Czech Language News

Fall 2000 | North American Association of Teachers of Czech | Number Fifteen
http://www.language.brown.edu/NAATC/index.html | ISSN 1085-2950

Message from the President

Dear Members and Friends of NAATC,

This year's plenary meeting will once again be held in conjunction with the American Association of Teachers of Slavic and East European Languages (AATSEEL), convening December 27-30 in Washington, D.C. On December 29 at 1:00 Malynne Sternstein chairs our panel, a new contribution to the theme "Czech Literature in Transition" which we have sponsored for three years now. There will be two papers on Topol's *Sestra*, to be given by Karin Beck of Columbia University and Yvonne Howell of the University of Richmond, one on Kundera presented by Anna L. Kleshelskaya of Dartmouth College, and one on a Czechoslovak-Soviet topic, presented by David Lightfoot of the University of Toronto. For information on AATSEEL, see its site at http://clover.slavic.pitt.edu/~djb/aatseel.html#program.

We look forward to seeing you there, especially as our NAATC organizational business meeting will directly follow the panel at 3:00. On the agenda stands the election of new officers. The business meeting is also an excellent opportunity for members to submit 2001 dues!

If you are not a subscriber to the Ohio State Czech list managed by Jeff Holdeman, consider sending in your address. Material shared and questions asked (on problems in translation, Czech language and linguistics, folklore, literature) are amusing and interesting—a delightful contrast to the barrage of bureaucratic email we are obliged to receive. Write to listproc@lists.acs.ohio-state.edu, saying SUBSCRIBE CZECH - (your email address).

A short note on the health of the Czech language in the South: The presence in American universities of Czech and Slovak student-athletes, many of them excellent students as well as tennis players, enhances our presence in European Studies consortia where any edge, however tiny, may tip the balance toward giving us release time and staffing to teach Czech language. At Tulane we are working for the creation of a faculty slot in cross-disciplinary colloquia in Central European social institutions. If we should acquire this extra person, it could enable us to offer a full year of beginning Czech each and every year. The indirect path to legitimacy requires patience and ingenuity, but the results can make it all worthwhile.

We could perhaps find inspiration in the song by Václav Poustka in *Dobytí severního pólu* (Smoljak/Svěrák):

Tam, kde hynou vlci
tam, kde hynou sobi
Čech se přizpůsobí
Čech se přizpůsobí

George Cummins, Tulane University

Contents

Developing Teaching Materials with the Czech National Corpus[1]

Neil Bermel,
University of Sheffield

It takes an impressive linear footage of reference works to get me through my days as a Czech teacher. First comes a series of dictionaries, bilingual (English-Czech, Czech-English) and monolingual, plus a couple of grammars, my current favorite being the Masaryk University *Příruční mluvnice češtiny*. I also use dictionaries of phrases and proverbs, orthography manuals, advanced language textbooks and *skripta*…. And like most teachers, I also have a group of good-natured native speakers to hand, both locally and by e-mail, who answer my endless stream of questions.

Even with all these books and consultants, it's never quite enough. Who among us, native speakers included, has not questioned his or her judgment, doubted the accuracy of a dictionary based on conservative 1950s usage, wondered if a reference grammar was telling the whole story? Those of us who are not native speakers know what it feels like to search in vain for the case or preposition governed by a particular verb or noun, to write a handout or correct a student assignment wondering if, even once we've gotten the grammar "right," whether a Czech would really say it that way.

This article discusses one computerized reference work that can offer us some assistance: the Czech National Corpus (*Český národní korpus*, hereafter ČNK). It has the potential to keep us better informed about developments in the language, and to help us solve the occasional linguistic puzzle from the comfort of our own computers.

[T]he ČNK supplies the sort of information no grammar ever will. It shows how the language is actually being used "on the ground" by writers, journalists, scientists, politicians, and ordinary

The Czech National Corpus

A computerized corpus, for those not familiar with the term, is a substantial body of text in electronic format, which can be searched and sorted to yield additional information about its content. The ČNK is the first large-scale corpus project of its sort for Czech, and is the only one that really rivals the corpora developed for English and other languages in size and depth.[2] It currently contains well over 100 million words of contemporary Czech and is still growing. Separate corpora within it contain large samples of journalistic language, fiction, scientific prose, spoken Czech, and language from a variety of historical periods.

Basic information about the ČNK is at http://ucnk.ff.cuni.cz. The ČNK is the *raison d'être* of the Czech National Corpus Institute (*Ústav českého národního korpusu* or ÚČNK) affiliated with the Philosophical Faculty at Charles University, and is directed by Professor František Čermák. It came into existence in the years just after the Revolution, and its remit is to create and maintain a series of linked text corpora that give a picture of contemporary Czech usage, and to develop programs that provide access to these corpora. A good source of information on the history and development of the ČNK is Michal Šulc's short monograph *Korpusová lingvistika* (Karolinum, 1999).

For teachers like me, who speak what is hopefully a passable version of "foreigner Czech", but who are constantly writing handouts and exercises that demand a steady supply of examples of good usage, the ČNK supplies the sort of information no grammar ever will. It shows how the language is actually being used "on the ground" by writers, journalists, scientists, politicians, and ordinary citizens.

The ČNK is available in two formats. The first is web-based, offering limited access to some 20 million words, while the second (the full 100 million words) makes somewhat greater technical demands. In the latter case, use of the Corpus implies consent to the ČNK's license, which states that you cannot use information gleaned from the Corpus for profit without the express permission of the Institute, and that you will credit the ČNK in any materials developed from it.

Querying the Corpus

To use the ČNK you need to master a very limited sort of computer language. It consists, for all intents and purposes, of five or six symbols that help you include and exclude certain combinations and alternatives—letting you search for a word while saying, for example, "spell it with either an *s* or a *z*", "give me any case with any ending", or even "give me examples that don't have this ending." This sort of flexibility is crucial for searching in a heavily inflected language like Czech. Among the symbols are question marks (?) for optional characters, periods (.) for "wild card" characters, asterisks (*) for strings of more than one wild card, square brackets ([ ]) for lists of alternative characters and a circumflex (^) for excluded characters.

The site pages say that your computer must be able to work with the ISO Latin 2 coding (8859-2), but for those with newer computers this warning may be needlessly off-putting. I've found that if your computer handles Czech characters in most programs without any difficulty, you should simply be able to switch keyboards and type in the field. If you see Czech characters when you do so, you should have no trouble getting a correct answer from the Corpus. I've tried the Web version on both a Macintosh running OS9 and a PC running Windows 98, and have used both Netscape Navigator 4 and Internet Explorer 5. In neither case was any special set-up required, nor did I have to meddle with the pre-set encodings.

Even if you can't type in Czech and don't feel like solving the problem just now, you can still use the Corpus without it. For many queries, you can simply substitute a period for the accented character before entering it. For example, if you cannot type

čarodějnice

---

[1] I am deeply grateful to Professor František Čermák and Mgr. Michal Šulc of the ÚČNK for their comments and suggestions.

[2] In many ways, of course, users of the ČNK are far more fortunate than users of these older English corpora. Many of the English ones are twenty or thirty years old, and today's user is consequently stuck with the limitations of computer technology of that time.

you could try

.arod.jnice

which will yield the same or a very similar result.

In its response, the corpus processor will supply examples of the word requested, plus a certain amount of context to the left and right of each example. You can set this context to be small (20 characters) or relatively extensive (60 characters). I usually request 40 to 50 characters of context; I find that's the right amount to get an idea what the text is about without making each citation too wide for the screen.

Public Web-based access to the Corpus

As a public service, the ÚČNK provides web-based access to one corpus, called PUBLIC, consisting of about 20 million words drawn from a subset of the full 100-million word corpus of written Czech. From the main page of the corpus, you click on the link Veřejný přístup and then the link WWW rozhraní k ČNK. This brings you to a page with a field at the top, where you type in the word you are looking for with any modifying symbols; this is called your "query." When you submit your query by pressing the "Hledej" button, the computer returns the first twenty entries in the PUBLIC corpus that match it. There is no limit to the number of requests you can make, although you will only ever get the first twenty items it matches. At the moment, if you perform the search several times, the processor will always throw up the same twenty examples, but according to Professor Čermák, there are plans to remove this restriction, so that you get a different 20 examples each time you search. To set the amount of context, you type a number between 20 and 60 in the box labelled "v rozsahu _____ znaků".

[T]he whole process took under fifteen minutes—surely less time than it would have taken me to locate lists of words, look them up in enough dictionaries to collect examples, and pester my long-suffering native speaker friends to fill in the gaps.

To test the web-based interface, I chose a couple of my favorite bugbears: places where grammars and dictionaries usually fail me by being too succinct, too dogmatic and conservative, or simply out of date. Examples are the range of prepositions and cases used with a particular verb, or what happens with certain neologisms.

Tracking the locative plural

I decided first of all to look at the masculine and neuter locative plural of nouns, where there is a well-known alternation between forms in *-ích* and *-ách*. I have always found this a difficult case to get a handle on, as the words involved aren't among the most frequent words in the language, and don't crop up all that often in this particular form. When giving examples of this sort to students in explanations or exercises, it can be important to get the right register, so I decided I would prefer to have examples from texts.

Following the example given on the instructions page (http://ucnk.ff.cuni.cz/navod.html) I typed in

.+íčkách

to find out what happens with words in *-íček* in the locative plural. The period means "any letter"; the plus sign, "repeat that as many times as necessary." In other words, my request said, "give me words with any number and kind of characters at the beginning and these specific characters at the end." The reply was:

Požádali jste o vyhledání slova ".+íčkách" a bylo nalezeno 117 výskytů. Výpis konkordancí - vypíše se pouze 20 výskytů:

I then sifted through the twenty examples it had given me, eliminating 12 feminine forms like *svíčkách* and *příčkách,* and was left with *víčkách, vozíčkách, slovíčkách, měsíčkách, klubíčkách* and *žebříčkách*.

I then went back and asked for

.+íčcích

and was given the first 20 of 77 finds. Here were *balíčcích, vozíčcích, klíčcích, žebříčcích, koníčcích, pytlíčcích* and even *kamarádíčcích*. There was one overlap with the previous example, meaning that I now had examples with both *žebříčkách* and *žebříčcích*.

I then repeated the search on *.+ečkách* and *.+ečcích*, with respectively 119 and 38 finds, including forms like *šatečkách, kolečkách, hrobečkách, svazečcích, válečcích, rámečcích, oblečcích, čepečcích, stařečcích, konečcích*, and *koberečcích*.

Forging ahead, I tried searching on *.+ákách and .+íkách,* but here I ran into difficulties. Aside from the fact that *-ák* is found largely with words common in spoken Prague Czech, I found that when I looked for the more literary endings in *-ácích* and *-ících,* the answers from the Corpus were swamped with forms of adjectives and participles, making the results useless. The examples I gathered from here nonetheless included *teplákách, manšestrákách, gumákách, barákách, dřevákách, modrákách, truhlíkách, vozíkách,* and a number of place names.

The search was therefore not an unqualified success, but it did give me a lot of valuable information. I had lots of examples of both types of locative plural usage, and the whole process took under fifteen minutes—surely less time than it would have taken me to locate lists of words, look them up in enough dictionaries to collect examples, and pester my long-suffering native speaker friends to fill in the gaps.[3]

I was also able to get a sense of the sort of contexts where the *-ách* ending appeared. With masculine nouns, it did tend to occur in less formal texts and when someone was being quoted speaking informally, although—as we would expect—with neuter nouns in *-ečko* and *-íčko* it was universal. The more formal *-ích* ending did occur with masculine nouns in some quotations, but was far more frequent in ordinary narrative prose. These statistics would not hold up in court (I'd need to use the full access version to increase the sample size and set the answers in context), but they would certainly help me to answer questions from students more confidently.

[3] With full access to the Corpus, a tag-based search would also have been successful here (see below).

Examples of government

The best translating dictionaries supply not only an equivalent of the word that you are looking for, but some indication of how it is used, i.e. what prepositions and/or cases it governs. It is much harder to find complete examples in dictionaries that illustrate usage. Looking up 'search', I found among other words *pátrat;* my current favorite dictionary, Fronek's *Anglicko-český slovník*, gave me *pátrat po kom/čem,* but with no examples of usage. This is understandable; Fronek is a one-volume dictionary, and space considerations preclude extensive examples. Looking in a big monolingual dictionary like the eight-volume *Slovník spisovného jazyka českého*, I did find examples, but mostly short ones that do not make illuminating reading, such as *pátrat po ztraceném dítěti, pátrat po tajemství, všechno pátrání bylo marné*.

Moving to the Corpus, I constructed a more complex query to retrieve all the forms of a given word. For the verb *pátrat*, the possible active forms are *pátrám, pátráš, pátrá, pátráme, pátráte, pátrají, pátral, pátrala, pátralo, pátraly, pátrali, pátrej, pátrejte, pátrejme*. If we want to pick up future tense forms, all we need is the infinitive *pátrat*. Arranging these forms in a chart according to letter position, we get:

1	2	3	4	5
P	Á	T	R	Á
P	Á	T	R	Á
P	Á	T	R	Á
P	Á	T	R	Á
P	Á	T	R	Á
P	Á	T	R	A
P	Á	T	R	A
P	Á	T	R	A
P	Á	T	R	A
P	Á	T	R	A
P	Á	T	R	A
P	Á	T	R	A
P	Á	T	R	E
P	Á	T	R	E
P	Á	T	R	E

For all forms the first four letters are the same. We can just type them in as

pátr

The fifth letter can be *á, a* or *e,* so we will show this by typing all three forms in square brackets to indicate alternatives:

pátr[áae]

The next character could be one of several—if in fact there is another character. I typed the possibilities in a second set of brackets and then put a question mark afterward, to show that this character might be optional:

pátr[aáe][štlj?]

I finished by adding a period and an asterisk, just in case there were any letters beyond the sixth place:

*pátr[aáe][štlj?].**

Another option would have been to just stop after the first set of brackets and type period-asterisk for "any combination of characters":

*pátr[aáe].**

This option, however, would have pulled in other related words and forms as well, such as *pátrán, pátrání, pátraný, pátravý, pátrač, pátračka, pátradlo, pátrací*, which is why I chose the route I did.

Unfortunately, the search did not come out exactly as I had hoped, as the results seem to have missed the form *pátrá*. This is an inevitable outcome when receiving only twenty forms out of possible thousands in the Corpus, and I probably exacerbated the problem because my search did not include capital P, thus eliminating any sentences beginning with *Pátrá*. I suspected that *pátrá/Pátrá* would be a common form, so I ran a separate search on it. In both instances I asked for the maximum context—60 characters on either side—to increase the possibility of getting full sentences.

There were respectively 316 and 142 examples. And the 20 that the web interface provided for each query were indeed more entertaining than those in my dictionary. There were some fairly basic sentences:

Ke zranění nedošlo, po muži se <pátrá> .

Policie zatím po zloději marně <pátrá>.

Začala jsem neobratně <pátrat> po důvodu peněz. [sic?]

Vadí mi, že ti dnešní mladí málo <pátrají>, co bylo dřív.

Policie ČR <pátrá> po zmizelém podnikateli i po zmizelých botách.

Nyní se <pátrá> po dalších pěti ženách, které v okolí Gloucesteru zmizely.

There were also some more complex and interesting sentences, which would be suitable for advanced learners and for homework assignments:

Policie , která po pachatelích <pátrá>, škodu odhadla na 200 tisíc korun.

Vtahuje mne to stejně jako každého, kdo <pátrá> po tajemství ztracených civilizací.

Záchranný vrtulník se vznášel nízko nad hladinou a <pátral> po mrtvém těle.

Ruští historici začali <pátrat> po ostatcích amerických vojáků padlých na ruském Dálném východě.

Německé úřady nyní oficiálně <pátrají> už jen po třech domnělých členech RAF…

One thing that comes across strongly in these examples is how closely this word is associated with official searches and investigations: police, army, government offices. And yet this very basic fact of usage and cultural context is not at all clear from the examples I found in print dictionaries. If I had used sentences based purely on the phrases in the dictionary and given them to a native speaker to check, I would have come out with "correct Czech" but would have lost this essential point.

Use of neologisms

One of the many places where print reference works often fall down is the provision of neologisms. What if, for instance, we want to know how the word "web" is used in Czech? It won't be found in any printed dictionary (yet); there are undoubtedly specialized computer glossaries on the web itself that will tell me what web gurus think, but I may not know where they are, and I may not want to take their word for it anyway. The ČNK is an obvious source of information.

I had already observed that in speech the WWW is often referred to by the word *web,* and was reasonably sure that this word is fully declinable in spoken Czech, i.e. *s webem, na webu,* etc. I was less sure whether this usage was found in written Czech to any extent. Then there was the matter of the full term *World Wide Web*. Does the longer term decline as well (*na world wide webu*)? Are either of them capitalized?

Since the web interface will only give me twenty examples, I decided not to cast a wide net and sift through it. Instead, I did several very restricted searches on the words *web, Web, webu, Webu, webem, Webem*.

Sure enough, the ČNK obliged me by supplying examples of five out of these six terms (absent was *webem*). All of these except *Web* (67 occurrences) occurred fewer than 20 times in the Corpus, so I was able to look at complete findings for these forms. I learned the following:

• W or w? Czechs use both small and large W's, although the capitals seem more common when the full name (*World Wide Web*) is used and small ones seem more common when the shortened form (*web*) is used.

• Inflected or non-inflected? The word *web* is fully inflectable, although it also appears as an indeclinable noun. Thus we find *prostřednictvím webu* but also *v prostředí World Wide Web,* as well as *agresivnímu bratříčkovi - world wide webu* and *na síť Web*. It seems that declinability is more common with *Web* and *web* than with *World Wide Web* and *world wide web*. Capitalization and declinability do not seem to be related.

• Integrated into the language or not? To a certain extent, the ability to decline and to appear uncapitalized is an indication of the word's increasing acceptability as an ordinary Czech lexeme. However, I observed a few places where it is surrounded by quotation marks or explained by a Czech paraphrase (*celosvětová pavučina*), testifying to its recent origin.

The web interface turned out to be a useful tool for answering casual questions quickly and with a reasonable degree of accuracy. However, if I were carrying out research on any of these words or problems, I would be well advised to use the full-access version instead, which can let me see a larger selection of examples and set small samples against the background of the whole.

The web interface turned out to be a useful tool for answering casual questions quickly and with a reasonable degree of accuracy. However, if I were carrying out research on any of these words or problems, I would be well advised to use the

Full access to the Corpus

The second level of access to the ČNK is more demanding, both academically and technically. Access over the Internet is granted for a set period of months to scholars who apply to the Institute with a project description. The user is provided with a password and user name, and can then download and install the software necessary to connect directly to the Institute's server. This level provides access to a much larger corpus, called SYN2000, removes the twenty-item restriction, and allows much more varied types of searching than with the web-based version. At the moment, the software only works on Windows machines, although Macintosh-compatible software is planned.

The program is called GCQP, for Graphical User Interface for the Corpus Query Processor. Like the web interface, you input queries consisting of full words or partial words; you can partially specify any missing characters or leave them completely unspecified. Unlike the web interface, GCQP lets you play with your results after you get them. You can remove unwanted examples, sort the results in various ways, and save them for future work. You also get certain statistics on what you have done, meaning that you can easily keep track of the overall context in which your findings are set. A further advantage is that you can see much longer stretches of context by selecting individual examples. The GCQP has a Czech and an English interface that you can easily switch between, and a detailed on-line manual is now provided via the Institute's web page.

Much of the GCQP's power derives from the fact that in the corpora themselves, every word is labelled and described. You cannot access any of this information through the web version's processor, but you can when using the GCQP. There are several types of labelling that are used.

First, all the forms of a particular word are associated with each other. This means that, theoretically, the database has the information that *prosbami* and *prosbu* are forms of the same word *prosba*. This "basic" form (the nominative singular, infinitive, etc.) is called the *lemma*, and the corpus is said to be *lemmatized* if it contains this information linking disparate forms of the same word.

Second, each form contains a marker that describes its various grammatical pigeonholes, including gender, number, case, tense, person and so forth. The information attached to each form is called its *tag*, hence the term *tagged corpus*.

With my approved account and the software installed on my computer, I logged onto the ÚČNK computer and tried a few more searches.

Sorting forms by spelling or lemma

When the GCQP returns its list of examples to you, it does not sort them in any way; it simply displays them in the order it found them in the corpus. This did not matter very much when we were dealing with only twenty examples, but when the numbers run into the hundreds, you want a way of organizing them. The query processor allows you to reshuffle or sort (*třídit*) your examples into different orders. There are two basic ways to sort an unwieldy list of

forms. One way is to sort them by spelling (left to right or right to left); a second is to sort them by lemma.

Going back to the examples above with *pátrat*, I could run this search in the GCQP. Here I combined two disparate searches into one using the vertical bar command ("either/or"), which seems not to work in the web version:

*pátrá|pátr[aáe].**

The query says, "show me all examples of *pátrá* and other forms beginning with *pátr-* and continuing." The response brought up 913 examples from the PUBLIC corpus, which I then sorted through.

First, I decided to get rid of extraneous forms that were not really part of the verb. Going to the *simple sort/ jednoduché třídění* command, I set the parameters for the sort. I wanted it to look at the words it found, not at what was around them, so I asked it to sort by what is called *KWIC odleva/from left*. (KWIC is, for some reason, the term for the target word.)

I then set the method for sorting, asking it to sort the KWIC by lemma—in other words, by the "head word" each form was classed under. This grouped together all the forms of the verb *pátrat*, and I then went through erasing all the others using the *smazání vybraných* command. The 447 examples remaining were all of the verb *pátrat*.

I then went through and sorted again on *KWIC odleva* using the *word/slovo* designation. This essentially alphabetized the list for me and put all of one particular kind of form together, letting me easily pick through for nice first-person examples, past-tense examples, etc.

I began to envisage a series of exercises built around the valences of common words, which could be used to expand students' vocabulary

Searching by lemma

An easier way to approach this same search would be simply to search on the lemma *pátrat* itself. To do this, you use the format

[lemma="pátrat"]

in the query field, instead of just typing in the word *pátrat* and hitting return. You could then proceed as before to sort the different terms from each other. In most instances the lemma is intuitive, being the nominative singular of a noun or adjective, and the infinitive of a verb.

Sorting by context

You can also sort by the context of the word. For instance, I thought it would be interesting to check the valence of the word *klíč,* which is often described as having two possible complements: *klíč od čeho (od domu, od bytu)* for actual, physical locks or *klíč k čemu (ke cvičení, k hádance)*for metaphorically locked objects. But one also frequently hears *klíč do čeho*. I ran a search on the word *klíč* in the larger SYN2000 corpus. Here I decided that nominative-accusative singular forms would probably be sufficient, so I just input the basic word. Sure enough, the query processor returned 2101 examples—far too many to simply leaf through.

I then asked it to sort my results by what appears to the right of them. I chose the *simple sort/jednoduché třídění* command, and asked it to sort by what is called *right context/pravý kontext*. This means that the program looks just to the right of the search word in each example and re-files all its examples them in that order. Since prepositional phrases usually follow the noun directly, this separated out all the examples of *klíč od čeho*, *klíč k čemu* and *klíč do čeho*. I asked it to sort up to four places to the right of the word, such that *k* and *ke* will be separated out.

Once my sort was done, I listed through a screen at a time using the *page down/o stránku dolů* command, skipping past examples of *klíč a, klíč byl, klíč visel* and so forth that did not interest me. It took a while for me to delete all of these, not least because every time I hit *delete selected rows/vyjmout vybrané řádky* the GCQP hopped back to the first screen of examples. In the end, I had 630 examples with the prepositions I had wanted, plus a few interesting additions that I will discuss later.

There were 47 examples with *klíč do čeho*. Most of them were false leads, as the *do* depended on the verb, not the noun (*dát klíč do schránky, když se mu podaří klíč do 15 minut vynést z kruhu, strčit klíč dokapsy)*. However, there were a few interesting and notable examples. I found *klíč do ráje, klíč do nejvyšší soutěže* and *přišel, jako kdyby znal klíč do všech zámků světa*. Clearly three examples out of 630 means this is a marginal construction, and to judge by the examples I found, in standard Czech it does not concern real locks, only metaphorical locks or passages. I would judge this construction as unnecessary baggage for students, then, regardless of how it is used in the spoken language.

With the preposition *k* I found further examples bringing me to line 451 of my search, meaning there were 404 such examples when I subtracted the 47 for *do*. Among them were *klíč k blahobytu, k bezpečnosti, k chování ODA, k ekonomickému růstu, k pochopení,k reformě,k úspěchu*. For *klíč od,* I found *klíč od bran města, od kostela, od hlavního vchodu, od trezoru, od pokoje, od propasti,* giving a further 120 examples.

The unanticipated bonus was 28 examples with *klíč pro co*, as in *klíč pro druhou komoru parlamentu, pro dělení dotace, pro místnost číslo padesát, pro obnovení růstu ekonomiky, pro pochopení židovského výkladu Bible*. There was one example with *proti: klíč proti chybnému připojení konektoru*. Interestingly enough, the ever-more common preposition *pro* seems to be creeping into places where perfectly good equivalents like *k* and *od* already exist in Czech.

What can a teacher do with examples like this? More advanced students can be asked to write sentences with these phrases, to complete sentences with *klíč* that you begin, or to react to statements using a phrase with *klíč* from a list. As I worked my way through this search, I began to envisage a series of exercises built around the valences of common words, which could be used to expand students' vocabulary.

Sorting and searching by tag

One of the more entertaining things you can do with the Corpus is to sort your results by tag, essentially separating out different grammatical forms from each other. However, you have to control the results fairly carefully. Manual tagging of a corpus is an error-prone exercise, but with corpora as large as the ČNK it is not even an option.

In an ambitious move, the ÚČNK, in cooperation with the Applied Linguistics Institute of the Charles University Mathematics and Physics Faculty, have written programmes that have automatically tagged the corpora. The results are at times erratic, but fascinating nonetheless.

To see the tags once you've performed a search, go to the *View/Zobrazení* menu, and choose *Atributy/Attributes*. Click the radio button for *Tag* and close the box. You will now see a fifteen-place code that gives all the morphological information for each form. A guide to the tagging used in the Corpus is available at the Institute for Applied Linguistics web site (http://ufal.ms.mff.cuni.cz/), and further information can be found in the online manual, or by following the *Morfologie* link on the ÚČNK web page.

A typical noun tag might look like this: *kráva NNFS1-----A----*. This means that *kráva* is a *NN* noun, of *F* feminine gender in the *S* singular *1* nominative case. The blanks indicate unused categories, and the A means that it is affirmative (no negation). *Umějí*, on the other hand has the tag *VB-P---3P-AA---*. This tells us that it is a *VB* verb in the *P* plural using the *3P* third-person plural ending, and also *AA* affirmative, active voice.

For instance, searching on the word *trhu* 'market (gen./dat./voc./loc. sg.)', I found 4086 examples, which I was quickly able to sort by case. To do this, I went to the *jednoduché třídění/simple sort* command, but asked it to sort tags, instead of lemmas or words. Here it is important to sort a decently large number of positions—I used 15—to make sure it gets the information it needs. The GCQP quickly sorted it into three codes:

NNIS2-----A----
NNIS3-----A----
NNIS6-----A----

representing a noun (inanimate masculine, singular) in the genitive, dative and locative cases respectively. (There were, apparently, no examples of vocatives, which is not surprising.) I was then presented with a wealth of examples, almost too many, so I used the *Reduction/Redukce* command, asking it to choose 400 random rows, which I felt was a manageable amount. (If I wanted more after that, I knew I could always choose *Back to previous/Zpět k předchozímu,* which restores the previous screen.)

Another way to use this feature is to search directly on the morphological tag. This would be useful, for example, in looking at the use of oblique-case numerals, a subject which inevitably comes up in advanced Czech classes, and for which I am heartily tired of trotting out the same lists of examples. First I needed to find out what a tag for a numeral would look like. Instead of trying to make up a probable tag using the manual, I took a shortcut. I did a search on *třemi*, and then asked the processor to display the tags for the queried form, which I copied down:

ClXP7----------

Referring to the manual, I found that this indicated a *Cl* cardinal number of *X* indeterminate gender which was *P* plural and *7* in the instrumental case.

Consulting the manual, I found that tag searches are entered in the query field in the format

[tag="XXXXXXXXXXXXXXX"]

where the actual query replaces the X's.[4] As the manual warns you, you have to make sure to fill all the blanks in the tag with periods (or the period-asterisk). I decided I wanted to see all plural numerals in the instrumental case, so I input the tag as:

[tag="ClXP7.*"]

This threw up 5980 entries in the SYN2000 corpus, which I sorted by lemma, asking for 9 positions (*jednoduché třídění, počet tříděných pozic = 7*) worth of sorting. This separated out *dvěma, oběma, třemi* and *čtyřmi*, as well as sorting by the word following it. (It will not retrieve most numerals of 5 and over, probably because they are not treated as plurals, although it did retrieve numerals like *s dvěma sty*.)

The examples I found were once again far more useful than anything I could have found in a grammar. I found not only tired cliches (*s dvěma malými dětmi*) but also numerous fixed phrases with instrumental numbers (*mezi čtyřmi stěnami, mezi dvěma extrémy/póly*) that are handy for students to know and common enough to feel useful. I also found, interfiled under the lemma *rok*, examples of *před dvěma lety* and *před dvěma roky*. Given time and interest, it would have been possible to excerpt these examples and analyze them to see where there was any logic to the choice or not and what the rough proportions were.

The hitch in using the tags comes with the more difficult—and more interesting—word forms. Buoyed by my success with numerals, I decided to look at a troublesome word like *den*, which has variation in its forms that always leaves me hesitating (when should I use *dní* and when *dnů*?). However, some of the forms of den coincide with real or potential forms of dno, and the tagging software had not been able to distinguish them. Sometimes the tags did not distinguish between dative and locative forms, labeling all of them dative. I then tried to search on adjective forms, using the word červený, and found a significant number of them were mis-tagged in some way.[5] It seems safe to assume that the automatic tagging will improve greatly as time goes on, and that the features that depend on it will become correspondingly more reliable and useful.

Tracking sources in the Corpus

The GCQP also allows you to see the sources of texts. This might be useful if I were getting more deeply into the way words are used. For instance, I could check where and when the various forms of the word web started to appear, and whether there is any difference between technical usage and common usage.

To look at this, I made a chart of all the possible forms of the word:

---

[4] There is a menu-driven way of doing this, but I found it easier to enter the commands manually in the search field.

[5] Predictably, this problem was most evident in places where case syncretism is frequent (as in the masculine and neuter nominative and accusative singular) and where the features assigning case are not easily stated (cf. a preposition that takes one and only one case appearing in close proximity to the target word).

W	e	b
w	e	b
W	e	b
w	e	b
W	e	b
w	e	b
W	e	b
w	e	b

The first letter could be either W or w:

[Ww]

The next two were always eb:

[Ww]eb

It might end there, or might continue with either an e or a u:

[Ww]eb|[Ww]eb[eu]

And the final term is a possible, but not obligatory m, giving me a query of:

[Ww]eb|[Ww]eb[eu]m?

The PUBLIC corpus had 140 results. I then asked it to sort the results by form, and unclicked the ignore case/ignorovat velikost box so that it would separate out Web from web.

Once it had completed the sort, I went to the View/Zobrazení menu and chose Structures/Zdroje. In the dialogue box, I put a check next to doc and chose OK. Over to the left of the screen, next to each line, there appeared a code much like this one:

<doc S|PUB|1998|pr980722>

This shows the genre, year and exact source of the citation. Unfortunately, there isn't as of yet any guide as to how exactly to read the notation, so it's less helpful than it might be. (The ÚČNK confirms that a guide is in the planning stages.) However, one can guess at certain abbreviations. PUB is most likely journalism, with the final entry being a name of a journal or paper and a date (here Právo for 22 July 1998). My guess was confirmed by looking at other entries.

From this search we can gather a few facts. Forms like webu are noticeably more frequent in 1999, and especially in newspapers; Lidové noviny seems especially fond of it. However, an earlier example from LN in 1997 runs: jak pomocí "World Wide Web" úspěšně provozovat reklamu a marketing… Individual journals may or may not have a policy about how this word is written; in one, Ikaros, it appears as webu, Webu or the uninflected World Wide Web. This seems to confirm what we proposed earlier: that there is considerable variation in the use of this word, and that it is evolving from an undeclinable, foreign proper noun into a declined, domesticated ordinary noun.

I have only scratched the surface here of what can be done with the GCQP. As I worked through these examples using the GCQP, I had a tantalizing glimpse of its power to provide accurate, timely, real-life data on a wide range of intriguing research topics in Czech morphology, syntax, registers and discourse. Research is, in fact, where the GCQP comes into its own; using it only for the occasional casual query would be like installing a spanking new, fully-equipped kitchen in which you only ever used the toaster.

u
u
e
e
e
e

Conclusions

The Czech National Corpus can be a valuable tool for non-native-speaker teachers of Czech as a foreign language, and native-speaker teachers as well. The basics take a few minutes to master, but the results are indisputable: accurate, up-to-date citations, and, for those who aren't afraid to interpret it for themselves, information on a broad variety of linguistic themes that can give us a broader outlook on current usage. It's there; it's free: I for one will certainly be using it.

Czech Cultural Studies Workshop

Jindrich Toman,
the University of Michigan
at Ann Arbor

The first annual Czech Cultural Studies Workshop, sponsored by the Davidson Institute, the Department of Slavic Languages & Literatures, the Institute for the Humanities, all of the University of Michigan, and the North American Association of Teachers of Czech, took place from April 14-16, 2000, at the University of Michigan, Ann Arbor.

Panels on History/Society, Literature, Nation/ Symbols, Economy & Culture/Past & Present, were enthusiastically attended. Participants were doctoral candidates and junior scholars in Czech Studies from more than 10 institutions, including Columbia University, Northwestern University, Princeton University, The College of New Jersey, University of California at Berkeley, University of California at