František Čermák
Praga

MINOR LANGUAGES IN TODAY'S EUROPE: THE CZECH CASE

Prispevek najprej obravnava jezikovna stanja v evropskih državah in s tem povezane težave. Kot del analize dejavnikov, ki prispevajo k različnim jezikovnim stanjem, so omenjeni nekateri družbeni in politični dejavniki; dodana je tudi kvantitativna razdelitev malih jezikov. Sledi kratek oris češkega jezikovnega stanja; opisane so potrebe po novi jezikovni politiki. Avtor trdi, da brez obsežnega jezikovnega korpusa in analize le-tega ni možen nadaljnji razvoj jezika in jezikovne politike, ki bi temeljil na trdni in prepričljivi osnovi. Na koncu opiše češko rešitev problema, katere vrh so dosegli s projektom Češki nacionalni korpus.

In its first part, the contribution surveys the language situation in European countries and some of the problems connected to it. As part of the analysis of factors contributing to the various situations, some sociological and political factors are taken up; a quantitative classification of minor languages is added. Against this background the Czech language situation is briefly sketched and some desiderata for a new language policy suggested. It is argued that no future development of a language and language policy, based on a sound and convincing basis, is possible without a large language corpus and its analysis. Finally, the Czech solution to this problem is described, its culmination being the Czech National Corpus project.

1 Minor Languages in Europe

Some 96 % of mankind uses only 4 % of existing 6000–7000 languages, while the rest use small languages (Crystal 2000), whatever »small language« might mean. A pessimistic estimate for the world is that about half of existing languages will die in the foreseeable future (see also *Foundation for Endangered Languages, www.ogmios.org*). The situation in Europe seems to be much better, however. But also here, a mixture of factors contributes to a language being in good health, while another is not so well. A brief look at what is, traditionally, considered Europe follows; it leaves out, e.g., Russia because of lack of information on its present language policy, which used to be very bad, destructive and caused the death of a number of languages in the past (it is now estimated that only in its European part, there still exist some 56 languages). Language death is usually due to the replacement of the old language by a new one with social and political prestige. In other words, the old language ceased to fulfil the social needs of the community or

its last speaker died (naturally or due to genocide). Two decisive sets of factors can, at least, be singled out here: official support on the one hand, and the natives' determination to continue using their language if it fulfils its basic functions, on the other. However, this rather simple situation may, if considered carefully, not seem so simple at all. Leaving aside such atrocious facts as genocide, displacement or natural disaster, we need to mention a more civil and pragmatic factor. Namely, there are often parents who feel that using a small/minor or indigenous language at school will hold their children back, and that better chances for them are linked with a high-prestige language, such as English, a case noted in many countries now. A phenomenon of globalization, but not quite. So let us first have a brief look at both types of factors.

1.1 Psycho-Political Aspects

Virtual language death is now rare in Europe. Facts stating that, for example, the last Dalmatian speaker in Croatia died on June 10th 1898 (Encyclopaedia Britannica) seem to be part of history now. The contemporary European situation may be presented as a scalar image, with two opposing **extremes**, where, all things considered, the language situation is either very bad or, at the opposite end, very good. Thus, the situation may seem (**1**) **very bad** for Byelorussian (7,5 million speakers) where people seem to lack the will to continue using their own language, but also for some other languages, such as Ruthenian (1,5 million), Kashubian (0,2 million), Breton (0,5 million) or Arumanian (0,1 million), which are all doing rather badly, mostly due to scant support from their stronger neighbours. A special case is Irish Gaelic (with only 0,26 million speakers being fluent in it) and Lower Sorbian (5000?), despite all kinds of official support.

The neutral **middle ground** belongs to (**2**) most European minor languages, although there are differences among them too. Many receive institutional and/or state support, having their own Academias, newspapers, church services etc., although they are sometimes used on territories extending over the official state boundaries: these include (**2A**) Scots (88.000? speakers), Frisian (over 1 million), Basque (approx. 0,6 million), and Saami (approx. 20.000?). While this subgroup seems to be doing well, other Eastern and Central European languages (**2B**) feel a need for language legislation, a dubious call for protectionism with no certainty of being really useful.

On the other hand, (**3**) some minor languages are **doing very well**, such as Luxembourgian (0,3 million speakers), which was only recently promoted to official language status and given recognition, considerable investments etc., although this move may seem to be somewhat artificial, especially to native Germans. A very special case is Cornish, whose last native speaker died in 1877; yet the language has been revived and is dutifully supported by the EU, having now about 150 speakers.

A different case to be considered outside the three groups mentioned is (**4**) Romany (5? million speakers), where there is a growing split between dialects with little or no contact (e.g. in Spain, UK or Eastern Europe) because of its scattered existence in many countries and lack of common or adjacent territory. Yet, in some countries at least, there is a support to be found for this language. A similar case, where the split seems to widen due to political reasons, is to be found in Bosnian (4 million?), which is recognized as an independent language by some.

1.2 Quantitative Aspects

Using the *Ethnologue* survey (*www.sil.org/ethnologue*), one may distinguish at least three groups among the minor European languages which have:

(**A**) **under or around five million speakers:** Albanian, Allemanisch, Catalan, Croatian, Danish, Finnish, Galician, Norwegian (17 % Nynorsk), Sicilian, Slovak, etc.

(**B**) **under or around three million speakers:** Latvian, Lithuanian, Macedonian, Piedmontese, Slovene, Venetian, etc.

(**C**) **under or around one million speakers:** Basque, Breton, Corsican, Estonian, Frisian, Friulian, Gaelic, Icelandic, Kashubian, Ligurian, Lombardic, Luxembourgian, Provencal, Saami, Sardinian, Upper Sorbian, Welsh, etc.

Comparing the figures in the above survey, it is evident that there is no critical mass leading to either language growth or language decline: some minor languages are doing very well (Icelandic 0,23 million speakers), while some other, although having a larger population of speakers, such as Byelorussian, are not so well.

2 The Czech Language Situation

By the above standards, Czech may not be counted among very minor languages, but it is a minor language in its own right, with some special features as well as problems. It is not my aim here to describe to what degree diglossia can be found in Czech, a fact which is to be viewed as a result of ill-conceived and elitist language policy in Czechia over the past decades. Leaving aside the eastern, Moravian territory, a much smaller linguistic area (!!) than Bohemia, most Czech speakers use, in their everyday speech, *Common Czech (obecná čeština)* as their first language, which is rather different from traditional dialects. This is the language they have been born into, to be told only six years later, when entering school, that there is something wrong with their language and that the correct language is to be expected of the teacher.

Obviously, this is a situation which is now difficult to accept, despite all kinds of allegations coming from academic »old-timers«, who will never know better, about the prestigious, beautiful and ideal shape of official *Literary Czech (spisovná čeština).* In this situation of a largely artificial and prescribed standard meant to supercede the kind of natural vernacular used by the majority of the natives, the

obvious question to be asked is what kind of psychological and sociological attitudes the population could adopt towards the standard. Not very positive, I suppose. Yet not much research has been undertaken along these lines, and the only official support to be found is that given to the literary standard. However, this officially promoted standard is, in fact, far from ideal, being now more and more reserved for official written use only. It has been amply demonstrated that the literary standard hardly fulfils half of the communicative functions which a language should perform (in Jakobson's sense, see, e.g., Čermák 1996). Despite the recent misguided and ill-conceived spelling reform, the basis of which was prepared by Communist »old-timers« at the Academy and which at first met with a loud and hostile public refusal; it now seems that a certain slackening of the hitherto rigid power relations may be observed.

Due to a growing realisation that there is more than one, namely the rigidly codified, form of the language, books are now also being written and published in the *Common Czech* standard. It is obvious that each of the standards, the largely spoken *Common Czech*, the basically written *Literary Czech*, and the spoken local *dialects* (in Moravia, mostly), has its own place and a function to perform, and all should be recognized and supported. It is still an open question whether this quasi-diglossic situation will find its solution in a changed and more open attitude of the official codifiers of the literary standard (Čermák 1987, 1993, Sgall – Hronek – Horecký – Stich 1992).

But it now seems that a thorough new research, based on fresh and ample language facts and their contexts, is needed (Labov 1972). In the light of this knowledge, much of the old tradition of grammars, dictionaries and textbooks, based on the idea of *codification,* may look suspicious and problematic as these publications have not really been founded on sufficient data and knowledge of usage. Instead, prescriptive policy of declarations and stipulations has been used, often based on wishful thinking of a handful of officially recognized linguists and officials.

Times have changed, however, and the old idea of *purism* (Thomas 1991), pathetically and pathologically hostile to any foreign borrowings, and that of the subsequent *protectionism* seem now both futile and meaningless. Languages have always influenced each other and there is no putting a stop to it, all the more so in an era of globalization. Instead, reasonable and viable goals to set a revised language policy should be re-defined, and inevitable and practical *linguistic division of labour* recognized, in which a native language co-exists with major or international languages. A new and, hopefully, **reasonable and democratic, language policy** will need to give up dictatorial and prescriptive practices, become open to observation and make a practice of qualified recommendation. After all, major language engineering attempts of the past, such as the introduction of *samnorsk* in Norway or the misplaced official support to *katharevousa* in Greece, have failed, although they were made with the best of linguistic intentions. Perhaps, the two really **basic functions** a language must fulfil are the following: *intergenerational*

linguistic socialization (outside family) and giving its community *a sense of identity*. Which other functions, apart from these basic ones, can be fulfilled by a minor language is, naturally, a matter of research and funds available.

One can think of a number of more important matters to which the official support should go, instead of, for example, to the latest problematic and unsuccessful spelling reform, a deplorable case in recent Czech history. Among some other obvious things, it seems necessary to promote, through language awareness, *authenticity and expression of identification with the community*, *increase in one's own language prestige, access to electronic technology and computer literacy*, with all these finding a solid basis in a *modern descriptive, not prescriptive, dictionary*. Provided a large database for new research is available, one should also take into consideration the vexing problem of *terminology*, for instance, as the largest part of any modern language consists of terminology, a fact which many linguists still refuse to recognize.

3 The Czech Language (Electronic) Project

Instead of adhering to the old practice of academic and official decision-making and prescription, some people opted for an alternative. They felt that any decisions and judgements concerning the language have to be based on modern data, now obviously computer corpus data. The importance of large corpora has become obvious, and they have become indispensable in every area of life in many countries. The multiple goals for the build-up of the *Czech National Corpus* (CNC), one of which was collecting the data for a new and badly needed dictionary of Czech, are now supplemented by the obvious benefits to be gained from the project: contributing to the preservation of the language and establishing a unique source of information. It is also obvious that a national corpus, an idea which is becoming very popular in many countries, should be made as open as possible and viewed as one of the truly national institutions.

The project began to take shape with the foundation of the Department of the Czech National Corpus, established in 1994 at Charles University; it is supported by a number of institutions and state grants.

3.1 The Czech National Corpus: Overall Design

The general framework of the Czech National Corpus project includes, in fact, more than one corpus being made at the same time. Apart from the *synchronic corpus SYN2000*, whose first but not final goal was to arrive at 100 million running words (wordforms) of contemporary written Czech in a representative form, there is now an extensively tagged corpus of authentic Spoken Czech (*Prague Spoken Corpus*) available, as well as a *Corpus of old Czech* covering all stages up to the present time which is still growing; some attempts have been made to establish a *dialect corpus*, too (Čermák 1997, 1999, 2000).

The synchronic corpus, based on multiple research with a bias towards language reception, is viewed as representative, covering, in a balanced and well-argued manner, a range of 20th century language. Most of it, however, comes from the last decade of the 20th century. Since CNC is an academic project, it has been decided to make it accessible free of charge to any academic, researcher or student (for details, visit *http://ucnk.ff.cuni.cz*), and it is also planned to be widely used by schools as it happens to be the largest electronic information source for the country. In addition to limited public access, however, full access can be obtained by signing an agreement. Needless to say that the community of researchers, both from the Czech Republic and abroad, and students who gladly take this opportunity, is steadily growing.

Obviously, such a project, once established and invested into, should be considered a standing task with an open end needing permanent support. It is not only necessary to keep one's language under constant linguistic surveillance from which society as a whole will benefit, but also to record language changes as well as to link it with its past. To have such a vast repository and data bank is of paramount importance for any language.

References

David CRYSTAL, 2000: *Language Death*. Cambridge: Cambridge U. P.

František ČERMÁK, 1987: Relations of Spoken and Written Czech (With Special Reference to the Varying Degree of Acceptability of Spoken Elements in Written Language). *Wiener Slawistischer Almanach* 20. 133–150.

– – 1993: Spoken Czech. *Varieties of Czech. Studies in Czech Sociolinguistics.* Ed. E. Eckert. Amsterdam – Atlanta: Rodopi. 27–41.

– – 1995: Jazykový korpus: Prostředek a zdroj poznání. *Slovo a slovesnost* 56. 119–140.

– – 1996: Obecná a spisovná čeština: Poměr, funkce a metodologie. *Spisovnost a nespisovnost dnes*. Ed. R. Šrámek. Brno. 14–18.

– – 1997: Czech National Corpus: A Case in Many Contexts. *International Journal of Corpus Linguistics* 2/2. 181–197.

– – 1999: Czech National Corpus: Its Character, Goal and Background. *Text, Speech, Dialogue, Proceedings of the First Workshop on Text, Speech, Dialogue-TSD'98.* Eds. P. Sojka – V. Matoušek – K. Pala – I. Kopeček. Brno: Masaryk University. 9–14.

– – 2000a: Linguistics, Corpora and Information. *PALC '99: Practical Applications in Language Corpora, Łodź Studies in Language*. Eds. B. Lewandowska-Tomaszcyk – P. J. Melia – P. Lang. Frankfurt am M. – Berlin.193–201.

– – 2000b: *Český národní korpus: Úvod a příručka uživatele*. Eds. J. Kocek – M. Kopřivová – K. Kučera. Praha: FFUK, Ústav českého národní korpusu.

William LABOV, 1972: The Study of Language in its Social Context. *Language and Social Context*. Ed. P. P. Giglioli. Harmondsworth: Penguin Books. 283–307.

Petr SGALL – Jiří HRONEK, 1992: *Čeština bez příkras*. Praha: H+H.

Petr Sgall – Jiří Hronek – Alexandr Stich – Ján Horecký, 1992: *Variation in Language. Code-Switching in Czech as a Challenge for Sociolinguistics*. Amsterdam: J. Benjamins.
Zdeněk Starý, 1995: *Ve jménu funkce a intervence.* Praha: Karolinumillion.
George Thomas, 1991: *Linguistic Purismillion*. London – New York: Longman.

MALI JEZIKI V DANAŠNJI EVROPI: ČEŠKI PRIMER

Povzetek

Različne probleme, ki se tičejo lastnega jezika, lahko rešimo le s pomočjo zanesljivejših podatkov. S tem lahko zavržemo nekdanje posplošene trditve jezikoslovcev, ki se niso mogli opirati na zadostne podatke. Specifično rešitev, ki je pogoj za nadaljnje raziskave, najdemo v velikem, reprezentativnem korpusu, ki bi moral biti odprt tako za uporabnike kot v prihodnost.