-	gram	gram
inanimate	+	-a	+	ząb	tooth
inanimate	-	-u	-	garaż	garage
inanimate	-	-u	+	pokój	room
inanimate	-	-a	-	bicz	whip
inanimate	-	-a	+	nóż	knife
animal	+	-a	-	krab	crab
animal	+	-a	+	wół	ox
animal	-	-a	-	koń	horse
animal	-	-a	+	wąż	snake
virile	+	-a	-	pasierb	stepson
virile	+	-a	+	majster	master
virile	-	-a	-	słuchacz	listener
virile	-	-a	+	cudoziemiec	foreigner
mixed	+	-y	-	poeta	poet
mixed	-	-y	-	kierowca	driver
etc.					

excluded from the test data because the nascent system is primarily intended for journalistic prose, where relatively few vocative forms are expected to occur. In addition, inflectional forms that might not be semantically valid (e.g., plurals for collectives) were permitted; this bit of overgeneration is irrelevant since Polish text will be only be analyzed, not generated, by this system.[8]

[8] I must emphasize that the morphological analyzer of Polish built through the Boas knowledge-elicitation system is not intended to compete with morphological analyzers designed expressly for Polish. Polish is simply being used as a test case for a system that is intended for languages for which there are few or no available machine resources.

2.1. Paradigm 1

Paradigm 1 includes *alternating inanimate Masc. nouns with Gen. Sg. in -u and no vowel shifts*. The Primary Example (used as the first, fully-specified example provided to the morphological learner) is *telefon* 'telephone', whose inflectional forms are shown below.

Sg.	*Nom.*	telefon	*Pl.*	*Nom.*	telefon**y**
	Acc.	telefon		*Acc.*	telefon**y**
	Gen.	telefon**u**		*Gen.*	telefon**ów**
	Dat.	telefon**owi**		*Dat.*	telefon**om**
	Loc	telefon**ie**		*Loc.*	telefon**ach**
	Instr.	telefon**em**		*Instr.*	telefon**ami**

All inflectional forms in this paradigm are trivial except:

- The Loc. Sg. depends upon the stem-final consonant, some of which undergo alternations.

Final Consonant	**Loc. Sg. Ending**	**Consonant Alternations**
b, p, f, w, m, n, s, z	-ie	
t, d, st, zm	-ie	t c, d dz, st śc, zm źm
ł, r, sł	-e	ł l, r rz, sł śl
g, k, ch	-u	

- The Instr. Sg. depends upon the stem-final consonant; two velars have an idiosyncratic ending.

Final Consonant	**Instr. Sg. Ending**
b, p, f, w, m, n, s, z, t, d, st, zm, ł, r, sł, ch	-em
g, k	-iem

- The Nom. Pl. depends upon the stem-final consonant; two velars have an idiosyncratic ending.

Final Consonant	Nom. Pl. Ending
b, p, f, w, m, n, s, z, t, d, st, zm, ł, r, sł, ch	-y
g, k	-i

Based on just the Primary Example, the learner knows nothing about the inflectional details presented in the bulleted list above; therefore, more examples must be provided. It is not necessary to provide all the inflectional forms of each additional example—only those that cannot be predicted based on the forms of the Primary Example. Therefore, the paradigm specification for Paradigm 1 will consist of a fully-specified Primary Example plus partially specified additional examples. Once the necessary data have been provided, the morphological learner creates rules which then must be tested on a series of new examples that cover each of the slight inflectional variations found in the paradigm.

Preparing for the teach-test-debug cycle required collecting letter-specific examples that decline precisely alike—inventories not found in available resources. A subset of the words collected for Paradigm 1 is listed below, arranged in a four-tier manner based primarily on the form of the Loc. Sg.

The group labeled LOC. SG. IN -IE contains consonants that alternate phonetically but not graphotactically; their Loc. Sg. ending is *-ie*. Words ending in all of these consonants should be covered by the rules generated for the Primary Example. However, there is one complication: since this paradigm will ultimately contain certain letter-specific rules for the Loc. Sg. (described below), the generalization of rules to any consonant gets corrupted during the process of machine learning. Therefore, Loc. Sg. examples for at least one other stem-final letter from this first group had to be added as a control during the test-debug loop.

The group labeled CONSONANT ALTERNATION AND LOC. SG. IN -IE contains words with letter-specific consonant alternations; their Loc. Sg. is *-ie*. At a minimum, the following three forms had to be provided for a word ending in each letter: the Nom. Sg. (the base form), the Loc. Sg. (the unpredictable, mutated form), and at least one other non-mutated inflectional form, which counters overgeneralization of the mutation.

The group labeled CONSONANT ALTERNATION AND LOC. SG. IN -E contains more words with letter-specific consonant alternations; their Loc. Sg. is in *-e* (as opposed to *-ie*, as above). The data requirements for the previous group apply here as well.

The final group contains words with stem-final velars, which have the following special properties, referred to hereafter as VELAR PECULIARITIES: Loc. Sg. in *-u*; Instr. Sg. in *-iem* for stem-final *g/k* but in *-em* for stem-final *ch*; and Nom. Pl. in *-i* for stem-final *g/k* but in *-y* for stem-final *ch*.[9]

Loc. Sing. in -ie

b	p	f	w	m	n	s	z
pogrzeb	sklep	aperitif	motyw	tłum	telefon	adres	nakaz
herb	postęp	klif	krzew	film	egzamin	autobus	obraz

Consonant Alternation and Loc. Sing. in -ie

t/c	d/dz	st/ść	zm/źm
akcent	sad	most	komunizm
plakat	wypad	list	socjalizm

Consonant Alternation and Loc. Sing. in -e

ł/l	r/rz	sł/śl
artykuł	teatr	pomysł
kawał	kolor	zmysł

Velars

g	k	ch
pociąg	bank	dach
brzeg	atak	śmiech

Apart from being machine tractable, this method of paradigm organization has a number of interesting properties relating to language

[9] In order to preserve clarity of presentation, some glosses have have been moved to footnotes. Letter-by-letter glosses for this set of examples are as follows: **b**: funeral, coat of arms; **p**: store, ruse; **f**: aperitif, cliff; **w**: motive, bush; **m**: crowd, film; **n**: telephone, exam; **s**: address, bus; **z**: order, painting; **t**: accent; poster; **d**: orchard, outing; **st**: bridge, letter; **zm**: communism, socialism; **ł**: article, chunk; **r**: theater, color; **sł**: idea, sense; **g**: train, bank; **k**: bank, attack; **ch**: roof, laughter.

description, teaching, and learning. First, since each group is so highly specified, even someone who knows nothing about Polish could decline any of the words in this paradigm based on the Primary Example plus the description of each group. Second, having multiple examples for every stem-final letter provides rich material for practice (for reasons of space, only two examples were listed for each stem-final consonant, but many more were collected for this project). Third, the four-tier layout of this paradigm organizes the notable letter-specific variations in a visually memorable way. Finally, this highly precise definition of a paradigm reinforces the factors that affect paradigm membership in Polish: alternating/non-alternating, animate/inanimate, vowel shifts/no vowel shifts, etc.

I am not suggesting that linguistically insightful generalizations be dismissed in the descriptive and pedagogical realms. Rather, I am suggesting a division of labor of the type that has long been discussed with respect to the one-stem verb system in Slavic: it is good and helpful to know the theory, but when it comes to brute memorization and the mysterious process of internalizing a second language, simple rote patterns have their place.

2.2. Paradigm 2

Paradigm 2 includes alternating inanimate Masc. nouns with Gen. Sg. in *-u* and vowel shifts. The Primary Example is *grób* 'grave':

Sg.	*Nom.*	grób	*Pl.*	*Nom.*	groby
	Acc.	grób		*Acc.*	groby
	Gen.	grob**u**		*Gen.*	grob**ów**
	Dat.	grob**owi**		*Dat.*	grob**om**
	Loc	grob**ie**		*Loc.*	grob**ach**
	Instr.	grob**em**		*Instr.*	grob**ami**

This paradigm has all the same properties as Paradigm 1 except that there are vowel shifts whose occurrence is not predictable based on the citation form (e.g., *grób* has vowel shifts but graphotactically similar *mózg* does not). The vowel shifts occur in all inflectional forms except the Nom. Sg. and the Acc. Sg., which are identical. It was not possible to find examples representing all consonant alternations in combination

with all vowel shifts; therefore, this paradigm currently covers only the combinations for which examples were found.[10]

Vowel Shifts

Nom./Acc. Sg.	Other Forms
ó	o
ė	
ie	
a	e

Loc. Sg. in *-ie*

b	p	w
grób → grobie dąb → dębie	półwysep → półwyspie	parów → parowie rów → rowie

n	z
len → lnie sen → śnie	mróz → mrozie nawóz → nawozie

Consonant Alternation and Loc. Sg. in -ie

t	d/dzi	zd/źdz
obrót → obrocie odwrót → odwrocie	lód → lodzie błąd → błędzie	dojazd → dojeździe najazd → najeździe

Consonant Alternation and Loc. Sg. in -e

ł/l	r/rz
stół → stole dół → dole	bór → borze cukier → cukrze

Velars

k	g	ch
budynek → budynku pakunek → pakunku	róg → rogu okrąg → okręgu	mech → mchu

[10] Glosses for the next set of examples are: **b**: grave, oak; **p**: peninsula; **w**: ravine, ditch; **n**: sleep, flax; **z**: frost, fertilizer; **t**: revolution; retreat; **d**: ice, mistake; **zd**: approach, invasion; **ł**: table, pit; **r**: forest, sugar; **k**: building, package; **g**: horn, circle; **ch**: moss.

2.3. Paradigm 3

Paradigm 3 includes *alternating virile nouns with no vowel shifts*. The Primary Example is *pasierb* 'stepson':

Sg.	*Nom.*	pasierb	*Pl.*	*Nom.*	pasierb**owie** pasierb**i**
	Acc.	pasierb**a**		*Acc.*	pasierb**ów**
	Gen.	pasierb**a**		*Gen.*	pasierb**ów**
	Dat.	pasierb**owi**		*Dat.*	pasierb**om**
	Loc.	pasierb**ie**		*Loc.*	pasierb**ach**
	Instr.	pasierb**em**		*Instr.*	pasierb**ami**

In this paradigm, all of the consonant alternations and velar peculiarities discussed above are still in effect, making the four-tier system of presentation relevant. (I omit the examples here as this convention has been amply illustrated in the earlier paradigms.) This virile paradigm differs from inanimate Paradigm 1 in the following ways:

- The Acc. Sg./Pl. coincides with the Gen. Sg./Pl.
- The Nom. Pl. ending depends both upon the word-final consonant and on idiosyncrasies of the word itself. There are five possibilities: (i) *owie*, (ii) *i* (iii) *y*, (iv) *owie* or *i* (v) *owie* or *y*.[11]

Word-final consonant	**Nom. Pl. endings**
b, f, w, m, n, z, t	*owie* or *i* or both
p, ch	*i* only
d, ł	*owie* only
r, k, g	*owie* or *y* or both

Which of the Nom. Pl. endings will be valid for a given word is largely unpredictable, as shown by the sample words below (grammar books of Polish often present conflicting information regarding licit variants). The rest of the alternating consonants in this paradigm show a similar degree of unpredictability.

[11] Two points deserve mention. First, *i/y* are allomorphs in complementary distribution. Second, although there are no *letters* that exclusively permit *y*, there are *words* ending in *r, k,* and *g* that exclusively permit *y*.

	Cit. Form	*owie*	*i*	Gloss
b	Arab	Arabowie	Arabi	Arab
	pasierb	——	pasierbi	stepson
p	biskup	——	biskupi	bishop
	chłop	——	chłopi	peasant
f	filozof	filozofowie	——	philosopher
	szef	——	szefi	boss
w	Nowakow	Nowakowie	——	*proper names*
	Bogusław	——	Bogusławi	
m	kum	kumowie		godfather
	agronom	agronomowie	agronomi	agronomist
n	kapitan	kapitanowie	——	captain
	kuzyn	——	kuzyni	cousin
	piastun	piastunowie	piastuni	guardian

For purposes of the morphological learner—and, later, the morphological analyzer—this unpredictability raises no problems: both variants (*owie* and the correct one of the *i/y* allomorphs) will be permitted for every word. This bit of overgeneration is irrelevant since the analyzer will only be analyzing, not producing, inflectional forms. However, since the morphological learner has no way to predict which of the *i/y* allomorphs is used with a given word-final consonant, explicit examples of the Nom. Pl. for each word-final consonant had to be provided. As regards human consumption, no method of paradigm delineation can ease the brute memorization required to produce *filozofowie* in the same breath as *szefi*.

2.4. Paradigm 4

Paradigm 4 includes non-alternating inanimate Masc. nouns with Gen. Sg. in -a and no vowel shifts. The Primary Example is bicz 'whip':

Sg.	*Nom.*	bicz	*Pl.*	*Nom.*	bicze
	Acc.	bicz		*Acc.*	bicze
	Gen.	bicza		*Gen.*	biczy
	Dat.	bicz**owi**		*Dat.*	bicz**om**
	Loc.	biczu		*Loc.*	bicz**ach**
	Instr.	bicz**em**		*Instr.*	bicz**ami**

Unlike alternating consonants, non-alternating consonants show no Loc. Sg. mutations, making the four-tier system of organization unnecessary. However, a spelling rule comes into play: word-final letters written with a diacritic lose their diacritic and are followed by *i* when a vocalic ending is added.

ń + u niu	ć + u ciu
ń + owi niowi	ć + owi ciowi
etc.	*etc.*

One other complication arises in this paradigm: the Gen. Pl. ending depends upon both the final consonant and the lexical properties of the given word, just as we saw for the Nom. Pl. of virile nouns in Paradigm 3. For purposes of testing the morphological learner, I limited the Gen. Pl. endings to one for each stem-final consonant; however, building a complete analyzer would require the same type of overgeneration allowed in Paradigm 3.

Final Cons.	**Gen.Pl. End.**	**Cit. Forms**	**Glosses**
cz	y	klucz, bicz	key, whip
sz	y	kapelusz, klawisz	hat, key (of a piano)
rz	y	ołtarz, korytarz	altar, corridor
ż	y	krzyż	cross
l	i	parasol, badyl	umbrella, stalk
ść	i	liść	leaf
ń	i	kamień, strumień	rock, stream
j	ów	kij, liszaj	stick, lichen
ch	ów	brzuch, kielich	stomach, glass
szcz	ów	płaszcz	overcoat

This paradigm emphasizes the need to deal with spelling conventions, not just phonetic properties of words, when establishing paradigms. It also reinforces the need for creating an inventory of letter-specific examples to teach and test the morphological learner.

3. Discussion

The rules created through this machine learning process are bi-directional, meaning that they can be used for generation as well as parsing. However, generation imposes a restriction that analysis does not: for generation, each word must be associated with one and only one paradigm.[12] Accordingly, nominal entries in computational lexicons of Polish would have to be expanded to ensure unambiguous paradigm assignment. One option would be to create a full inventory of inherent features and mark each noun in the lexicon for the relevant ones: MASC / FEM / NEUT; INANIMATE / ANIMATE NON-VIRILE / VIRILE; GEN.SG. IN A / U; ALTERNATING / NON-ALTERNATING, etc. The specific combination of inherent features, in conjunction with the spelling of the citation form, would place a noun in a specific paradigm. Another option would be to first create the full inventory of paradigms then manually assign each noun in the lexicon to one of them. Irregular or truly unpredictable forms would have to be listed as exceptions under either approach.

The paper has shown that although the computational approach increases the number of paradigms used to describe Polish declension, it circumvents the necessity of incorporating layers of umbrella rules. Thus, the computational canvas is much larger, but also much clearer. I believe that this degree of clarity and explicitness could be fruitfully incorporated into formal descriptions of, and pedagogical approaches to, Polish inflection.

[12] For parsing, the important thing is that the correct analysis is among those posited. Therefore, having multiple analyses is not necessarily problematic, especially since various means can be use to filter the analyses. Of course, having a single analysis is best and could be achieved using the type of lexical expansion described here. However, the task of fully specifying the entire lexicon would be too time consuming for the typical user of our system.

References

McShane, M. Forthcoming. "Polish inflection fit for man and machine," in *Memoranda in computer and cognitive science, Computing Research Laboratory,* New Mexico State University.

Nirenburg, S. 1998. "Project Boas: 'A linguist in a box' as a multi-purpose language resource," in *Proceedings of COLING '98.*

Nirenburg, S. and V. Raskin. 1998. "Universal Grammar and Lexis for Quick Ramp-Up of MT Systems," in *COLING-ACL '98* (36th Annual Meeting of the Association for Computational Linguistics), vol. II, 975-979.

Oflazer, K. 1996. "Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction," in *Computational Linguistics,* 22(1):73-90.

Oflazer, K. and S. Nirenburg. 1999. "Practical Bootstrapping of Morphological Analyzers," in *Proceedings of the Workshop on Computational Natural Language Learning at EACC '99,* Bergen, Norway.

Oflazer, K., S. Nirenburg, and M. McShane. Forthcoming. "Bootstrapping morphological analyzers by combining human elicitation and machine learning," in *Computational Linguistics.*

Westfal, S. 1956. *A study in Polish morphology: the Genitive singular masculine*. The Hague, Mouton.

Computing Research Laboratory
Box 30001/MSC 3 CRL
New Mexico State University
Las Cruces, NM 88003
marge@crl.nmsu.edu