Thirteen

Marjorie McShane

Out of the Box

1. Introduction

I was recently reminded of the lightbulb puzzle, whose solution I had long since forgotten. The jist: there are three switches outside a room connected to three lightbulbs inside. You must determine which switch goes with which bulb but may only peek inside the room once. In response to my babbling about states of switches, the Keeper of the Answer remarked: "Nope, you're still in the box." Indeed, reaching the solution—which involves turning on one bulb and allowing it to heat up before turning it off, subsequently using temperature as the missing distinguishing feature—requires a radical shift in thinking, or crawling out of the box.[1]

Crawling out of the box struck me as the perfect analogy for my recent reanalysis of Polish morphology, the results of which have little in common with traditional approaches. Whereas traditional approaches are driven by linguistic or lexicographic principles, my approach was driven by a practical computational application. Since this application *forced* a departure from tradition, one might amend the analogy to my being kicked out of the box. Once out, however, I realized that my newly gained insights into Polish morphology might have direct application in other realms, like formal language description and pedagogy. This paper describes a few of those insights, to be found only in the great expanse outside of the box.

2. The Computational Task

The necessity of taking a new approach to Polish inflection derived from the following task: to teach a machine-learning program all the patterns of inflection in Polish solely by feeding it sample inflectional paradigms—making no reference to umbrella rules covering spelling conventions, consonant and vowel alternations, semantic classes, etc.[2] Despite the apparent simplicity of the task,

[1] The solution in more detail: First, turn on switch 1, allow that bulb to heat up, then turn it off. Next, turn on switch 2 and leave it on. Then go in the room and check the bulbs: the hot off bulb is connected to switch 1, the on bulb to switch 2, and the cold off bulb to switch 3.

[2] Polish served as a test case during development of the morphological learning algorithm for the Boas knowledge-acquisition module of the Expedition Project, supported

Laura A. Janda, Steven Franks, and Ronald Feldstein, eds. *Where One's Tongue Rules Well: A Festschrift for Charles E. Townsend. Indiana Slavic Studies* 13: 147–55, 2002.

it posed a significant challenge because traditional accounts of Polish inflection define paradigm loosely, collapsing into a single paradigm lexical items with significant inflectional variation. There is good reason for this: since many types of rules cut across not only paradigms within a part of speech but also across parts of speech, they are most succintly accounted for globally from the outset then assumed ever after. However, this almost universally adopted approach to Polish inflection blurs distinctions that are crucial for machine processing and, I will argue, for human understanding as well.

3. Traditional Descriptive and Lexicographic Approaches

Although the computational task extended to all open-class parts of speech, here I focus on verbal inflection, or conjugation.[3] The available descriptions of Polish conjugation occupy extreme points along a spectrum ranging from bunching to splitting paradigms. On the bunching end are general grammars and reference books (e.g., Bielec 1998, Janecki 2000, Kaipio 1977, Kaleta 1995) that delineate 3 or 4 major paradigms with various numbers of subtypes: the *ę–esz, ę–isz/ysz, am–asz,* and *em–esz* paradigms, the last two of which are conflated in three-paradigm schemes.[4] These cover the basic, productive patterns of conjugation, glossing over details and, for the most part, barring exceptions. Paradigm membership is based solely on present-tense forms, with all other tenses and all non-indicative moods being dealt with cross-paradigmatically. This linguistically elegant, streamlined, and academically satisfying approach will be referred to hereafter as the traditional descriptive approach.

At the splitting end of the spectrum is Mędak's comprehensive dictionary of Polish verb forms (*Słownik form koniugacyjnych czasowników polskich*), which covers 17,000 verbs divided into 334 paradigms. Here, the diagnostics for paradigm membership include all tenses and moods, alternations, variations, etc. (but even Mędak resorts to noting certain types of exceptions in footnotes in order to keep the number of paradigms from skyrocketing). This dictionary contains exhaustive information about Polish verb forms, but, as a dictionary following lexicographic principles, it presents this information in a manner suitable only for reference. It does not attempt to organize material in an insightful manner for learners.

---

by Department of Defense Contract MDA904-92-C-5189 (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 (McShane *et al.* 2000, Nirenburg and Raskin 1998, Oflazer *et al.*, Forthcoming). Although other approaches to machine learning may permit the inclusion of global rules, this one does not.

[3] McShane (In press) discusses nominal inflection within this framework.

[4] Bielec and Kaleta are reference grammars. Janecki and Kaipio are manuals of fully-inflected verb forms that are prefaced with a description/categorization of verb forms.

Some notable features that underscore Mędak's goal of reference rather than pedagogy are the following. (i) The index contains some 17,000 verbs, organized alphabetically and referenced by the paradigm number; however, there are no inventories of verbs belonging to each paradigm—a feature that would be invaluable for training purposes.[5] (ii) The paradigms are ordered in a manner whose logic is not readily apparent. All imperfective paradigms are listed in section 1 (using numbers 100–199.34) and all perfective paradigms are listed in section 2 (using numers 200–299.89), but there is no predictable numerical correspondence between imperfectives and perfectives sharing the same stem. This lack of correspondence is justified by the fact that not all imperfectives in a given 100-series paradigm have perfectives in the same 200-series paradigm; however, it loses certain generalizations that could—indeed, must—be exploited by learners, like the formal correlation between imperfective and perfective verbs with a common stem. Mędak's approach will be referred to hereafter as the lexicographic approach because the presentation is driven by lexicographic practices like the unabridged listing of forms, indexing, and the adherence to formal rather than conceptual organizational principles.

Both the traditional descriptive approach and the lexicographic approach are valid in terms of their stated or implied goals. The problem is that goals occupying the middle of the spectrum have not been touched in the literature—for example, the goal of offering relatively exhaustive coverage of Polish inflection organized in a conceptually elucidating manner. Although in my experience this goal arose for computational purposes, it is relevant for advanced students of Polish as well. That is, the perfectly explicit input and multiple comparable examples demanded by a machine learning program are also required by human learners seeking active mastery of inflectional patterns. In more personal terms, I found that the information I was feeding to the computer was precisely the information I wanted as a student of Polish. Thus, this new approach to describing Polish morphology is fit for both man and machine.[6]

4. Overturning Old Conventions

The world of Polish conjugation can be divided up in countless ways. Devising a new way involved overturning some long-accepted conventions and replacing them with new ones, whose justification derives from practical experience. The following are some novel organizing principles for the current approach to paradigm delineation.

---

[5] A large part of my work on Polish conjugation involved manually creating inventories of verbs belonging to each paradigm, based on Mędak's index.

[6] The full realization of this approach to inflectional patterns for Polish open-class items (nouns, verbs, adjectives and adverbs) can be found McShane (2001).

ALL TENSES AND MOODS AFFECT PARADIGM DELINEATION. While Mędak's paradigms cover all tenses and moods, traditional descriptive approaches do not—they base paradigms on the present tense alone, leaving all other tenses and non-indicative moods to be dealt with separately. The justification for this is that global rules provide fairly good coverage of non-present tenses and non-indicative moods. But although global rules are a good starting point for learning, when it comes down to the drudgery of internalizing word forms, having fully specified paradigms at hand is indispensable—far more reliable than attempting to apply rules from a dozen different chapters of a textbook and hoping they hold without exception.

THE INFINITIVE IS THE ONLY BASE FORM. This seemingly obvious tenet is, in fact, thwarted by traditional descriptive treatments of Polish conjugation. That is, the infinitive and the present (or future, for perfectives) stem are considered equally basic and are assumed known from the outset. Such an assumption simplifies life for the descriptive linguist but is unhelpful for human learners, who look up a verb in the dictionary, find the infinitive and want to know what to do with it. In Mędak's treatment, most paradigms cover only one type of infinitive. Similarly, in my treatment the top-level diagnostic for paradigm delineation is the infinitival ending; from there, further paradigm division proceeds. My grouping of citation forms, along with their number of members, is as follows: *-ać* (23), *-ąć* (6), *-awać* (1), *-c* (10), *-eć* (13), *-ić* (17), *-ieć* (13), *-iwać* (2), *-nąć* (8)[7], *-ować* (2), *-ść* (16), *-uć* (1), *-yć* (6), *-ywać* (2), *-źć* (5). Although this total of 125 paradigms might seem excessive, it includes the majority of verbs that are generally thrown into the irregular trash heap (like most verbs in *-ść*, for example, since they show suppletion). The bottom line is, whether one is building a computational system or studying a language, the irregular verbs and small inflectional groups must be covered, and there is no benefit in systematically avoiding the messiness they introduce.

LAYERS OF PHONOLOGICAL, SEMANTIC, AND SPELLING RULES ARE REFERRED TO OVERTLY IN EACH PARADIGM, RATHER THAN ASSUMED GLOBALLY. One means by which traditional descriptive approaches achieve succinctness is by initially presenting phonological, semantic and spelling rules as global rules to be taken as a given forever after.[8] Knowledge of such rules

[7] Actually, there is a huge number of paradigms in *-nąć*, deriving from the fact that *ną* may be dropped, retained, or optional in various subsets of forms. This is the only group of verbs for which I resort to a paradigm/subparadigm division: the seven basic paradigms are underspecified—excluding the past, conditional, and verbal noun—and these missing portions are presented as subparadigmatic patches.

[8] Such rules include: vowel and consonant alternations, and the inflectional forms in which they occur; which vocalic endings follow which stem-final consonants; the effects of a virile subject on conjugation; etc.

can be helpful, but incorporating them on the fly is far from trivial. The current approach makes overt reference to such rules in the paradigm description.

MULTIPLE EXAMPLES OF EACH PARADIGM ARE NECESSARY (IF AVAILABLE). A shortcoming of most grammars and textbooks is their paucity of examples. If repetition is truly the mother of learning, learners need entities to repeat. Furthermore, there should be no guesswork involved in deciding that some new word takes one set of endings or another—that information should be unambiguously provided.

DIAGNOSTICS ARE THE KEY. Fundamental to mastering a complex system is clearly stating the sources of complexity. Therefore, each section in my inventory begins with a summary table showing the parameters that divide infinitives of the given type into paradigms. Table 1 shows the relevant diagnostics for the relatively simple group of verbs in *-ąć*. In some instances, pfv. and impfv. verbs are collapsed into a single paradigm. In other instances, separate paradigms are required.[9]

Table 1. VERBS IN *-ąć*

Type	Present/Future	Imperative	Other Properties	Primary Ex.
1	nę, niesz, ną	nij		żąć/nażąć
2	nę, niesz, ną	ij	suppletion	ciąć/dociąć
3	nę, niesz, ną	ij	prefix alt.	giąć/zgiąć
4	nę, niesz, ną	nij	suppletion	/przysiąć
5	mę, miesz, mą	mij		/zadąć
6	mę, miesz, mą	mij	prefix alt.	/zdjąć
7	mę, miesz, mą	∅	z ~ ź	/wziąć

Explicitly stating the relevant diagnostics for infinitives of each type serves as a map through the maze of Polish inflection: it says "If you have an infinitive ending in X, here's what to look out for."

5. Reducing Mędak's Inventory

The obvious question is, if one takes a paradigmatic approach to Polish inflection and wants to achieve broad coverage, is it possible to reduce Mędak's 336 paradigms to a smaller inventory? It turns out that it *is* possible with the loss of only a small degree of precision. Below are the compromises I made in order to reduce Mędak's inventory by about 2/3.

[9] The notation */verb* indicates that the paradigm includes only pfv. verbs; the notation *verb/* indicates that the paradigm only includes impfv. verbs.

Mędak places impfv. and pfv. verbs with the same stem in different paradigms because imperfectives and perfectives show two fundamental inflectional differences: (i) the pfv. has no present tense and no compound future tense—what looks like the present tense has future meaning; (ii) the impfv. and pfv. have different inventories of participles. Under my approach, impfv. and pfv. verbs are placed in the same paradigm if they behave the same except for these two factors because, for purposes of learning formal patterns of inflection, capturing the overlap is more valuable than focusing on the differences (which can be noted using simple formatting conventions, as shown below).

Another main reason for paradigm splitting in Mędak is the presence, absence, or rarity of certain participial forms. While a comprehensive lexicographic description warrants splitting paradigms based on such details, this is too fine-grained to be useful to non-native speakers.[10] Glossing over such absences in the inventory of participial forms permits a significant reduction in the number of paradigms.

Another source of paradigm splitting is the existence of variant forms, which are quite common in Polish. For a non-native speaker, learning one valid form should be more than sufficient for active use, and recognizing other forms passively should pose no great obstacles once the general notion of variation has been presented. For these reasons, variations within paradigms are only selectively listed.

6. A Sample Paradigm

A sample of my approach to paradigm presentation is given in the table opposite; it is Type 1 among the *-ąć* verbs. As shown in the heading line, its only distinguishing features are the present tense in *nę, niesz, ną* and the imperative in *nij*.

The inflectional pattern is shown using the primary examples *żąć/nażąć*. Tenses and non-indicative moods are indicated, but values for person and number are not, since any learner past the first few months of study will know the routine. Only single-word forms are shown since multi-word entities (e.g., *będę żął*em) are entirely predictable based on the single-word forms. The goal for the human variant of the paradigm tables is, after all, to provide all unpredictable information in the most succinct and organized way possible. For the machine variant, a script was written to convert these tables into fully-specified inventories of impfv. and pfv. single- and multi-word forms.

[10] The matter of missing forms can be glossed over in the machine application as well because the morphological analyzer resulting from the learning algorithm will only decode, not produce, inflected forms.

TYPE 1 -ĄĆ ⇳ NĘ, NIESZ, NĄ ⇳ NIJ

ŻĄĆ/NAŻĄĆ

PRES./ FUT.	PAST: MASC.	PAST: FEM.	PAST: NEUT.
żnę	żąłem	żęłam	żęłom
żniesz	żąłeś	żęłaś	żęłoś
żnie	żął	żęła	żęło
żniemy	żęliśmy | żęłyśmy[11]	żęłyśmy	żęłyśmy
żniecie	żęliście | żęłyście	żęłyście	żęłyście
żną	żęli | żęły	żęły	żęły

IMPERATIVE	CONDIT. MASC.	CONDIT. FEM.	CONDIT. NEUT.
-	żąłbym	żęłabym	żęłobym
żnij	żąłbyś	żęłbyś	żęłobyś
niech żnie	żąłby	żęłby	żęłoby
żnijmy	żęlibyśmy / żęłybyśmy	żęłybyśmy	żęłybyśmy
żnijcie	żęlibyście / żęłybyście	żęłybyście	żęłybyście
niech żną	żęliby / żęłyby	żęłyby	żęłyby

	IMPERFECTIVE	PERFECTIVE
PRES. ACTIVE	żnący, -a, -e, -y, -e	—
PRES. PASSIVE		—
PAST ACTIVE	—	—
PAST PASSIVE	żęty, -a, -e, -ci, -te	nażęty, -a, -e, -ci, -te
PRESENT GERUND	żnąc	—
PAST GERUND	—	nażąwszy
VERBAL NOUN	żęcie	nażęcie
IMPERS. PAST	żęto	nażęto

	INFINITIVE	*ENGLISH*		*INFINITIVE*	*ENGLISH*
	żąć	to reap, mow		**począć**	to conceive
	nażąć	to reap		napocząć	to cut
	kląć	to curse		rozpocząć	to begin
	wykląć	to curse		spocząć	to sit
	przekląć	to curse		odpocząć	to rest
	skląć	to swear at			

The first two sections show finite forms. The PRES./FUT. label for the first set of forms implies that the impfv. present is formally the same as the pfv. future. The genders are separated for the past tense and conditional since spelling

[11] Masculine virile forms in the past and conditional use *li* whereas masculine non-virile forms use *ły*.

rules (like the *ą/ę* alternation) come into play for some verbs. Although certain past and conditional forms are rare in the neuter, they are listed since they could potentially occur: for instance, a neuter machine in a sci-fi book might be a first-person speaker.

The third section contains participles, gerunds, and verbal nouns, with different columns for impfv. and pfv. verbs. The present forms can only apply to impfv. verbs and the past forms to pfv. verbs. The shorthand for showing different endings for different person/number/gender combinations (e.g., *żnący, -a, -e, -y, -e*) will be readily understood by humans and specified for machines in the conversion script.

The last section contains additional examples of verbs belonging to the paradigm. Verbs with the same stem are grouped together, with the impfv. (if there is one) in boldface. No attempt at full coverage was made; the idea is to present a sampling for training purposes. Since the nuances of prefixed verbs are often hard to capture in translation, I followed the lead of large bilingual dictionaries and ignored them, especially since the translations are merely intended to provide an anchor for learners. Mędak's paradigm numbers are noted in gray to the left of each Polish infinitive, should a user want to see his full treatment and comments. In this example, I collapse two of Mędak's impfv. paradigms and four of his pfv. ones—reducing 6 paradigms to 1. The differences I overlooked in doing so are minor: for example, unlike #137, #199.6 has no passive participle; unlike #272, #299.25 has a rare verbal noun and impersonal past, and #299.87 has a rare Masc. Personal Pl. passive participle.

7. Concluding Remarks

The main point of this essay lies not in the details of this particular method of slicing up the world of Polish inflection, but in the larger implications of attempting such a reanalysis in the first place. Even before being cornered into reanalysis by computational necessity, I understood that current approaches to Polish inflection did not fulfill my needs as an advanced learner. However, unless force of necessity had exerted itself, it is unlikely that I would have devoted the time and energy to developing a new approach. After all, working on theory or linguistically sophisticated description promises greater rewards... or at least more familiar ones. In retrospect, however, this project offered its own brand of *kaif*—giving me free rein to overturn accepted tenets of scholarship, revealing an unexpected overlap between the disparate realms of natural language proceessing and pedagogy, and resulting in a final product that could potentially have widespread application. Such is life outside the box.

References

Bielec, Dana. (1998) *Polish: An essential grammar.* London and New York: Routledge.

Janecki, Klara. (2000) *301 Polish verbs.* Barron's Educational Series, Inc.

Kaipio, Clara. (1977) *201 Polish verbs.* Barron's Educational Series, Inc.

Kaleta, Zofia. (1995) *Gramatyka języka polskiego dla cudzoziemców.* Nakładem Uniwersytetu Jagiellońskiego.

McShane, Marjorie. (2001) "Polish inflection fit for man and machine". *Memoranda in computer and cognitive science,* Computing Research Laboratory, New Mexico State University. Available at http://crl.nmsu.edu/~marge/webPersonal/polishBook.html.

McShane, Marjorie. (In press) "One formal approach leads to another". Steven Franks, Tracy King and Michael Yadroff eds., *Formal approaches to Slavic linguistics: the Bloomington meeting, 2000.* Ann Arbor: Michigan Slavic Publications.

McShane, Marjorie, Stephen Helmreich, Sergei Nirenburg and Victor Raskin. (2000) "Slavic as testing grounds for a linguistic knowledge elicitation system". Tracy King and Irina Sekerina eds., *Formal approaches to Slavic linguistics: the Philadelphia meeting, 1999*, 279–95. Ann Arbor: Michigan Slavic Publications.

Mędak, Stanisław. (1997) *Słownik form koniugacyjnych czasowników polskich.* Kraków: Universitas.

Nirenburg, Sergei and Victor Raskin. (1998) "Universal grammar and lexis for quick ramp-up of MT systems". *Proceedings of COLING-ACL '98* (36th annual meeting of the association for computational linguistics), vol. II, 975–79.

Oflazer, Kemal, Sergei Nirenburg and Marjorie McShane. (Forthcoming) "Bootstrapping morphological analyzers by combining human elicitation and machine learning". *Computational linguistics.*

Computing Research Laboratory
MSC 3 CRL, P.O. Box 30001
New Mexico State University
Las Cruces, NM 88003-8001
marge@crl.nmsu.edu