

Standard and standardized verb forms in the Czech National Corpus¹

Neil Bermel

University of Sheffield

n.bermel@sheffield.ac.uk

1. Introduction

The relationship between the Czech standard code and various competing non-standard varieties presents definitional problems for linguists. On the one hand, the terms “standard” and “non-standard” are a convenient shorthand for explaining what “should” and “should not” be used in formal prose and speech. On the other hand, linguists have long recognized that the boundary between the standard and other varieties is blurred, and that *the standard* is not so much a monolithic construct as a collection of prescriptions and tendencies of varying strengths. Special difficulties occur with forms that undergo *standardization*, or admission (by various more- or less-acknowledged authorities) to the canon of standard forms.

The tension inherent in these definitions has been of concern to František Čermák, who has devoted several articles to the subject (see, for example, Čermák 1987, 1993, 1997, Čermák and Sgall 1997). In Čermák 1987, he calls attention to standard Czech’s “rigorous artificial codification” (135) and yet alongside that, notes that features of the spoken code do occur in the written language, especially in genres such as personal letters, modern fiction, and persuasive journalistic texts (140).

A second major interest of Čermák’s has been tools for and methods of describing actual Czech usage, as opposed to purely prescriptive approaches based on intuition. This line of research has resulted in the creation of the Czech National Corpus (CNC) at Charles University.

This contribution thus touches on two of Čermák’s interests in examining an example of standards and standardization as reflected in the CNC.

¹ Research for this article was supported by a travel grant from the British Academy, whose support is gratefully acknowledged.

2. The research problem

The received view of the Czech language situation is that it is stratified into Standard Czech (*spisovná čeština*, hereafter SC) and a variety of non-standard codes.

SC is the result of over a century of prescriptive activity from c. 1780-1920, in which a relatively permissive Renaissance Czech written norm was slowly revived and reinterpreted, supplanting the norms of Baroque Czech and creating a supradialectal standard that excluded many features of the contemporary spoken codes. It assumed its most conservative form by the last quarter of the nineteenth century, as reflected in the handbooks and manuals of c. 1880-1940. From that point on, a slow liberalization took hold, reintroducing into SC some of the “spoken” features that had been eliminated in previous generations. SC is used primarily in writing, but it also has a spoken variant that occurs in formal speech situations.

The most prominent spoken code is Common Czech (*obecná čeština*, hereafter CC), spoken in the western part of the country. The non-standard varieties represent native dialects or regional interdialects and are the codes people grow up speaking or learn to speak in ordinary, non-formal speech situations. There are commonly accepted ways of transcribing them, but these transcriptions do not constitute acceptable, unmarked written discourse for the vast majority of situations.

Even if we disregard those markers of “spokenness” or “writtenness” that are artefacts of the media in which they usually appear, there are still numerous places in Czech where purely formal distinctions separate the standard and non-standard codes on all levels of the language. If we consider only SC and CC, then in most instances there are two contrasting variants available, although places with three or four are not unknown.² Typically one form is associated with SC and the other with CC, meaning that one is prescribed for formal and written language, while the other is found in informal, spoken communication. The existence of two forms side by side indicates their *complementarity*, with each having its own domain of usage and consequently register-based or stylistic restrictions. In a categorizing approach, we would label one of these features as e.g. *SC* and the other as *CC*, or one as *bookish/colloquial* and the other as *neutral*, and claim that this description allowed us to predict usage.

² If we were to consider the full range of widely-spoken dialects, that number would increase dramatically.

In this study, we will examine two such related features from Czech morphology to see whether this sort of description is useful, adequate, and reflective of actual attested forms in written texts.³

3. The features

Czech verbal morphology is riddled with decision points where two forms are available, one genetically SC and the other CC. We will examine two forms of the non-past tense, the first-person singular (1 sg.) and the third-person plural (3 pl.). In the traditional six-class system, these points occur in verb classes I, V and VI.⁴

These three classes have non-past endings as follows:

	sg.	pl.
1. (“I/we”)	-u <i>or</i> -i	-eme
2. (“you”)	-eš	-ete
3. (“he/she/they”)	-e	-ou <i>or</i> -í

The deciding factor as to which ending applies in the 1 sg. and 3 pl. was traditionally whether or not the stem ends in a historically soft consonant. In classical SC, fronting occurred after a soft consonant, yielding *-i* instead of *-u* and *-í* instead of *-ou*. In CC and most other spoken varieties, a non-fronted variant was reintroduced by analogy. The 1 sg. form is common to most dialects, whereas the long vowel of the 3 pl. is variously [u:], [o:] or [ou], with the latter found in CC. As a result, verbs in the three classes have the following variants, which I will refer to respectively as the *fronted* and *analogous* forms:

³ This division into CC and SC is followed in e.g. Hammer 1985 and 1993 and Kravčičinová and Bednářová 1968, although the latter contains evidence of the inadequacy of this description.

⁴ This scheme is based on the infinitive and past tense stems, and is found in the *Průruční mluvnice češtiny* (PMČ) and many older handbooks. It provides a short-hand for distinguishing between the types represented by *krýt* and *kupovat*. In the discussion in section 4, I have treated all the handbooks using this same scheme, regardless of what actual numbering and division appears in the works themselves.

Table 1. Forms and verb classes

	SC 1 sg., 3 pl.	CC 1 sg., 3 pl.	gloss
I.	kryji, kryjí	kryju, kryjou	‘I cover, they cover’
V.	maži, maží	mažu, mažou	‘I spread, they spread’
VI.	kupuji, kupují	kupuju, kupujou	‘I buy, they buy’

Classes I and V are closed and represent a relatively small number of verbs (20-30) of high frequency. Class VI is very productive. Most borrowed verbs end up in this class, as do many imperfective verbs formed from their perfective counterparts. It includes a large group of low-frequency words, and covers the gamut from those technical and specialist in nature, through basic vocabulary items, to those perceived as slang or jargon.

4. History of standardization of *-u*, *-ou*

The changing status of the two analogous endings is shown in descriptions from the 1930s to the current day.

Trávníček 1941 gives a picture of the state of affairs in interwar Czech. He describes these three classes exclusively with their SC endings (1941: 96, 99, 113, 118), although he makes an exception for the forms *lžu*, *stůňu* ‘I lie, I am ill’ and *lžou*, *stůňou* ‘they lie, they are ill’. In a general note, he mentions “in folk and colloquial speech the forms *piju*, *píšu*, *vážu* ‘I drink, I write, I bind’ are often found, which not infrequently creep into the standard language. The rule, though, aside from those two forms [*lžu*, *stůňu* –NB] is to use the ending *-i*.” For the 3 pl. he is more categorical: “The forms *pijou*, *píšou*... ‘they drink, they write...’ are folk and colloquial” (90).

Trávníček 1951 gives a subtly different description. In his overview of conjugation, he recommends the use of *-u* after hard and ambiguous consonants and after *ň* and *ř* (*stůňu*, *ořu*). In contrast, “*-i* is used after *j*: *kryji*, *piji*, *žiji* ‘I cover, I drink, I live’; *-i* || rarely *-u* after *š*, *ž*, *č*: *píši*||*píšu*, *váži*||*vážu*, *pláči*||*pláču* ‘I write, I bind, I cry’. Only *lžu* ‘I lie’ has *-u* alone.” The 3 pl. is described as having the same distribution (565). This revision seems to permit and codify the use of the analogous forms alongside the fronted forms for class V verbs, but without assigning any register or stylistic differences. For the remaining verbs, no variation is admitted.

In places, then, this new description is more progressive in admitting a role for some analogous forms in the standard language, while elsewhere it retreats to a conservative position, refusing to acknowledge these forms for many other verbs.

Havránek and Jedlička's widely used *Česká mluvnice* recommends analogous forms in class V, with the following note:

Only verbs in -š, -ž, -č such as *píši, káži, dokáži, táži se, piší...* 'I write, I bid, I manage, I query, they write...' (alongside *píšu, kážu, dokážu, tážu se, pišou...*) can have -i in the 1 sg. and -í in the 3 pl. alongside -u, -ou; rarer are *češi, maži, váži, pláči, skáči, češí...* 'I comb, I spread, I bind, I cry, I jump, they comb...' alongside the more frequent *češu, mažu, vážu, pláču, skáču, češou...*

Verbs with other endings have only -u, -ou here, such as *ořu, pářu, stůňu, pošlu, kašlu, ořou...* 'I plow, I undo, I am ill, I shall send, I cough, they plow...'; at the same time some verbs in -š and -ž have only -u and -ou, such as *lžu, lžou* 'I lie, they lie', or at least as a rule, such as *hryžu, kloužu, klušu, koušu, křešu, kyšou, lížu, řežu, tešu, vržou, hryžou...* 'I gnaw, I glide, I trot, I bite, I strike, they pickle, I lick, I cut, I hew, they squeak, they gnaw...'

Since CC predominantly has -u, -ou (*mažu, mažou* 'I spread, they spread'), these endings are frequent with verbs whose meaning concerns daily life; thus we have consistently – even in SC – *klušu, -ou, koušu, -ou, kloužu, -ou, řežu, -ou*, 'I/they trot, I/they bite, I/they slip, I/they cut' etc. and regularly *pláču, -ou, mažu, -ou, vážu, -ou* 'I/they cry, I/they spread, I/they bind' and so forth, while only rarely, in bookish and archaic language, do we find *pláči, -í, maži, -í, váži, -í* etc. (1963: 274-275).

Classes I and VI are only given with fronted endings, although there is a note to the interested:

In CC these forms predominantly have the ending -u, -ou; thus it is making its way into the colloquial standard language for verbs whose meaning concerns daily life, or which are common in conversational phrases, such as *zuju se, obujou se, plejou, vylejou, děkuju, kupuju* 'I take

off my shoes, they put on their shoes, they weed, they pour out, (I) thank you, I buy' (1963: 285-286).⁵

The 1986 “Academy” grammar (*Mluvnice češtiny*, hereafter AMČ) further raises the profile of analogous variants. For class V verbs, the fronted forms *maži*, *maží* are termed *výrazně knižní a zastarávající* ‘markedly bookish and obsolescent’, while the analogous *mažu*, *mažou* forms are deemed neutral. For class I and VI verbs, the AMČ notes that “in these verbs as well, the endings *-u*, *-ou* are beginning to predominate in the 1 sg. and 3. pl...” (438-439). They later expand on this, saying:

...alongside the stylistically neutral flective morphs *-i* in the 1 sg. and *-í* in the 3 pl., the variant endings *-u* and *-ou*, which are evaluated as colloquial, are making inroads. Only bookish verbs and [...] verbs with an inherent sign of colloquiality or expressivity, in some cases those decidedly non-standard in character, have just one of the two endings, depending on which one fits with the lexeme’s overall stylistic colouration.

Note: In the contemporary language, the colloquial variants *-u*, *-ou* are coming to occupy a distinctly firm position [...] and especially for verbs that are frequently used and that indicate common, everyday life experiences, there is a tendency for them to become the neutral variants. With these verbs, the younger generation especially feels the original endings *-i*, *-í* to be nearly bookish. At the same time, the process of stylistic neutralization of colloquial variants is occurring more rapidly in the 1 sg. than in the 3 pl. of the non-past (458).

In the AMČ, the fronted class V *maži*, *maží* forms are marked “bookish”, while the analogous class I forms *kryju*, *kryjou* and the analogous class VI forms *kupuju*, *kupujou* are marked “colloquial”. Other forms (*mažu*, *mažou*, *kryji*, *kryjí*, *kupuji*, *kupují*) are left unmarked (430, 457).

Havránek and Jedlička’s shorter *Stručná mluvnice česká* (SMČ) was revised by the second, surviving author in the mid-1990s. Its account more or less agrees with

⁵ This explanation appeals explicitly to the notion of a “colloquial standard” Czech, which supposedly combined features of SC with non-standard features that were not markedly limited to individual spoken varieties. No stable colloquial standard with describable features has ever actually been shown to exist, although many scholars have posited its existence *a priori*.

the AMČ, giving preference to certain of the fronted forms while noting the gradual, lexically-based ascendance of the analogous ones (1998:127).

The 1995 Brno grammar (*Příruční mluvnice češtiny*, hereafter PMČ), is more cautious than the AMČ and SMČ. It remarks on class I verbs that:

In the 1 sg. and the 3 pl. they can have either the more common, colloquial endings *-u, -ou*: *kryju, oni kryjou*, or the stylistically higher *-i, -í*: *kryji, oni kryjí* (regional and substandard *kreju, oni krejou*). In colloquial Czech after soft consonants the endings *-i, -í* are thus giving way to *-u, -ou* (1995: 327).

For class V verbs, PMČ views the analogous forms as basic, noting that “some verbs of the *mazat* type keep the endings *-i, -í* in the 1 sg. and 3 pl. for higher style: *Táží se vás* ‘They are querying you’; *Oni píší* ‘They are writing’ etc.” (331). Some greater hesitation is shown over class VI verbs, where the authors remark:

For this type, the 1 sg. and 3 pl. non-past undergoes a shift in favor of the endings *-u, -ou* as opposed to *-i, -í*. [...] In higher style and for reasons of dissimilation (so as not to repeat *-u-* in two/three syllables in a row), the original ending is, however, retained: *kupuji//kupuju, kupují//kupujou; děkuji//děkujou, děkují//děkujou; studuji//studuju, studují//studujou; pracuji//pracujou, pracují//pracujou* (332).

We would be remiss if we did not include the most influential grammar of the modern age: the spell checker bundled with Microsoft Word. Although shipped without any explanations, its cheery red underlining yields quick and definitive answers as to whether we have spelled words correctly. My version, which goes with Word 2002, essentially follows the recommendations found in Trávníček’s works, with a slightly more liberal interpretation of class V verbs. For class I and VI, it does not admit the analogous forms, while for class V it admits both forms, even where some are fairly archaic. Thus it permits

píši, píší, dokáži, dokáží and *píšu, píšou, dokážu, dokážou*;

maži, maží, češi, češí and *mažu, mažou, češu, češou*;

hryži, hryží, kouši, kouší and *hryžu, hryžou, koušu, koušou*;

kryji, kryjí, piji, pijí but not *kryju, kryjou, piju, pijou*;

kupuji, kupují, studuji, studují but not *kupuju, kupujou, studuju, studujou*.

Against a background of gradual liberalization in scholarly and popular grammars, the Microsoft spell checker is a throwback to an earlier age, and one that could lead to a recrudescence of forms labelled by grammarians as bookish or stilted.

These forms have not come in for much attention in the scholarly literature, although they are often treated in passing in works on language variation in Czech. Čermák 1987, for example, develops a labelling scheme for acceptability of CC variants, and assigns *biju*, *bijou*, *kupuju*, *kupujou* to class A/a, meaning they are “accepted (and used) currently as normal” in written texts, and “occurring always or in most instances” in the spoken code (1987:142). Čermák mentions elsewhere (1987:140) that these spoken elements occur primarily in certain text types – informal correspondence and fiction – and are largely absent from others. Based on the fact that he assigns many non-standard features to this A/a category, it can be presumed that he was not extending this view of “accepted and currently used” across the board to SC texts of all types, but meant it to apply to certain registers and/or text types.

What then can we conclude about the status of these forms in SC from the literature? There seems to be potential for differentiation:

- by *verb class* (class V is most susceptible to analogous forms; class I least so);
- by *person* (1 sg. is more susceptible to analogous endings; 3 pl. less so);
- by *lexical register* (lower ones more susceptible to analogous endings; higher ones less so);
- by *communicative situation* or *textual register* (highly formal ones are less susceptible to analogous endings).

However, there is no independent evidence to show how these features should be ranked.

5. The Czech National Corpus

The CNC currently consists of a series of linked corpora, the “heart” of which is the 100-million word SYN2000 corpus. Described in detail in Kučera 2002 and elsewhere in this volume, SYN2000 provides researchers with a convenient way to gather data on usage patterns of contemporary written Czech. Its *proportional representativity* means that the formula for text inclusion is based on empirical research into what the Czech populace reads, in what amounts. The SYN2000 corpus thus yields a picture of the sort of language that the typical reader is exposed to.

The approximate breakdown of SYN2000 is: 60% contemporary press, 25% contemporary specialist literature drawn from a variety of fields, and 15% twentieth-century fiction (Kučera 2002:247-248). Source information is available for every citation, and for every search we can learn the approximate “distribution” of forms across the corpus, which is given in the form of a bar graph. The bar graph shows fiction first, followed by technical and journalistic texts.

SYN2000 does not provide data about the spoken language, nor does it provide information on unpublished sources (personal correspondence, essays, handouts, etc.). All its sources have in theory been subject to some level of editorial revision. It would thus be unlikely to yield, for example, copious examples of the instrumental plural in *-ma*, which is perceived as a feature of spoken Czech and would only occur in printed texts deliberately stylized after spoken language, such as dialogue in fiction.

Features like our verb forms, however, make a fruitful field for study with SYN2000. The acceptability of these variants is not clear; their status differs from handbook to handbook, and although there is clearly some stylistic differentiation between them, no one is quite sure what it consists of. Searching SYN2000 should yield information about how Czechs see these verbs in the printed world around them.

6. Approach

The current study required an overview of the scope of variation, as well as detailed individual pictures of variation. I thus conducted “horizontal” searches to determine how frequently these forms occurred in the corpus, as well as “vertical” searches to “drill down” into variation in individual forms.

In the vertical searches, I queried the SYN2000 corpus for four relevant forms of the following verbs: *absolvovat* ‘graduate/complete’, *archivovat* ‘archive’, *citovat* ‘cite’, *česat* ‘comb’, *děkovat* ‘thank’, *dokázat* ‘prove’, *charakterizovat* ‘characterize’, *koncertovat* ‘play (a) concert(s)’, *kousat* ‘bite’, *kupovat* ‘buy’, *milovat* ‘love’, *nakupovat* ‘shop’, *namazat* ‘spread’, *napsat* ‘write’, *naservírovat* ‘serve up’, *organizovat* ‘organize’, *reprezentovat* ‘represent’, *vypít* ‘drink up’, *výt* ‘howl’, *vytunelovat* ‘asset-strip’, *zabít* ‘kill’, *zakrýt* ‘cover’, *zamilovat se* ‘fall in love’.

These verbs provide a sampling of items in the lexicon. Class I and V verbs are frequently prefixed, so a selection of prefixed and unprefixed forms were checked against the entire set of verbs derived from that stem (e.g. *-psát*, *-kázat*). Class VI verbs were chosen to evaluate claims about their register determining their usage:

nakupovat, *kupovat*, *milovat*, *zamilovat se* come from daily language; *archivovat*, *absolvovat*, *reprezentovat*, *charakterizovat*, *organizovat* come from scholarly language; and *naservírovat*, *koncertovat*, *vytunelovat* come from colloquial language that would be likely to be found in a newspaper of the mid- to late 1990s. *Děkovat* ‘thank’ is a special case, as the forms *děkuji/děkuju* ‘I thank you’ is the most common way of expressing gratitude.

7. Scope of variation

Each word in SYN2000 is automatically assigned a sixteen-place tag that gives nearly complete morphological and partial syntactic information, and is also assigned to a *lemma*, or headword. Searching by tag and lemma, both full and partial, is possible. The “horizontal” searches took the form

`[(word="*uju") & (tag="VB.S...1.*") & (lemma="*ovat")]`

This query looks for forms with three features. They end in the combination *-uju*, are tagged as 1 sg. non-past verb forms, and are associated with lemmas that end in *-ovat*. With minor modifications to the first two criteria, this same search finds 3 pl. forms or fronted forms, and with minor modifications to the first and third criteria, it isolates forms from other verb classes.

Table 2. Scope of forms in SYN2000

Class	Form	Tokens (%)	Distinct forms (tokens/form)	Distribution in SYN2000
I	all	11,571 (100%)	376 (30.8 t/f)	
	-ji	906 (7.8%)	69 (13.1 t/f)	Uneven, with some small, sharp peaks ¼ way through.
	-ju	1634 (14.1%)	92 (17.8 t/f)	Uneven, with sharp peaks in first ¼.
	-jí	8678 (75.0%)	155 (60.0 t/f)	Uneven, with a few sharp peaks one-third of the way through.
	-jou	353 (3.1%)	60 (5.9 t/f)	Uneven, with large peaks in first ¼, then occasional small peaks.
V	all	9152 (100%)	324 (28.2 t/f)	
	-i	460 (5.0%)	42 (11.0 t/f)	Slightly uneven, with a few spikes in first 1/3.

-u	2629 (28.7%)	104 (25.3 t/f)	Uneven, with numerous spikes in first ¼, low frequency thereafter.
-í	3584 (39.2%)	69 (51.9 t/f)	Slightly uneven, with little in the first ¼ and even spikes elsewhere.
-ou	2479 (27.1%)	109 (22.7 t/f)	Balanced, with even spikes scattered throughout.
VI	all	227,341 (100%)	6976 (32.6 t/f)
-uji	26,307 (11.6%)	1,515 (17.4 t/f)	Uneven: high in beginning, sharp fall-off, gradually decreasing to low.
-uju	8,232 (3.6%)	1,018 (8.1 t/f)	Uneven: high in beginning, isolated smaller peaks thereafter.
-ují	191,913 (84.4%)	4,005 (47.9 t/f)	Uneven: first low; then peaks, and regular, even distribution thereafter.
-ujou	889 (0.4%)	438 (2.0 t/f)	Very uneven: high peaks in beginning, isolated peaks thereafter.

The overwhelming majority of tokens with these endings are found in class VI verbs (227,341 out of 248,064, or 91.6%.) Despite the far larger number of tokens in class VI, the number of tokens per distinct verb form is, on average, similar to those for classes I and V.⁶ Of a total of 7676 distinct forms, 6976, or 91.8%, come from class VI verbs.

The 3 pl. forms found in CC are significantly more frequent in SYN2000 than those from other dialects. Class I has 353 forms in *-jou* vs. 21 in *-jú* and 2 in *-jó*. Class V has 2479 forms in *-ou* vs. 5 in *-ú* and 3 in *-ó*. Class VI has 889 forms in *-ujou* vs. 48 in *-ujú* and none in *-ujó*. Many of the forms in *-ú* are citations from Slovak, where this is the standard 3 pl. ending. Those forms in *-ú* and *-ó* representing Czech dialects occur in highly stylized texts, where a large percentage of the forms are non-standard in shape. This is not, for the most part, true of the forms in *-ou*; they may occur in direct speech, but the surrounding forms are by and large standard in shape.

The “distribution” graph provides a rough idea of whether the tokens are spread evenly through the corpus or concentrated in one part of it (i.e. in one or more

⁶ Lists of forms generated by ticking the “remove duplicates” box during a general sort will contain one entry representing all tokens of e.g. *Považuji* and *považuji* and a second entry for the negated form, e.g. *nepovažuji* and *Nepovažuji*. If this function is used consistently, the patterns discerned should remain valid.

particular types of texts). None of our features appear exclusively in one portion of the corpus, but there are noticeable bumps and lumps in the distribution. These asymmetries can result from what I will call a *functional co-occurrence* or from a *stylistic co-occurrence*.

As an example of *functional co-occurrence*, we can take the verbs *dělat* ‘do/make’ and *mít* ‘have’, which show no variation in the 1 sg. and 3 pl. The forms *dělám* ‘I do/make’ (1603) and *mám* ‘I have’ (20,834) are distributed differently from the forms *dělají* ‘they do/make’ (3715) and *mají* ‘they have’ (64,287), with a ratio of 1 sg. forms to 3 pl. forms that is respectively 3:7 and 1:3. The 1 sg. forms have an uneven distribution, with noticeably more forms in the first quarter of the corpus than elsewhere. The 3 pl. forms show slightly lower frequency in the first quarter of the corpus, but are by and large evenly distributed through the corpus. It is not surprising that 1 sg. forms are more common in some text types than others: in fiction, for example, as opposed to journalistic texts and technical literature. The overall higher frequency of 3 pl. forms is also predictable, given the greater proportion of explanatory and descriptive texts in the corpus.

Stylistic co-occurrence happens when a particular form is more appropriate to a given register. For example, we would expect to find more examples of non-standard forms in texts that contain direct quotes from the spoken language or stylizations of speech, so that the *-u* and *-ou* endings should be more common in literary and journalistic texts than in technical literature. Analogous forms are indeed more common in the first quarter of the corpus than fronted forms, suggesting that there is more going on here than a simple functional co-occurrence.

Three trends discerned here coincide with the questions raised in section 4:

(1) There are significant differences between the frequencies of 1 sg. and 3 pl. forms. This reflects a difference between the spoken language and published written texts, which include less reference to the first person.

(2) Analogous forms overall are significantly less popular than fronted forms in the written Czech of the SYN2000 corpus.

(3) Class I and VI verbs are distributed differently from class V verbs, with analogous endings enjoying noticeably higher popularity in the latter class. Despite this, the fronted 3 pl. form is frequent in this class for some verbs (as per trend 1).

8. Variation in individual verbs

The picture within individual verbs supports the conclusions drawn above, with significant variation explainable through points signalled in the grammars. Searches employed either simple queries like

[Nn]akupuji|[Nn]enakupuji

.*[Kk]upujou

which retrieve respectively all tokens of *nakupuji* ‘I shop’, both affirmative and negative, and all forms ending in *-kupujou*, including *kupujou*, *nakupujou*, *vykupujou*, etc. Sometimes more complex queries were needed, such as

[(word=".*i") & (lemma=".*pít") & (tag="VB.*")]

which retrieves all forms ending in *-i* from the group of lemmas ending in *-pít* that are non-past forms (i.e. not past participles)⁷.

All class I verbs tested support the generalization that analogous forms predominate in the 1 sg., while fronted forms predominate in the 3 pl.:

Table 3. Class I verbs

	1 sg. <i>-u</i>	1 sg. <i>-i</i>	3 pl. <i>-ou</i>	3 pl. <i>-í</i>
<i>zakrýt</i> ‘cover (up)’	5 (100%)	0	1 (3%)	37 (97%)
<i>-krýt (-křejt)</i> ‘cover’	22 (55%) ⁸	18 ⁹ (45%)	5 (1%)	538 (99%)
<i>vypít</i> ‘drink up’	31 (82%)	7 (18%)	4 (6%)	58 (94%)
<i>-pít</i> ‘drink’	319 (78%)	92 (22%)	58 (9%)	588 (91%)
<i>zabít</i> ‘kill’	197 (95%)	11 (5%)	56 (25%)	164 (75%)
<i>-bít</i> ‘hit’	258 (93%)	19 (7%)	94 (16%)	485 (84%)
<i>výt</i> ‘howl’	6 (86%)	1 (14%)	9 (100%)	0
<i>-výt</i> ‘howl’	7 (87%)	1 (13%)	9 (89%)	1 (11%)

Vypít and *zabít* are frequently used in daily language (the form *zabiju* ‘I’ll kill’ is often used expressively), while *výt* and *zakrýt* appear in higher style texts.

In class V verbs, the pattern is more complex:

⁷ For the most part, these searches or variations on them sufficed to isolate the needed forms. Occasionally, manual sorting of results proved to be a necessary or expedient measure.

⁸ Includes 3 examples of *-křejtu*, with a CC alternation in the root.

⁹ Of which 8 examples are from a single sonnet.

Table 4. Class V verbs

	1. sg. -u	1. sg. -i	3. pl. -ou	3. pl. -í
<i>napsat</i> ‘write’	424 (85%)	73 (15%)	51 (33%)	103 (67%)
<i>-psát</i> ‘write’	1732 (73%)	631 (27%)	373 (19%)	1547 (81%)
<i>dokázat</i> ‘prove’	1205 (87%)	176 (13%)	1120 (32%)	2337 (68%)
<i>-kázat</i> ‘preach’	1601 (87%)	236 (13%)	1370 (32%)	2877 (68%)
<i>namazat</i> ‘spread’	13 (100%)	0	2 (100%)	0
<i>-mazat</i> ‘spread’	42 (100%)	0	60 (98%)	1 (2%)
<i>česat</i> ‘comb’	14 (100%)	0	14 (100%)	0
<i>-česat</i> ‘comb’	19 (100%)	0	19 (100%)	0
<i>kousat</i> ‘bite’	14 (93%)	1 (7%)	31 (100%)	0
<i>-kousat</i> ‘bite’	21 (95%)	1 (5%)	40 (100%)	0

There are two distinct groups here: for verbs in *-kázat* and *-psát*, the old fronted forms are very much alive, and predominate in the 3 pl. form. For those in *-česat*, *-kousat*, and *-mazat*, the old fronted forms are barely represented. The nature of these actions (literary/cultural vs. physical) may contribute to this difference.

Stylistic differences come to bear in considering class VI verbs:

Table 5. Class VI verbs

	1. sg. -u	1. sg. -i	3. pl. -ou	3. pl. -í
<i>děkovat</i> ‘thank’	945 (34%)	2104 (66%)	1 (0%)*	370 (100%)*
<i>-děkovat</i> ‘thank’	953 (31%)	2120 (69%)	2 (0%)*	386 (100%)*
<i>nakupovat</i> ‘shop’	11 (23%)	36 (77%)	2 (0%)*	499 (100%)*
<i>kupovat</i> ‘buy’	64 (36%)	112 (64%)	13 (2%)	661 (98%)
<i>-kupovat</i> ‘buy’	77 (33%)	159 (67%)	15 (1%)	1356 (99%)
<i>zamilovat se</i> ‘fall in love’	15 (65%)	8 (35%)	3 (4%)	65 (96%)
<i>milovat</i> ‘love’	814 (48%)	892 (52%)	29 (5%)	602 (95%)
<i>-milovat</i> ‘love’	830 (48%)	892 (52%)	32 (5%)	677 (95%)
<i>vytunelovat</i> ‘asset-strip’	0	1 (100%)	0	4 (100%)
<i>naservírovat</i> ‘serve up’	3 (100%)	0	0	6 (100%)
<i>-servírovat</i> ‘serve’	5 (50%)	5 (50%)	1 (2%)	62 (98%)

<i>koncertovat</i> ‘perform’	1 (17%)	5 (83%)	0	39 (100%)
<i>charakterizovat</i> ‘characterize’	0	4 (100%)	0	340 (100%)
<i>archivovat</i> ‘archive’	3 ¹⁰ (75%)	1 (25%)	0	16 (100%)
<i>absolvovat</i> ‘graduate, complete’	4 (14%)	24 (86%)	0	409 (100%)
<i>citovat</i> ‘cite’	10 (3%)	276 (97%)	0	89 (100%)
<i>organizovat</i> ‘organize’	4 (21%)	15 (79%)	2 (0%)*	334 (100%)*
<i>-organizovat</i> ‘organize’	8 (31%)	18 (69%)	3 (1%)	373 (99%)
<i>reprezentovat</i> ‘represent’	3 (20%)	12 (80%)	1 (0%)*	215 (100%)*

* = attested tokens in -ou are so few as to round to 0%.

For stylistically neutral verbs representing everyday actions, there is a preference for the fronted form in the 1 sg., but the analogous form is well-represented. The verb *děkovat*, whose use is said to be idiosyncratic because of the pragmatic functions discussed above, conforms nonetheless to the pattern for this group. More colloquial or expressive words likewise show a certain percentage of 1 sg. forms in *-u*. Verbs from a higher stylistic register have far fewer 1 sg. forms in *-u* (with the exception of *organizovat* ‘organize’, apparently forming a bridge between the two categories). Few verbs show any significant usage of the 3 pl. form in *-ou*.

9. Conclusions

In the current analysis, I have, for reasons of space, confined myself to data obtained by automatic processing of results. Use of the relatively crude contextual “meter” offered by the distribution graphs nonetheless confirmed that there are clear differences in the usage of variants across text types that exceed simple functional co-occurrence. A more finely-grained picture will have to await a more detailed look at individual examples, and must consider the contexts in which these examples appear. We can nonetheless draw several clear conclusions.

Many a grammar has pointed to the shift underway from fronted to analogous endings, and these observations are largely borne out. The shift is already complete in the written usage of many class V verbs, and there is evidence of variation in class I. Many class VI verbs remain unaffected by this shift, although the most numerically frequent of them already show evidence of it. Change seems to be proceeding lexeme by lexeme through the system; how far it will go remains to be seen. Čermák’s high

¹⁰ All in one article.

acceptability rating for the analogous endings is not consistently reflected yet in usage, at least for certain verb classes.

The stylistic value and usage pattern of lexemes contributes to their susceptibility to analogous endings, as noted in most of the recent grammars. Expressivity and pragmatic functions in the spoken language also play a role, with e.g. *miluju* ‘I love’ and *děkuju* ‘I thank’ better represented than other analogous forms.

Only Trávníček 1941 and the AMČ point to a difference between the 3 pl. and the 1 sg., one which is regularly reflected in these data. In certain verb classes, the differences are pronounced, with 1 sg. analogous endings appearing 30 to 70 percent of the time, while 3 pl. analogous endings are nearly absent. Regardless of the frequency of both forms in the spoken language, only the former seem to be making substantial headway in the written language across the majority of verb classes. This confirms the results of Bermel 2000 (55-61, 103-108), which found a substantial difference in the frequency of these forms in literary dialogue.

The most misleading prescription turns out to be the one implicit in current spell-checkers, which disallows a large number of analogous forms in common written usage (*děkuju, miluju*), while prescribing some fronted forms that are relatively infrequent (*zabiji, vypiji*).

To sum up, the *-u* forms appear to be fully *standardized*: they appear frequently in texts described as standard written Czech and compete with the traditional *-i* ending. There is evidence that their prosperity is limited, so far, to certain verb classes and certain classes of lexical items, which indicates that this ending for the most part still has a particular stylistic “flavour” to it, and further research will probably show them to occur in certain textual situations with greater frequency. The *-ou* forms, on the other hand, are at best still *standardizing*. Outside class V, they are infrequent and tend to be found in dialogue or direct citation. Many verbs show no examples of them at all in SYN2000. Despite the common heritage of these two forms and their apparent equivalence for many Czech grammarians, corpus-based research shows deep differences in their usage in the contemporary written language.

References

Bermel, Neil. 2000. *Register variation and language standards in Czech*. Munich: Lincom Europa.

Čermák, František. 1987. Relations of spoken and written Czech. *Wiener slawistischer Almanach* 20: 133-150.

---. 1993. Spoken Czech. In: *Varieties of Czech*. Eckert, Eva, ed., 27-41. Atlanta: Rodopi.

---. 1995. Prague School of Linguistics today. *Linguistica Pragensia* 5: 1-15.

---. 1997. Obecná čeština: je součástí české diglosie? *Jazykovědné aktuality* 34: 34-43.

Čermák, František and Sgall, Petr. 1997. Výzkum mluvené češtiny: jeho situace a potřeby. *Slovo a slovesnost* 58: 15-26.

Český národní korpus - SYN2000. Praha: Ústav Českého národního korpusu FF UK. <<http://ucnk.ff.cuni.cz>>.

Hammer, Louise. 1985. *Prague Colloquial Czech: A case study in code switching*. Indiana University: unpublished PhD dissertation.

Hammer, Louise. 1993. The function of code switching in Prague Colloquial Czech. In: *Varieties of Czech*. Eckert, Eva, ed., 63-78. Atlanta: Rodopi.

Havránek, Bohuslav and Jedlička, Alois. 1963. *Česká mluvnice*. Praha: Státní pedagogické nakladatelství.

Havránek, Bohuslav and Jedlička, Alois. 1998. *Stručná mluvnice česká*. 26. vydání. Praha: Fortuna.

Kravčičinová, K. and Bednářová, B. 1968. Z výzkumu běžné mluvené češtiny. *Slavica Pragensia* 10: 305-320.

Kučera, Karel. 2002. The Czech National Corpus: Principles, design, and results. *Literary and linguistic computing* 17 (2): 245-257.

Mluvnice češtiny (= AMČ). 1985. Díl 2: Tvarosloví. Praha: Academia.

Příruční mluvnice češtiny (= PMČ). 1995. Praha: Nakladatelství Lidové noviny.

Trávníček, František. 1941. *Stručná mluvnice česká*. Praha: František Borový.

Trávníček, František. 1951. *Mluvnice spisovné češtiny. Část I: Hláskosloví – Tvoření slov – Tvarosloví*. Praha: Slovanské nakladatelství.