Human Language Technologies – The Baltic Perspective
K. Muischnek and K. Müürisep (Eds.)

doi:10.3233/978-1-61499-912-6-158

Czech & Slovak Corpus Resources Go (not only) Latvian

Michal ŠKRABAL[a,1] and Vladimír BENKO[b]
[a] *Charles University in Prague, Institute of the Czech National Corpus*
[b] *Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics & Comenius University in Bratislava, UNESCO Chair in Plurilingual and Multicultural Communication*

Abstract. As Latvian can still be considered an under-resourced language, several corpora and corpus tools that can be used for its linguistic research are presented in the paper, namely: the *InterCorp* and *Araneum Lettonicum* corpora along with the *Treq* database, a word-sketch grammar for Latvian and the *Morfio* tool.

Keywords. Baltic languages, parallel and comparable corpora, translation equivalents, sketch grammar

1. Introduction

In our paper, we would like to introduce several corpora and corpus tools that can be used for the linguistic research of Baltic[2] languages, possibly with other follow-up NLP applications. Specifically, they are two corpora: the Latvian component of the *InterCorp* parallel corpus (2) and *Araneum Lettonicum* (3) and two corpus tools based on these corpora: *Treq* (4) and a word-sketch grammar for Latvian (5), along with *Morfio* (6) in the near future. We want to make them known to the professional public and encourage everybody to use them, as they are generally accessible, free-of-charge tools. Our attention is focused mainly on Latvian, the other two Baltic languages are also however taken into account. The motivation is twofold: firstly, we consider Latvian to still be an under-resourced language,[3] and secondly, we would consider our collections of corpora to be incomplete without the Baltic languages.

In principle, the instruments described below are – after necessary modifications – applicable to any language (and available upon request, as well as the data). In our paper, we offer both creator- and user-oriented perspectives, as all of these sources are

[1] Corresponding author: Charles University, Faculty of Arts, Institute of the Czech National Corpus, Panská 7, 110 00 Praha 1, Czech Republic; e-mail: michal.skrabal@ff.cuni.cz.

[2] By Baltic languages we mean not only Lithuanian and Latvian but – considering the HLT Baltic's focus – the languages of all three Baltic states, i.e. also including Estonian.

[3] Until recently only a few corpora of Latvian were available, in the lead of the balanced corpus of contemporary Latvian *LVK2013* (4,5M tokens), updated this year to *LVK2018* (10M tokens). Besides that, there are some corpora available at the Sketch Engine portal, including the web-crawled *lvtenten* (530M tokens) or the *EUR-Lex Latvian* version (325M tokens). A complete list of Latvian corpora and corpus tools can be found at http://www.korpuss.lv/ and https://www.sketchengine.eu/user-guide/user-manual/corpora/by-language/latvian-text-corpora/.

actively used in the compilation of the emerging Latvian-Czech dictionary. Nevertheless, the scope of their possible application is much wider (translation, translatology, language pedagogy, etc.).

2. *InterCorp*

InterCorp (*IC*) is a large parallel synchronic corpus under continuous construction at the Institute of the Czech National Corpus since 2005 [1; 2]. It is available via the *KonText* interface on the website https://www.korpus.cz/. Unlike other parallel corpora, in particular the web-crawled ones, *IC* also includes literary texts with manually corrected OCR and sentence alignment. In addition, there are several "collections" consisting of texts which were only processed automatically, not manually. These include the following types of texts:

- legal texts of the European Union from the *Acquis Communautaire* corpus;
- journalistic articles and news published by *Project Syndicate* and *VoxEurop*;
- proceedings of the European Parliament dated 2007–2011 from the *Europarl* corpus;
- movie subtitles from the *Open Subtitles* database;
- the Bible.

The up-to-date version *IC v10* contains, besides Czech as the pivot language, other 39 languages that are, however, unevenly represented. Texts in more than half of the languages are provided with morphological annotation (23 out of 39) and lemmatized (20 out of 39), including both Latvian and Estonian, whereas Lithuanian texts are neither annotated nor lemmatized. The total size of *IC v10* is more than 1.48 billion running words / 1.87 billion tokens.

Table 1. Baltic components of *InterCorp v10* (thousands of tokens)

ISO Code	**Et**	**lt**	**lv**
Language	Estonian	Lithuanian	Latvian
PoS tagged & lemmatized	yes	no	yes
Fiction	0	358	2,025
Syndicate	0	0	0
PressEurop	0	0	0
Acquis	14,896	17,316	17,533
Europarl	10,899	11,213	11,682
Subtitles	10,298	558	280
Bible	0	471	0
Total	**36,093**	**29,916**	**31,521**

3. *Araneum Lettonicum*

The *Aranea* Project is targeted at the creation of a family of web corpora that could be used as a tool for teaching language, linguistics and translatology-related subjects, as well as for research in various areas of language studies. As all the corpora are being compiled by uniform methodology and have the same size, they can be conveniently used for contrastive research [3].

To create the Latvian web corpus a set of tools referred to as the "Brno pipeline" has been used. Its main components are the *SpiderLing* crawler that also incorporates modules for the boilerplate removal (*jusText*) and web page encoding detection (*Chared*), the near-duplicate removal tool (*Onion*) and the universal tokenizer (*Unitok*). All these tools are available from the Corpus tools web page (http://corpus.tools/).

The data has been crawled in two subsequent 14-day sessions during the summer of 2017 with the seed URSs being obtained by the *WebBootCAT* tool. The pace of the process is shown in Figure 1.

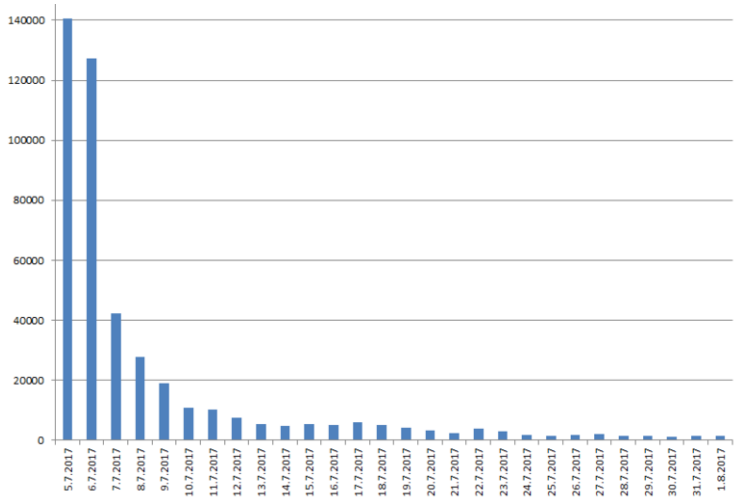


Figure 1. Crawling the Latvian data (deduplicated documents per day)

The yield had dropped dramatically at the end of the crawling, we therefore could not see any reason at that time to continue.

The downloaded data has been deduplicated at the document level (using 5-grams and 95% similarity threshold) and tokenized with the generic *Unitok* parameter file. The resulting vertical file has been PoS-tagged by the *LU MII Tagger* [4] using the customized *MULTEXT-East* tagset. The "native" tags have subsequently been mapped to *Universal Araneum Tagset* providing a parallel layer of "PoS-only" annotation. The result of this mapping is shown in Table 2.

Table 2. PoS mapping

PoS	Atag	%	count
noun	Nn	32.62	74,769,997
adjective	Aj	4.88	11,175,291
pronoun	Pn	6.33	14,518,319
numeral	Nm	0.84	1,928,764
verb	Vb	14.65	33,581,185
adverb	Av	4.73	10,830,559

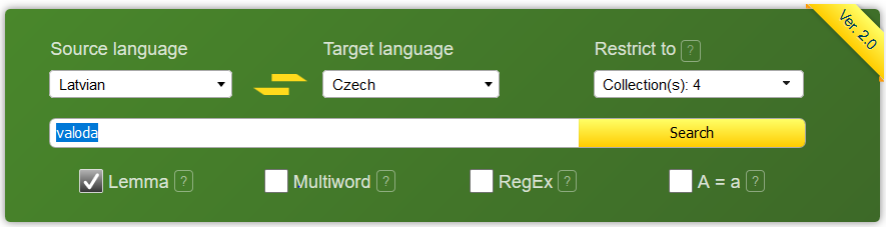
PoS	atag	%	count
preposition	Pp	4.20	9,636,825
conjunction	Cj	5.73	13,140,813
interjection	Ij	0.16	377,996
particle	Pt	1.47	3,363,914
unknown	Yy	24.37	55,861,074
	Total		229,184,737

Due to an error in the mapping procedure, the "Yy" tag (unknown) also accommodates punctuation, numbers and symbols, i.e. it does not show the real number of unrecognized word forms. This is expected to be fixed by the time of the conference.

The *Araneum Lettonicum* corpus along with the respective sketch grammar (see Ch. 5) is not publicly available online yet, but we expect it to be uploaded to the Sketch Engine website by the time of the conference.

4. Treq

Data from the *InterCorp* corpus (see Ch. 2) are further processed using automatic tools: original and translation texts are first word-to-word aligned using the *GIZA++* tool [5]. The aligned pairs of words are then sorted and summarized. The result of this automatic excerption is not revised in any way and is provided to users as a list of found equivalents of the given expression on website https://treq.korpus.cz.



▲ Frequency ▼	▲ Proportion ▼	▲ Latvian ▼	▲ Czech ▼
4004	77.0	valoda	jazyk
958	18.4	valoda	jazykový
77	1.5	valoda	řeč
49	0.9	valoda	znění
21	0.4	valoda	závazný
14	0.3	valoda	lingvistický
7	0.1	valoda	verze
7	0.1	valoda	jazykově
7	0.1	valoda	čeština
6	0.1	valoda	hovořící
6	0.1	valoda	slovo
4	0.1	valoda	1
3	0.1	valoda	výraz

Figure 2. Simple query in the *Treq* database (for lemma *valoda* "language").

Treq v2.0 [6] brings a number of improvements: in addition to a more user-friendly and clearer interface, it is now possible to enter multi-word expressions (bidirectionally) in order to get both one-word and multi-word results. With the implementation of multi-word units, the need to incorporate a query language that would allow the use of wild cards has become urgent: up to now *Treq* has only been searching for the exact string of characters. Furthermore, a second primary language (besides Czech), namely English, has been added, and in addition to the existing

bidirectional Czech-X lexicons, bidirectional English-X lexicons have also been generated from the *InterCorp* data. Thus, the possibility of using *Treq* is now opened up to a much wider audience as users are no longer limited by the need to master Czech.

5. The Latvian Sketch Grammar

The last resource we would like to present is the custom Latvian sketch grammar developed along with other sketch grammars for *Aranea* corpora, as suggested by [7]. Unlike most other grammars available at the Sketch Engine portal, "gramrel" names appearing in the headings of the respective tables do not indicate the syntactic but rather just the collocational, relationships. For example, we do not speak about subjects, objects or modifiers but rather only about the left-hand/right-hand noun or adjective collocates. Figure 3 depicts an example Latvian word sketch.

The "X" symbol in each table stands for the "keyword", i.e., *vīns* ("wine") in our case, and the PoS symbols at the left or right side of "X" indicate the collocates of the respective word class. The "Y" symbol stands for the "immediate" collocate of any word class.

As sketch grammars of this type can be written for (almost) any language, they can conveniently make use of the "bilingual sketch" functionality of Sketch Engine. Figure 4 shows an excerpt from such a Latvian-English bilingual sketch.

6. Conclusion and Further Work

All these tools are already available for users. Besides that, work on the Latvian version of the *Morfio* tool is ongoing. *Morfio* serves to give estimates of the extent and productivity of morphological models based on corpus data. It is therefore a tool which can be used in morphological research, especially for the study of derivation. Originally, it was created for Czech [8; 9], yet nothing prevents it from extending its functionality to other languages,[4] including the Baltic ones. For a fully-fledged and user-friendly non-Czech version of the tool, different tagsets must be implemented, and an inventory of relevant alternations (for both vocals and consonants) must be added. If the tool is to be used by the non-Czech research community, it would be appropriate to take the appropriate language mutation of the interface, including brief help and documentation.

Development of the tools described above does not stop, of course. Further improvements in the provided results can be expected in proportion to the increasing volume of data, the greater genre diversity of texts and the gradual improvement of automatic word alignment tools. However, they will never be comparable to manual alignment: a certain error rate is therefore inevitable and must be taken as the necessary tax for the possibility of effectively investigating a manually unmanageable volume of data.

[4] In fact, *Morfio* has already been successfully applied to the Polish part of *InterCorp* [10; 11]. Another language we want to test *Morfio* on is Latvian. As a data base, similarly to Polish, the *InterCorp* Latvian component can be used, possibly along with a representative corpus of contemporary Latvian LVK2018.

Regarding the web corpus, we would also like to increase its size, improve the morphosyntactic annotation and, naturally, complete the language collection within the *Aranea* family by the other two Baltic languages – Lithuanian and Estonian.

vīns **Araneum Lettonicum Parvum (Latvian, 17.08) 169 M freq** = 8,656 (51.18 per million)

X X		
	6,239	72.08
dzirkstīt	196	9.85
Sabile	64	7.94
dzert	132	7.92
vīnoga	53	7.79
pienene	36	7.46
malkot	33	7.34
sarkans	68	7.18
iedzert	34	7.00
deserts	27	6.95
Roza	34	6.93
pieliet	28	6.84
sauss	33	6.74
rabarbers	22	6.70
ābols	30	6.46
Francis	38	6.37
salds	25	6.35
oga	24	6.31
Gruzija	23	6.27
degustēt	15	6.22
itālis	23	6.20
izdzert	19	6.16
avene	14	6.01

X Y		
	5,120	59.15
glāze	310	9.89
darītava	171	9.83
degustācija	98	9.03
pudele	180	8.89
darīšana	77	8.43
pagrabs	73	8.03
dārzs	131	7.40
kalns	128	7.37
baudīšana	30	7.30
bārs	41	7.22
etiķis	29	7.19
pazinējs	21	6.93
cienītājs	35	6.83
gatavošana	32	6.79
skapis	29	6.74
korķis	20	6.71
raugs	20	6.63
šķirne	33	6.56
ražošana	77	6.39
garša	38	6.25
degustēšana	12	6.23
etiķete	15	6.11

Nn X		
	4,203	48.56
vīnoga	77	8.43
Sabile	74	8.00
pienene	38	7.94
deserts	28	7.26
rabarbers	23	7.09
Roza	36	6.96
oga	31	6.51
glāze	35	6.48
Gruzija	32	6.42
vīndaris	12	6.41
avene	15	6.40
Aleksis	13	6.35
ābols	31	6.26
itālis	30	6.24
Čīle	13	6.24
Abava	14	6.23
biķeris	10	6.16
Jaunzēlande	14	6.06
Vīna	13	6.04
pīlādzis	10	6.01
Francis	44	5.97
Marsala	8	5.95

X Nn		
	5,776	66.73
darītava	173	9.40
glāze	324	9.22
degustācija	103	8.70
pudele	195	8.16
darīšana	79	7.84
pagrabs	81	7.53
korķis	36	7.20
baudīšana	32	6.94
vīnoga	35	6.92
etiķis	30	6.86
pazinējs	22	6.67
bārs	42	6.65
dārzs	133	6.43
kalns	130	6.43
raugs	23	6.42
vīnzinis	15	6.35
etiķete	24	6.33
šķirne	45	6.13
skapis	29	6.09
gatavošana	33	6.08
muca	18	6.07
degustēšana	12	6.03

Aj X		
	1,492	17.24
sauss	40	7.84
pussauss	10	7.75
sarkans	75	7.68
salds	27	7.29
bordo	7	7.09
izsmalcināts	10	7.05
pussalds	6	7.02
sārts	6	6.71
gards	10	6.48
skābs	7	6.39
balts	36	6.34
izsmalcināta	6	6.26
kvalitatīvs	21	6.21
dārgs	22	6.20
lēts	19	6.03
garšīgs	7	5.97
viegls	20	5.92
smalks	8	5.64
vietējs	32	5.23
slavena	7	5.18
izcils	15	5.13
lielisks	18	5.04

X Aj		
	278	3.21
vietējs	6	2.36
augsts	6	1.51
jauna	9	0.42
jauns	8	0.06
labs	7	-0.15

Vb X/X Vb		
	8,344	96.40
dzirkstīt	206	9.14
malkot	55	7.48
degustēt	46	7.19
dzert	226	6.74
nobaudīt	47	6.59
iedzert	60	6.53
raudzēt	25	6.35
pieliet	34	6.18
ieliet	25	5.87
nedzert	21	5.83
izdzert	40	5.82
iemalkot	13	5.52
nogaršot	19	5.48
baudīt	113	5.41
cienāt	17	5.35
nodegustēt	11	5.34
garšot	31	5.33
rimt	20	5.31
baudītāt	10	5.26
pārliet	18	5.18
liet	28	5.12
saderēt	13	5.12

Av X/X Av		
	652	7.53
klāt	13	3.03
joprojām	10	1.64
nu	11	1.55
savukārt	14	1.45
kopā	28	1.31
pavisam	7	1.29
vispār	8	1.04
nedaudz	7	1.04
labi	22	1.01
šogad	8	0.82
kad	30	0.73
parasti	6	0.67
pāri	6	0.62
gandrīz	6	0.55
kur	17	0.47
vēl	28	0.25
īpaši	6	0.25
tad	23	0.16
tagad	9	0.08
bieži	7	0.08
tā	14	0.02
vismaz	6	0.01

Figure 3. Word sketch for *vīns* "wine"

Aj X			Aj X		
	1,492	17.24		24,516	40.48
sauss	40	7.84	sparkling	579	8.16
pussauss	10	7.75	red	2,093	6.09
sarkans	75	7.68	biodynamic	40	5.26
salds	27	7.29	Tuscan	41	5.11
bordo	7	7.09	fruity	43	5.04
izsmalcināts	10	7.05	full-bodied	29	5.03
pussalds	6	7.02	varietal	33	5.03
sārts	6	6.71	complimentary	98	4.95
gards	10	6.48	tannic	23	4.83
skābs	7	6.39	white	1,252	4.62
balts	36	6.34	Chilean	35	4.53
izsmalcināta	6	6.26	homemade	75	4.30
kvalitatīvs	21	6.21	fine	779	4.30
dārgs	22	6.20	vintage	96	4.30
lēts	19	6.03	kosher	29	4.24
garšīgs	7	5.97	claret	15	4.13
viegls	20	5.92	Sicilian	17	4.03
smalks	8	5.64	Portuguese	50	4.00
vietējs	32	5.23	plum	23	3.87
slavena	7	5.18	award-winning	64	3.84
izcils	15	5.13	world-class	49	3.83
lielisks	18	5.04	sweet	243	3.81

X Aj		
	278	3.21
vietējs	6	2.36
augsts	6	1.51
jauna	9	0.42
jauns	8	0.06
labs	7	-0.15

X Aj		
	7,317	12.08
varietal	11	4.50
fruity	13	4.06
approachable	9	3.79
juicy	10	2.83
ripe	15	2.50
drunk	20	2.39
sour	10	2.34
superb	17	2.16
appealing	9	2.15
crisp	11	2.13
charming	11	1.79
delicious	28	1.75
cherry	10	1.75
tasty	10	1.69
exotic	10	1.40
exceptional	17	1.19
renowned	9	1.19
retail	28	1.18
elegant	10	1.10
sweet	33	0.97
bold	12	0.97
distinctive	10	0.95

Vb X/X Vb			Vb X/X Vb		
	8,344	96.40		76,761	126.76
dzirkstīt	206	9.14	taste	1,726	6.97
malkot	55	7.48	mull	236	6.27
degustēt	46	7.19	drink	1,773	6.26
dzert	226	6.74	sip	281	6.14
nobaudīt	47	6.59	pair	582	6.02
iedzert	60	6.53	pour	524	5.10
raudzēt	25	6.35	ferment	130	5.09
pieliet	34	6.18	bottle	83	4.73
ieliet	25	5.87	sample	181	4.70
nedzert	21	5.83	fortify	91	4.42
izdzert	40	5.82	chill	85	4.20
iemalkot	13	5.52	cellar	43	4.13
nogaršot	19	5.48	craft	129	3.81
baudīt	113	5.41	spill	79	3.73
cienāt	17	5.35	decant	31	3.65
nodegustēt	11	5.34	complement	86	3.48
garšot	31	5.33	import	118	3.42
rimt	20	5.31	box	42	3.40
baudītāt	10	5.26	stock	59	3.37
pārliet	18	5.18	flavor	36	3.31
liet	28	5.12	age	217	3.30
saderēt	13	5.12	uncork	24	3.30

Av X/X Av			Av X/X Av		
	652	7.53		12,553	20.73
klāt	13	3.03	beautifully	36	2.96
joprojām	10	1.64	nicely	27	2.64
nu	11	1.55	moderately	10	2.60
savukārt	14	1.45	freely	36	2.27
kopā	28	1.31	locally	28	2.20
pavisam	7	1.29	intensely	9	2.11
vispār	8	1.04	proudly	11	2.08
nedaudz	7	1.04	specially	19	2.07
labi	22	1.01	reasonably	24	1.87
šogad	8	0.82	perfectly	42	1.60
kad	30	0.73	exclusively	24	1.58
parasti	6	0.67	traditionally	16	1.51
pāri	6	0.62	consistently	25	1.48
gandrīz	6	0.55	internationally	14	1.25
kur	17	0.47	annually	19	1.25
vēl	28	0.25	rarely	23	1.14
īpaši	6	0.25	surprisingly	9	1.10
tad	23	0.16	carefully	30	1.01
tagad	9	0.08	typically	39	0.93
bieži	7	0.08	naturally	23	0.90
tā	14	0.02	primarily	34	0.88
vismaz	6	0.01	slowly	23	0.77

Figure 4. Bilingual word sketch for *vīns* / *wine* (excerpt).

Acknowledgement

This work has been, in part, funded by the Slovak KEGA and VEGA Grant Agencies, Project No. K-16-022-00 and 2/0017/17, respectively. It was also supported by the European Regional Development Fund-Project “Creativity and

Adaptability as Conditions of the Success of Europe in an Interrelated World" (No. CZ.02.1.01/0.0/0.0/16_019/0000734). During its creation we used the tools developed within the Czech National Corpus project (LM2015044) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

References

[1] F. Čermák, A. Rosen, The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics* **3** (2012), 411–427.

[2] A. Rosen, InterCorp – a look behind the façade of a parallel corpus, in: E. Gruszczyńska, A. Leńko-Szymańska (eds.), *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora.* Instytut Lingwistyki Stosowanej, Warszawa, 2016, 21–40.

[3] V. Benko, Aranea: Yet Another Family of (Comparable) Web Corpora, in: P. Sojka, A. Horák, I. Kopeček, K. Pala (eds.), *Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8–12, 2014. Proceedings*, Springer International Publishing Switzerland, 2014, 257–264.

[4] P. Paikens, L. Rituma, L. Pretkalniņa, Morphological analysis with limited resources: Latvian example, in: S. Oepen, K. Hagen, J. B. Johannessen (eds.), *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA), 2013*, Linköping University Electronic Press, Linköping, 2013, 267–278.

[5] F. J. Och, H. Ney, A systematic comparison of various statistical alignment models. *Computational Linguistics* **1** (2003), 19–51.

[6] M. Škrabal, M. Vavřín, Databáze překladových ekvivalentů Treq. *Časopis pro moderní filologii* **2** (2017), 245–260.

[7] V. Benko, Compatible Sketch Grammars for Comparable Corpora, in: A. Abel, C. Vettori, N. Ralli (eds.): *Proceedings of the XVI EURALEX International Congress: The User in Focus*, Eurac Research, Bolzano, 2014, 417–430.

[8] V. Cvrček, P. Vondřička, *Morfio*. Ústav Českého národního korpusu FF UK, Praha, 2013. [available from http://morfio.korpus.cz]

[9] V. Cvrček, P. Vondřička, Nástroj pro slovotvornou analýzu jazykového korpusu, in: *Gramatika a korpus 2012*. Gaudeamus, Hradec Králové, 2013.

[10] A. J. Zasina, Konkurence koncovek -a a -u v genitivu singuláru neživotných maskulin v polštině, in: M. Stluka, M. Škrabal (eds.), *Lifka a czban – Sborník příspěvků k 70. narozeninám prof. Karla Kučery.* Nakladatelství Lidové noviny, Praha, 2017, 90−98.

[11] A. J. Zasina, M. Škrabal, *Morfio.pl – the possibilities for the application of Czech corpus tools to other languages*, in progress.