Tomáš Káňa, Hana Pelouškova

DIE WICHTIGSTEN ELEKTRONISCHEN KORPORA FÜR TSCHECHISCHE GERMANISTEN

Am Anfang der 60. Jahre begann in ersten kleinen Schritten die Linguistik den Computer als Arbeitsinstrument zu nutzen und bald entstanden die ersten, meist englischen, elektronischen Sprachkorpora. Seit damals haben sowohl die Computertechnologien als auch die Korpuslinguistik einen stürmischen Aufschwung erlebt.
In den 80. Jahren und vor allem in den 90. Jahren entstanden bereits große Korpora und mit ihnen verbundene lexikographische Projekte vieler Sprachen.
Hier möchten wir die wohl wichtigsten Instrumente für die Forschung und Unterricht der deutschen und tschechischen Sprache erwähnen. (Ausführliche Informationen über die einzelnen Korpora sind unter den angegebenen Adressen zu finden.) Näher wird hier das bisher einmalige Instrument der kontrastiven deutsch-tschechischen Sprachforschung vorgestellt: das Tschechisch-deutsche parallele Korpus.

I. DIE GRÖßTEN ELEKTRONISCHEN KORPORA DER DEUTSCHEN SPRACHE

Korpus des Instituts für deutsche Sprache, Mannheim

Im Rahmen des Projektes COSMAS (Corpus Search, Management and Analysis System) entstand in Mannheim das größte elektronische Korpus der deutschen Sprache, bekannt unter dem Namen „Das Mannheimer Korpus“. Es umfasst derzeit (2005) über 1.527 Millionen laufender Wortformen (dies entspricht etwa 3.817.000 Buchseiten).
Der Zugang zu der Recherche im Korpus erfolgt unbürokratisch und kostenlos per Anfrage und nach der Zuteilung des Benutzernamens und Kennworts.
Mehrere Informationen unter: **http://www.ids-mannheim.de/**

DWDS

Das digitale Wörterbuch der deutschen Sprache an der Berlin-Brandenburgischen Akademie der Wissenschaften in Berlin besteht aus einem „Kerncorpus“ und einem „erweitertem Corpus“.

Im „Kerncorpus“ sind etwa 100 Millionen Textwörtern, ausgewogen in der Auswahl der Textsorten und in der Streuung über das 20. Jahrhundert. (So kann der Benutzer die deutsche Sprache in ihren einzelnen Dekaden des 20. Jahrhuns erforschen.)
Das „erweiterte Corpus“ umfaßt neben dem Kerncorpus zusätzlich über eine Miliarde Textwörter, die aus „leicht zugänglichen und digital verfügbaren Texten“[1] besteht.
Die Recherche im DWDS ist mit oder ohne Anmeldung möglich, wobei bei der Recherche mit Anmeldung dem Benutzer der größte Teil des „Kerncorpuses“ zur Verfügung steht (ca. 102 Millionen Wörter aus Texten des ganzen 20. Jahrhunderts), bei der Recherche ohne Anmeldung sind es lediglich ca. 22 Millionen Wörter in Texte aus den Jahren 1900-1945.
Die Anmeldung erfolgt binnen Minuten, ist kostenlos und volkommen unbürokratisch.
Mehr dazu unter: **http://www.ids-mannheim.de/**

Bayerisches Archiv für Sprachsignale:
Das Institut für Phonetik und Sprachliche Kommunikation der Ludwig-Maximilians-Universität München erstellet Kopora der gesprochenen deutschen Sprache. Diese (besonders für Phonetiker interssante) Korpora können derzeit nur als CDs erworben werden. Ihre kurzen Segmente sind aber auch on-line unter **http://www.phonetik.uni-muenchen.de/Bas/BasProjectsdeu.html** zugänglich.

Wortschatz Lexikon
Das Wortschatz Lexikon erstellt das Institut für Informatik der Universität Leipzig. „In der aktuellen Wortschatz-Datenbank stecken 35 Millionen Sätze mit 500 Millionen laufenden Wörtern. Die jüngsten Beispielsätze stammen aus dem Frühjahr 2003. Mehr als 9 Millionen verschiedene Wörter und Wortgruppen können nachgeschlagen werden.“[2]
Die Texte wurden aus dem Internet („öffentlich zugänglichen Quelle“[3]) erhoben.
Das Lexikon ist frei zugänglich unter:
http://wortschatz.informatik.uni-leipzig.de/index.html

II. ELEKTRONISCHE KORPORA IN TSCHECHIEN

[1] http://www.dwds.de/pages/pages_info/dwds_info.htm
[2] http://wortschatz.informatik.uni-leipzig.de/index.html
[3] ibd.

1 Tschechische einsprachige Korpora

Von den Korpora der slawischen Sprachen ist zweifellos das Tschechische Nationalkorpus (ČNK) am umfangreichsten. Dieses Korpus entsteht seit 1994 auf dem Boden der Philosophischen Fakultät der Karlsuniversität in Prag, am Institut des Tschechischen Nationalkorpus unter der Leitung von Professor František Čermák.
Die wichtigsten Bestandteile des Tschechischen Nationalkorpus bilden das synchrone, mittlerweile über 120 Millionen Wörter umfassende Korpus SYN 2000, das diachrone, ca. 2,3 Millionen Wörter umfassende Korpus DIAKORP und zwei Korpora der gesprochenen tschechischen Gegenwartssprache – das ca. 700 000 Wörter umfassende „Prager gesprochene Korpus" und das ca. 500 000 Wörter umfassende „Brünner gesprochene Korpus".
Informationen über das Tschechische Nationalkorpus sind auf der Web-Seite: **http://ucnk.ff.cuni.cz** zu finden.

Außer dem Tschechischen Nationalkorpus gibt es auf dem Gebiet der Tschechischen Republik noch weitere Sprachkorpora, wie z.B. das morphologisch annotierte und fehlerfrei manuell disambiguierte Korpus der publizistischen Texte „DESAM" der Fakultät für Informatik der Masaryk-Universität in Brünn (Näheres unter: **http://www.fi.muni.cz/reports/files/older/FIMU-RS-97-09.pdf)** oder das syntaktisch annotierte Korpus „Prague Dependency Treebank" des Institutes für formale und angewandte Linguistik der Fakultät für Mathematik und Physik der Karlsuniversität in Prag. (Siehe dazu: **http://ufal.mff.cuni.cz/pdt/**)

2. Tschechisch in parallelen Korpora

Die tschechische Sprache ist auch in einigen parallelen Korpora vertreten. Von den mehr- oder vielsprachigen Korpora sind das europäische Projekt „Telri" (Platos Verfassung in 17 Sprachen) und das Projekt „Multext East" (Orwells Roman 1984 in 23 Sprachen) zu nennen.

Seit einigen Jahren existieren auch zweisprachige tschechisch-englische parallele Korpora – am Institut für formale und angewandte Linguistik der Fakultät für Mathematik und Physik der Karlsuniversität in Prag ist ein paralleles Korpus der publizistischen Texte aus der Zeitschrift Reader´s Digest und ihrer tschechischen Fassung Výběr entstanden. An der Philosophischen Fakultät der Masaryk-Universität in Brünn wurde Ende der

Neuzigerjahre ein literarisches, tschechisch-englisches Korpus „Kačenka“ erstellt. (**http://www.phil.muni.cz/angl/kacenka/kachna.html**)

Aufgrund der Tatsache, dass es bisher kein tschechisch-deutsches paralleles Korpus gab und dass sich seine Entstehung zum Zweck der kontrastiven Sprachforschung und des Fremdsprachenlernens als sehr dringend erwies, initiierte der Lehrstuhl für deutsche Sprache und Literatur der Pädagogischen Fakultät der Masaryk-Universität in Brünn ein Projekt, das seit 2001 läuft.

II. DAS TSCHECHISCH-DEUTSCHE PARALLELE KORPUS

Das Tschechisch-deutsche parallele Korpus (CNPK) besteht aus zwei einsprachigen Korpora, die miteinander verknüpft sind. Jede Parallele hat demnach die Standardeigenschaften eines einsprachigen Korpus.

Der Korpusmanager heißt “Manatee“, die Benutzerapplikation „Bonito“ (mehr dazu unter: **www.textforge.cz**). Unter dem gleichen Manager laufen auch „Das Tschechische Nationalkorpus“ und das „Slowakische Nationalkorpus“, wodurch die Korpusnutzung all jenen erleichtert wird, die es gewohnt sind mit den Nationalkorpora zu arbeiten.

1. Geschichte der Entstehung des Tschechisch-deutschen parallelen Korpus

Die Korpuserstellung ist ein langfristiger komplizierter Prozess, der zahlreiche Tätigkeiten einbezieht: Suche nach passenden Texten und ihren Parallelen, Elektronisierung (Einscannen) der Texte, Segmentierung langer Absätze, Katalogisierung der Texte, Export der Texte und vieles anderes mehr. Darüber hinaus erfordert die Erstellung und die Funktions- und Qualitätssicherung des Korpus Fachwissen im Bereich der Informatik und eine entsprechende technische Infrastruktur. Eine Zusammenarbeit mehrere Institute ist demnach unumgänglich.

Das Tschechisch-deutsche parallele Korpus wurde ab 1999 geplant und entsteht seit 2001 am Lehrstuhl für Deutsche Sprache und Literatur der Masaryk-Universität in Brünn. In beträchtlichem Maße sind bei der Erstellung des Korpus die Studenten dieses Lehrstuhls behilflich. Meistens führen sie technische Arbeiten durch: Scannen, „Reinigung“ der Texte, manuelles Alignment.

Ein wichtiger Partner, Berater und Unterstützer des Parallelen Korpus ist das Institut des Tschechischen Nationalkorpus an der Philosophischen Fakultät der Karlsuniversität, Prag.

Die technische Unterstützung und Bearbeitung, sowie die technische Verwaltung und Softwareentwicklung gewährleistet der Lehrstuhl für Informationstechnologien der Fakultät für Informatik der Masaryk-Universität, Brünn, neuerlich in Zusammenarbeit mit der Technischen Universität, Brünn.
Im Rahmen des Projektprogramms „Aktion Österreich – Tschechische Republik" beteiligte sich an der Korpuserstellung auch das Institut für Slawistik der Universität Wien.

2. Parameter des Korpus

2.1 Umfang

Zum 30.6.2005 beinhaltet das Korpus mehr als 3,2 Millionen Wörter im tschechischen Teil. Gemeinsam mit der deutschen Parallele sind es knapp über 6,8 Millionen Textwörter. Allein der Vergleich des Wortumfangs in beiden Parallelen führt zu einer interessanten Feststellung: der deutsche Text ist immer um ca. 15-20% länger als der tschechische, ungeachtet dessen, welche von den beiden Sprachen die Ausgangssprache war.
Für eine gängige linguistische Forschung (und die allgemeine Lexikographie) sollte das Korpus mindestens 5 Millionen Wörter im tschechischen Teil enthalten. Mit diesem Umfang wäre es jedoch nach wie vor zu klein für die Forschung auf einem Spezialgebiet und für spezialisierte Lexikographie.

2.2. Katalogisierung

Jeden in das Korpus eingegliederten Text begleitet die sog. äußere Annotation. Sie enthält möglichst detaillierte Angaben über den Text und seinen Ursprung. Im Einzelnen sind es folgende Informationen:

- Titel des Textes
- Name, Vorname und Geschlecht des Autors/ der Autorin, ggfs. aller Autoren
- bei Übersetzungen: Name, Vorname und Geschlecht des Übersetzers
- Verlagsangaben, ggfs. Internetseite
- Erscheinungsjahr des Originaltextes und der Übersetzung
- (grobe) stilistische Zuordnung: Belletristik, Fachtexte, Publizistik (siehe Tabelle oben)
- Themenbereich innerhalb der jeweiligen Stilebene (siehe Tabelle oben)

- Medium, von dem der Text gewonnen wurde: z.B. eingescannte Drucksache, Internet usw.
- Sprache des Textes
- Ausgangssprache (Originalsprache des Textes)
- Informationen über die Verarbeitung des Textes: welche Eingriffe notwendig waren für die „Reinigung" des Textes, wie z. B. Auslassung der Tabellen, Bilder, Seitenzahlen.
- Name der Person, die den Text für die Eingliederung in das Korpus vorbereitet hat.
- Datum der Bearbeitung

Diese Informationen ermöglichen Subkorpora nach konkreten Kriterien zu bilden (z. B. nur diejenigen Texte auszuwählen, die nach 2000 herausgegeben worden sind und von Frauen verfasst wurden). Dieselben Informationen sind auch im Katalog der Texte (.xls-Tabelle) enthalten. Alle Texte werden im Dokumentformat (.doc., .rtf. oder .txt.) auf CDs gespeichert und archiviert.

2.3 Korpustexte

2.3.1 Textsprache

Es handelt sich um ein **synchrones Sprachkorpus**. Alle Texte im Korpus wurden nach dem Jahr 1920 veröffentlicht. Der größere Teil von ihnen sogar erst nach 1950.
Bisher konnte auch die Prämisse der „Zweisprachigkeit" angehalten werden. Das bedeutet, dass eine der Parallelen ein Originaltext sein muss und die andere die Übersetzung gerade dieses Textes. Übersetzungen aus einer dritten Sprache konnten wir noch vermeiden.
Für das Gleichgewicht der beiden Sprachen soll auch gesorgt werden. Momentan überwiegen tschechische Originaltexte (63%).

2.3.2 Textsorten

Das Ziel des Vorhabens war und ist ein **allgemeines Korpus** zu erstellen. Da sich parallele Korpora ausschließlich auf geschriebene Texte konzentrieren müssen, wurde nach mehreren Beratungen ein Optimum für die Ausgewogenheit der funktionalen Textstile festgelegt: 50% Belletristik, 25% Publizistik und 25% Fachtexte.

Diese Stilebenen wurden dann in Themenbereiche aufgeteilt, die innerhalb der Stilebenen auch möglichst proportional vertreten werden sollen. Die folgende Tabelle stellt die Proportionalität der Texte im jetzigen Zustand dar. In Klammer werden Abkürzungen angeführt, die im Korpus als Attribute für die Sortierung der Konkordanzen oder Subkorpuserstellung dienen:

Belletristik (lit)	**63%**	
davon	Drama (dra)	9%
	Non-fiction (nfi)	4%
	Roman (nov)	87%
Publizistik (pub)	**17%**	
davon	Werbematerialien im weiteren Sinne (add)	3%
	Graue Zone zu Belletristik – Essay, Feuilleton... (ess)	0,5%
	informative Texte (tin)	96,5%
Fachtexte (sci)	**20%**	
davon	Administrative (adm)	3%
	Kunstwiss. (art)	2%
	Biologie (bio)	8%
	Handwerk (crf)	0%
	Ökonomie (eco)	1%
	Geschichte (his)	37%
	Chemie (che)	0,5%
	Industrie (ind)	2%
	Rechtswiss. (jus)	3%
	Philologie (lin)	0,5%
	Pädagogik (ped)	8%
	Religion (rel)	15%
	Soziologie u.a. (soc)	20%

2.4 Zusätzliche technische Eigenschaften

Alignment - das Korpus ist zur Gänze auf eine Länge des Satzes, höchstens eines kurzen Absatzes, manuell aligned (die entsprechenden Textpassagen sind einander zugeordnet). Automatisches Alignment hat sich in der Anfangsphase der Korpuserstellung nicht bewährt. Alignment zeigt sich so als der zeitlich anspruchvollste Schritt bei der Erweiterung des Korpus.

Lemmatisierung – Beide Parallelen sind automatisch lemmatisiert (den lexikalischen Grundformen zugeordnet).

Disambiguierung – Beide Parallelen sind bisher nur automatisch disambiguiert (zu ca. 10% müssen die Zuordnungen zu den richtigen Kategorien noch manuell angepasst werden).

Tagging – Im Korpus aktuell nur die tschechische Parallele automatisch getaggt (ermöglicht die Abfrage nach morphologischen Kategorien). Im deutschen Teil wurde mithilfe eines einfachen Taggers jedem Wort die Wortart zugeordnet.

3. Nutzung des Korpus

Zurzeit befindet sich das Korpus auf dem Server der Fakultät für Informatik in Brünn. Der Zugang zu ihm ist aus urheberrechtlichen Gründen nur autorisierten Benutzern möglich. Die Autorisierung kann nach Anfrage erteilt werden, allerdings ausschließlich für nicht kommerzielle Zwecke. (Kontaktpersonen sind die Verfasser dieses Artikels.)
Das Korpus ist in Probebetrieb und dient vor allem den Mitarbeitern und Studenten der Pädagogischen Fakultät als Quelle für authentisches Belegmaterial für Jahres-, Abschluss-, Magister- und Dissertationsarbeiten.
Bisher wurden folgende Themenbereiche erforscht:

- Der freie Dativ im Tschechischen und seine Äquivalente im Deutschen (Dissertationsarbeit)
- Die syntaktische und semantische Analyse der deutschen und tschechischen Präpositionen (Serie einiger Magister- und Jahresarbeiten, die sich immer monothematisch mit einer Präposition befassen. Bis jetzt wurden Präpositionalgefüge mit den Präpositionen *an, auf, für, durch, bei, von, um* und ihre tschechischen Äquivalente analysiert)
- Infinitivkonstruktionen als Transformationen der deutschen Nebensätze mit der Konjunktion *dass* und deren äquivalente Strukturen im Tschechischen (Magisterarbeit)
- Das Pronomen „*es*“, seine syntaktischen Funktionen und Äquivalente im Tschechischen (Magisterarbeit)
- Korrelate zu Ergänzungssätzen im Deutschen und im Tschechischen (Magisterarbeit)
- Übersetzung von Okkasionalismen im Werk „Fimfarum“ (Magisterarbeit)
- Die Frequenz der deutschen und tschechischen Satzbaupläne (Abschlussarbeit)
- Gründe für die unterschiedliche Länge der deutschen und tschechischen Texte (Abschlussarbeit)
- Der Ausdruck der Vorzeitigkeit in den deutschen und tschechischen Temporalsätzen (Jahresarbeit)
- Die Stellung der Partikeln im Deutschen und im Tschechischen (Jahresarbeit)

- Das Subjekt im Deutschen und im Tschechischen (Jahresarbeit)
- Ortsnamen und ihre Äquivalente (Jahresarbeit)

Weiterhin wird das Korpus als Quelle für diverse Lehr- und Studienmaterialien verwendet. Bisher schöpften daraus v.a. die linguistischen Fächer: v.a. Syntax, Morphologie, Lexikologie und Übersetzungsseminare. Es entstanden bereits einige Skripten (Syntax, Übersetzungsseminare), mehrere sind in Vorbereitung (Kontrastive Wortbildung, Lexikologie, Textlinguistik) oder in Planung (Morphologie, Stilistik).
Das Korpus wird nun von einigen wenigen Benutzern als ein individuelles Lehrmaterial verwendet. Um die Möglichkeiten der Korpusnutzung effektiver publik zu machen, wird im nächsten akademischen Jahr als Wahlfach die Einführung in die Korpuslinguistik angeboten.

4 Beispiele der Arbeit mit dem Korpus

Die Wahl der im Korpus vertretenen Sprache (Parallele) erfolgt durch eine Ikone rechts auf dem Bildschirm („cnpkz“ für die tschechische Parallele, „cnpkde“ für die deutsche Parallele).
In das Abfragefenster „New query/Nový dotaz“ wird die gesuchte Erscheinung geschrieben.

4.1 Abfrage

Prinzipiell hat der Benutzer drei Grundabfragen zur Wahl: **Form** (Wortform, Buchstabenkombination oder Position), **Lemma** und **Tag**.

4.1.1 Form

Die Abfrage nach der gesuchten Form erfolgt durch einfaches Eintippen in das Abfragefenster:

Wortform

Für die Abfrage einer einfachen Wortform sind keine Sonderzeichen notwendig. Hier das Beispiel der Abfrage „lump“ im Tschechischen und „Lump“ im Deutschen:

```
#------------------------------------------------------------
# Corpus  : cnpkcz
# Query   : lump
```

```
#------------------------------------------------------------

1:             Pak zmizel , <lump> . Nevěděl si asi
2:     bude asi také pěknej <lump> . ´ Musel skočit
3:         , mazanej jste , <lump> jste , uličník ,
4:  zas budeš stejně takový <lump> jako předtím . To
5:    z tebe nakonec stejný <lump> jako tvůj táta .
6:              v boj jde , <lump> si může hopsat .
7:             na zemi , je <lump> , kdo zůstal v
8:     chtíc nechtíc , když <lump> se zlatem zpil .
9:           mi přeje ; ten <lump> dole mě začíná mít
10:    - urostlý , navoněný <lump> z " lepší společnosti
11:         ten syčák , ten <lump> ! " vykřikl do
12:            . Ale byl to <lump> . Ein Halunke .cnpkde: Lump

#------------------------------------------------------------
# Corpus  : cnpkde
# Query   : lump
#------------------------------------------------------------

1:            " Riech einmal , <Lump> ! " Schwejk roch
2:      gebührt , verstanden , <Lump> ? Das ist der
3:     daß du ein ausgegorener <Lump> bist . " "
4:         verschwand er , der <Lump> . Er wußte sich
5:               Nase zu , der <Lump> . Ich habe noch
6:                 vor . " Der <Lump> ! " sagte K
7:       Käsemadrig hält , der <Lump> , der unverschämte !
8:              was , wenn der <Lump> dich umbringt ? Was
9:          Siehst du , dieser <Lump> hat wenigstens das Holz
10:             siehst du , du <Lump> , so gehts bei
11:             du hinten , du <Lump> , schneuz dich nicht
12:              " Gesteh , du <Lump> , daß do nur
13:              Na warte , du <Lump> , du Gauner !
14:        Wasser kriegen , du <Lump> , du ! "
15:              " Du bist ein <Lump> , Tomesch ! "
16:          . Tomesch ist ein <Lump> , stellte Prokop mit
17:             sind Sie , ein <Lump> sind Sie , ein
18:         wieder grad so ein <Lump> sein wie vorher .
19:            Aber er war ein <Lump> . Ein Halunke .
20:      gewiß auch ein feiner <Lump> sein . ' Er
21:        am Ende der gleiche <Lump> werden wie dein Papa
22:seiner Schandschnauze , der <Lump> " , brummt er
23:       So nicht ! Irgendein <Lump> hat uns die Mühle
            , ohne auch nur <Lump> , Trottel oder Saukerl
```

Position

Genauso einfach ist die Abfrage der „Position“ (Token).

Die Abfrage „!“ (Rufzeichen) hat ergeben, dass dieses in der deutschen Parallele häufiger vorkommt - 16.473 Mal, hingegen in der tschechischen – „nur“ 15.780 Mal.

```
#------------------------------------------------------------
# Korpus  : cnpkcz
# Query   : "!"
#------------------------------------------------------------
    jen na vašem odhodlání a odvaze <!> Jak vypadá program na vysokých
#------------------------------------------------------------
# Korpus  : cnpkde
# Query   : "!"
#------------------------------------------------------------
Entschlossenheit und ihren Mut an <!> Wie sieht ein Hochseilprogramm aus
?
```

Der Grund für diese Diskrepanz soll in einer eigenen Untersuchung thematisiert werden.

Hier bringen wir vier ad-hoc ausgewählte Parallelen mit unterschiedlichen Satzzeichensetzungen (diese Stellen sind unterstrichen):

```
cnpkde:
" Dir werde ich noch Keime zeigen ! Dir werde ich einen Schoß zeigen !
cnpkcz:
Já ti dám zárodky ! Já ti dám lůno .

cnpkde:
" Schnell , Abe , lauf , lauf ! " flüsterte Li .
cnpkcz:
" Honem , Abe , běž , běž , " šeptala Li .

cnpkde:
Dann muß ihm Tomesch Namen und Wohnung des Mädchens nennen und sich
verpflichten – nein : nur keine Versprechen von diesem Schuft !
cnpkcz:
Dále , mám ho jednoduše v hrsti : musí mně říci jméno a pobyt toho
děvčete a zavázat se – ne ; žádné sliby od takového ničemy .

cnpkde:
" Frau Oberköchin ! Frau Oberköchin ! " mahnte der Oberkellner , der
ihren Blick aufgefangen hatte .
cnpkcz: 1332141--1332160
" Paní vrchní kuchařko ! Paní vrchní kuchařko , " napomenul ji vrchní
číšník , který zachytil její pohled .
```

Buchstabenkombination

Für die Abfrage der Buchstabenkombination als Teil eines Wortes sind Sonderzeichen notwendig. Im Unterschied zum Mannheimer Korpus bedeutet „*“ *beliebige Zeichen* und „.“ *beliebige Anzahl der Zeichen*. Hier wurde nach der Erscheinung des tschechischen Diminutivsuffix „-íček“ gesucht.

Aus dem Kopf der Konkordanzen sind alle Schritte der Abfrage ersichtlich.

```
#-------------------------------------------------------------
# Corpus  : cnpkcz
# Label   :
# Query   : .*íček
# Expanded: ".*íček"
# Steps:
#   > Query   : ".*íček"
#   > Deleted : 400
# Size    : 9
# Context : 4 Positions to left, 4 Positions to right
# User    : Kana
# Date    : Wed, 19 Jan 2005 02:29PM
#-------------------------------------------------------------
            . zde stál malý <kostelíček> s věží a zvonem
          tuto odpověď za " <políček> " . Vládě Spojených
připomíná dodnes . Neslušný <mužíček> Když se na jižní
```

```
            si na blůzce o <knoflíček> víc nebo si významně
   Dášeňka už není bezmocný <uzlíček> s třesoucím se ocáskem
    tlamičce na patře černý <flíček> od čerta . Tak
             kufru , pane , <klíček> zavěsit pod košili ,
             neboj se , jsi <chlapíček> . Co stříháš ušima
```

Und hier die Abfrage des deutschen: Präfix „zer-“.

```
------------------------------------------------------------
# Corpus  : cnpkde
# Label   :
# Query   : zer.*
# Expanded: "zer.*"
# Steps:
#   > Query   : "zer.*"
#   > Deleted : 1348
# Size    : 5
# Context : 4 Positions to left, 4 Positions to right
# User    : Kana
# Date    : Wed, 19 Jan 2005 02:21PM
#------------------------------------------------------------
Reich der Luxemburger längst <zerfallen> ist und die politischen
    mit seinem Bruder Prokop <zerstritt> verwies er die brandenburgischen
     ihn mit Waffengewalt zu <zerschlagen> . Die öffentliche Meinung
   seit 1138 in Fürstentümer <zerfallen> war konnte der aus
   geschrieben das Reich sei <zerrissen> und verfallen und bedürfe
```

4.1.2 Lemma

Beide Parallelen des Korpus sind automatisch lemmatisiert. Die Syntax der Abfrage ist im Vergleich zum „Wort“, bzw. „Wortform“ komplizierter: [lemma="lump"]

```
#------------------------------------------------------------
# Corpus  : cnpkcz
# Query   : [lemma="lump"]
#------------------------------------------------------------
pomáhali Čechům vyřezat pár <lumpů> i u nás .
             . Pak zmizel , <lump> . Nevěděl si asi
       toho hodnýho od toho <lumpa> , zejména dnes ,
              " Čichni si , <lumpe> ! " Švejk si
           mít kázání k těm <lumpům> . V tomhle případě
   držkovala jako na ňákýho <lumpa> , když tady před
          , že dívka miluje <lumpa> a halunka , i
            ve slovníku . " <Lumpi> , " řekl tatínek
#------------------------------------------------------------
# Corpus  : cnpkde
# Label   :
# Query   : [lemma="Lump"]
# Expanded: [lemma="Lump"]
#------------------------------------------------------------
             helfen , ein paar <Lumpen> bei uns zu schlagen
   seiner Schandschnauze , der <Lump> " , brummt er
                 tu ich , ihr <Lumpen> . Ich verlier Zeit
               " Gesteh , du <Lump> , daß do nur
             Hause gehn , ihr <Lumpen> , es ist schon
```

4.1.3 Tag

In seiner On-Line Version ist das Korpus (nur) im tschechischen Teil automatisch getagged, jedoch noch nicht disambiguiert. Für die komplizierte Syntax der Abfrage ist ein Verzeichnis der morphologischen Zeichen notwendig. Dieses wird im vorbereiteten Korpusmanual in absehbarer Zeit publiziert.

Hier wird nach Adjektiv (k2) in Genitiv Singular und Plural (c2) gefragt:

```
#------------------------------------------------------------
Corpus  : cnpkde
# Label   :
# Query   : [tag="k2.*c2"]
# Expanded: [tag="k2.*c2"]
# Steps:
#   > Query   : [tag="k2.*c2"]
#   > Deleted : 2175
# Size    : 7
#------------------------------------------------------------
exkurze mají být během svého <pobytu> doprovázeni jedním či více
  jejího prasynovce Antonína <Burgundského>. Tím vymřela vedlejší
    kdo udává tón v produkci <šampaňského> vína . Virginie
   jeden dlouhý ocásek , pár <uší> od ohaře , čtyři malé nožičky
        synem , jenž byl tak <malého> vzrůstu , že jsem si myslel
  parková úprava , zejména u <starého> městského příkopu . Proto
           krásných iluzí a <lepších> vizí do nového roku Vám přejí
```

4.2 Kontexteinstellung

Der Kontext der Treffer in den Konkordanzzeilen kann beliebig erweitert oder verengt werden. Das Kriterium für die Kontexteinstellung ist die Anzahl der Positionen oder der Zeichen vor und nach dem Treffer. Die parallelen Textpassagen sind jedoch durch Alignment fix eingestellt. (Siehe weiter unter 4.4 Parallele Textabschnitte.)

4.3 Sortierungskriterien

Die Konkordanzzeilen können wir nach verschiedenen Kriterien sortieren: z.B. nach dem Stil des Textes, Jahrgang der Erscheinung, Geschlecht des Verfassers oder Übersetzers. Alle diese Kriterien beinhalten die sog. „Metainformationen“.

Darüber hinaus kann das Ergebnis nach Häufigkeiten, Zufallstreffer und anderen Kriterien eingeschränkt werden.

4.4 Parallele Textabschnitte

Zu jeder Konkordanzzeile (ungeachtet dessen, ob es sich um die Abfrage eines Wortes, Wortteiles, bzw. Position, Lemmas oder Tags handelt) können wir auch die entsprechende Textpassage in der anderen Sprache (in der oppositen Parallele) abrufen. Dies erfolgt

durch Drücken der Taste „F9“ und die sich entsprechenden Texte erscheinen in einem neuen Fenster.

Die Länge der parallelen Textpassagen kann noch sehr unterschiedlich sein. Der Grund dafür sind die bisher unzuverlässigen und unvollkommenen Aligner. Bittere Erfahrungen haben uns gelehrt alle Texte manuell zu alignen. So wird bei der Vorbereitung der Texte für das Korpus der „minimale Kontext“ beachtet - es heißt, dass alle Texte auf der Satzebene aligned sind. (Maßgebend ist der längere Satz in einer der Parallelen.) In Dramawerken und in Texten mit direkter Rede, wo die Sätze sehr kurz sind, werden mehrere Repliken zum „Minimalkontext“.

Als Beispiel suchten wir das deutsche Lexem „auseinandersetzen“ aus (hier nur die Infinitivform), für das die tschechische Sprache keine direkte lexikalische Entsprechung hat. Das Verb (das gesuchte Wort) im deutschen und seine funktionalen Äquivalente sind unterstrichen.

```
cnpkde:
Nach seiner allgemeinen Anerkennung als Römischer König und dem
Friedensschluß mit König Wladislaw von Polen und dem " Schiedsspruch vom
August 1412 konnte König Sigismund sich endlich mit dem seit Jahren
schwelenden Konflikt mit der Republik Venedig auseinandersetzen .
cnpkcz:
Poté co byl všeobecně uznán římsko - německým králem , uzavřel mír s
polským králem Vladislavem a v srpnu 1412 vyhlásil " budínský smírčí
rozsudek " , mohl se král Zikmund konečně vypořádat s letitým konfliktem
, který mìl s benátskou republikou .
--------------------
cnpkde:
" Der leeren Tasche inszenierte " gehört nicht nur zu den wenigen
erfolgreichen Kriegen des Luxemburgers sondern war auch ein
Überraschungscoup der blitzartig zum Erfolg führte bevor die Fürsten
sich eingehend mit der Problematik auseinandersetzen konnten .
cnpkcz:
" S prázdnou kapsou " , je jednou z mála úspěšných válek tohoto
Lucemburka . Překvapila bleskovým úspěšným úderem , dříve než se knížata
mohla s problematikou podrobně obeznámit .
--------------------
cnpkde:
Bedenken Sie , daß ich von Plato ausgegangen bin , mich durch die
Skepsis Humes , den Subjektivismus Kants durchgearbeitet , bei Comte ,
Mill , Brentano und so vielen anderen gelernt habe - wie viele noetische
Fragen gibt es da , mit denen ich mich habe auseinandersetzen müssen !
Also ganz kurz gesagt :
cnpkcz:
. Považte , že jsem rostl z Platóna , prokousal se skepsí Humovou ,
subjektivismem Kantovým , učil se u Comta , Milla , Brentana a u tolika
jiných - co tu je noetických otázek , které jsem si musel vyřídit ! Tož
docela stručně :
--------------------
cnpkde:
PLZAK : Kunst - damit müssen wir uns auseinandersetzen !
```

```
cnpkcz: 1046957--1046966
PLZÁK : Umění - tomu říkám slovo do pranice !
--------------------
cnpkde:
Nach Abschluß dieses ersten Grundkurses sind 5 Seminare geplant, die
sich mit den methodischen Fragen des differenzierten Unterrichts
auseinandersetzen , danach eventuell weitere follow - ups .
cnpkcz: 1586810--1586838
Po ukončení tohoto prvního základního kurzu je naplánováno pět seminářů
, které se budou zabývat metodickými otázkami diferencovaného vyučování
. , potom budou eventuelně následovat follow - ups .
--------------------
cnpkde:
In den Bereich fallen des Weiteren auch z.B. Seminare oder Publikationen
und Filme , die sich mit dem Fragenkomplex von Geschichte und Kultur der
Minderheiten und ihrer Identitätsbildung auseinandersetzen .
cnpkcz:
Do této oblasti patří dále semináře , publikace a filmy , které se
zabývají otázkami kultury a dějin menšin a jejich hledáním identity ve
společnosti.
```

5. Ausblick

Das Korpus wird nach wie vor erweitert und präzisiert werden. Der Umfang sollte bis 2006 auf fünf Millionen Wörter im tschechischen Teil steigen, wobei die Stilebenen entsprechend der gewünschten Parameter ausgewogen werden sollen. In weiterer Zukunft wird zudem mit manueller Disambiguierung beider Teile und mit dem Tagging der deutschen Parallele gerechnet.

Ein Teil des Tschechisch-deutschen parallelen Korpus wird ein Basisbestandteil des multilingualen Korpus „Intercorp“, das im Rahmen eines Forschungsvorhabens am Institut des tschechischen Nationalkorpus an der Karlsuniversität entsteht, und an dessen Erstellung auch die Autoren dieses Artikels als Partner mitwirken.

In absehbarer Zeit sollen auch alle Texte, die nicht urheberrechtlich geschützt sind, als paralleles tschechisch-deutsches Korpus dem breiten Publikum zugänglich gemacht werden. Gleichzeitig mit diesem Vorhaben soll auch das Korpusmanual herausgegeben werden.

Ein wichtiger Punkt, der ebenfalls bald gelöst werden wird, ist die Anpassung des Korpusmanagers für Sehbehinderte, die nicht imstande sind, mit der Maus zu arbeiten.

Eine weitere Priorität für die nähere Zukunft ist die Zusammenarbeit mit Partnern aus den deutschsprachigen Ländern (Slawisten/Bohemisten oder DaF-Germanisten), die sich an diesem Projekt (und anschließenden Projekten) beteiligen möchten.
Die Perspektiven einer möglichen Zusammenarbeit wären im Bereich der Erweiterung des Korpus, der kontrastiven Sprachforschung, in den lexikographischen Projekten und in der Implementierung des Korpus in das Fremdsprachenlernen z.B. in einer korpusunterstützten kontrastiven Lernergrammatik.

Literaturquellen

Atkinsová, B.T.S. - Clear, J. - Ostler, N.: Kritéria pro výstavbu korpusu. In: *Studie z korpusové lingvistiky*. Praha: Nakladatelství Karolinum, 2000. S.75-106.

Biber, D.: Reprezentativnost v projektu korpusu. In: *Studie z korpusové lingvistiky*. Praha: Nakladatelství Karolinum, 2000. S.107-136.

Čermák, F.: Jazykový korpus: Prostředek a zdroj poznání. In: *Slovo a Slovesnost*, 56, 1995. S.119-140.

Hlavičková, V.: K uplatnění počítačů v práci učitele cizích jazyků. In: *Cizí jazyky* roč. 38. 1993-94. S. 198.

Chafe, W.: Význam korpusové lingvistiky pro pochopení podstaty jazyka. In: *Studie z korpusové lingvistiky.* Praha: Nakladatelství Karolinum, 2000. S. 57-71.

Čermák, F. et al: *Český národní korpus. Úvod a příručka uživatele.* Praha: FF UK – ÚČNK, 2000.

Leech, G.: Korpusová lingvistika-současný stav oboru. In: *Studie z korpusové lingvistiky*. Praha: Nakladatelství Karolinum, 2000. S. 39-56.

Pelouškova, H. - Káňa, T.: Česko-německý paralelní korpus. In: *Teoretické východiská a perspektívy vyučovania cj.* Bratislava: Retaas, s r.o., 2004.

Pelouškova, H. - Káňa, T.: Das tschechisch-deutsche parallele Korpus als effektives Mittel in der Sprachforschung und im Fremdsprachenunterricht. In: *Könniggrätzer Linguistik - und Literaturtage*. Hradec Králové: Gaudeamus, 2003. S. 108-118.

Pelouškova, H. - Káňa, T.: Paralelení korpus jako zdroj autentického jazykového materiálu pro výzkum i výuku jazyků. In: *Cizí jazyky*, Plzeň: Fraus, 2002. S. 43-45.

Rychlý, P.: *Korpusové manažery a jejich efektivní implementace*. Disertationsarbeit. Brno 2000.

Štanglová, M.: CALL-CUU. Co vše se za tím skrývá? In: *Cizí jazyky* roč. 39. 1995-1996. S. 17-18.

Šulc, M.: *Korpusová lingvistika*. Praha: Nakladatelství Karolinum, 1999.

Ungermannová, M.: *Manuál pro značkování a desambiguaci slovních tvarů v korpusu DESAM.* Interní materiál FI MU. Brno 1999.

http://www.texforge.cz

Korpora im Internet:

http://www.ucnk.ff.cuni.cz

http://www.ids-mannheim.de/

http://www.phonetik.uni-muenchen.de/Bas/BasProjectsdeu.html zugänglich.

http://wortschatz.informatik.uni-leipzig.de/index.html

http://ucnk.ff.cuni.cz

http://ufal.mff.cuni.cz/pdt/

http://www.phil.muni.cz/angl/kacenka/kachna.html

Dieser Beitrag entstand im Rahmen des Forschungsvorhabens MSM 0021620823.

IN: Kolektiv autorů**:** *Brünner Hefte für Germanistik*. Masarykova Univerzita v Brně, 2005.