SBORNÍK PRACÍ FILOZOFICKÉ FAKULTY BRNĚNSKÉ UNIVERZITY
STUDIA MINORA FACULTATIS PHILOSOPHICAE UNIVERSITATIS BRUNENSIS
A 44, 1996

KLÁRA OSOLSOBĚ

FORMALE BESCHREIBUNG DER TSCHECHISCHEN MORPHOLOGIE UND IHRE ANWENDUNG AUF DEM GEBIET DER LINGUISTISCHEN DATENVERARBEITUNG

1. Einleitung

Dieser Artikel will die Ergebnisse der Arbeit auf dem Gebiet der linguistischen Datenverarbeitung präsentieren, vor allem die Ergebnisse der automatischen morphologischen Analyse und Synthese des Tschechischen und der automatischen Lemmatisierung, die im Rahmen der Untersuchungen des Kabinets für Computerlinguistik der Masaryk-Universität in Brno erzielt wurden.

In den slawischen Sprachen spielt die formale Morphologie eine wichtige Rolle. Die automatische Generierung und Interpretation der Wortformen ist die erste Stufe bei der Analyse der natürlichen Sprache. Eine dynamische Beschreibung der Formalmorphologie wirft ein neues Licht auf die bisherigen in den klassischen Grammatiken und Wörterbüchern benutzten Klassifikationen. Die morphologische Analyse ist der erste Schritt zur Lemmatisierung.

Einzelne Aufgaben bei der Lösung der gegebenen Problematik:

1. Algoritmische Beschreibung der tschechischen Morphologie
2. Erstellung eines repräsentativen Computer-Wörterbuchs des Tschechischen
3. Erstellung von Programmen, die die so gewonnene linguistische Basis bearbeiten können:
   a. Automatischer Korrektor (Spelling-checker)
   b. Automatischer Lemmatisator
   c. Automatischer morphologischer Analysator-Generator
   d. Automatischer Tagger

2. Algorithmische Beschreibung der tschechischen Flexionsmorphologie

Bei der formalen Beschreibung der tschechischen Flexionsmorphologie findet ein Modell morphologischer Analyse Anwendung, das diese als Prozeß einer zweistufigen Segmentierung der Wortform in drei formal genau definierbare Segmente sowie derer Identifizierung versteht.

Zum Ausgangspunkt für die algorithmische Beschreibung der tschechischen Flexionsmorphologie wurden die klassischen in den Grammatiken (Havránek, Jedlička, MČ 2) eingeführten Klassifikationen. Die Überführung solcher Beschreibungen in eine dynamische Form erforderte vor allem eine ausgedehnte Subklassifikation der klassischen Flexionsmuster, sowie eine detaillierte Analyse des Systems der Flexionsendungen bei den flektierbaren Wortarten. Zur Flexion wurden auch die folgenden Ableitungstypen gerechnet:

1. Adverbien, die paradigmatisch von Adjektiven abgeleitet werden;
2. possessive Adjektive, die von belebten Maskulina und Feminina abgeleitet werden;
3. Substantivierung und Adjektivierung von Verbalpartizipien;
4. komplexe Beschreibung der Deklination sowie der Ableitung der einzelnen Typen von Numeralien (vgl. Osolsobě, 1995).

Die Beschreibung der tschechischen Flexionsendungen und Flexionsmuster bildet eine geöffnete Struktur, die auf die Beschreibung von anderen Sprachen bzw. von Stadien oder Schichten einer Sprache transponiert werden kann (vgl. die praktischen Ergebnisse).

Segmentierung des Wortes für die Computerbeschreibung der natürlichen Sprache

Bei der Computeranalyse wird die Wortform in zwei Phasen in drei Teile gegliedert. In der ersten Phase wird vom Ende des Wortes die Flexionsendung (**T**) abgetrennt. Das übriggebliebene Segment, das wir als Stamm (**KM**) bezeichnen, wird weiter in zwei Teile gegliedert, und zwar das sgn. Intersegment (**IS**) und die Stammbasis (**KMZ**).

Die einzelnen Komponente der formalen Beschreibung der tschechischen Flexion

Wenn wir im weiteren von der Endung (**T**) sprechen werden, ist bei Substantiven, Adjektiven und Verben die in der Grammatik definierte Flexionsendung gemeint, die die einzelnen Bedeutungen der entsprechenden grammatischen Kategorien für die entsprechende Wortart trägt. Bei den unbestimmten Verbformen wird für die Endung (**T**) der Komplex aus ableitendem Suffix des Partizips und die Genusendung angesehen. Als Endung gilt auch die Nullendung. Die Nullendung (-0) schließt in unserer Fassung sowohl die Fälle mit unverändertem als auch mit verändertem Stamm in Kombination mit der Nullendung ein.

In der zweiten Phase wird der Stamm (**KM**) weiter in zwei Teile zerlegt. Sie werden Stammbasis (**KMZ**) und Intersegment (**IS**) genannt. Als Intersegment (**IS**) wird die Finalgruppe des Stammes (**FSK**) bezeichnet, die während der Flexion Änderungen erfährt.

Das Ziel der algorithmischen Beschreibung der tschechischen Formalmorphologie war es, eine dynamische Beschreibung von Deklination und Konjugation zu schaffen. Im Laufe der Arbeit wurden einige Fälle entdeckt, die sich auf der

Grenze zwischen Formalmorphologie und Wortableitung befinden. Es handelt sich um die oben genannten Typen (die paradigmatische Ableitung der Adverbien von Adjektiven und die Steigerung von beiden, die Ableitung der possessiven Adjektive von belebten Maskulina und Feminina, die Ableitung der adjektivisierten Partizipien von Verben, die komplexe Beschreibung der Deklination und die Ableitung der einzelnen Typen von Numeralien).

In dieser Beschreibung werden also nicht nur die Formalmorphologie sondern auch ausgewählte Typen der Wortbildung einbezogen. Der in Stammbasis (**KMZ**) und Intersegment (**IS**) zerlegte Stamm (**KM**) wird formbildende und ableitungsbildende Basis. Darauf hin unterscheidet man Intersegmente der ersten und der zweiten Stufe (**IS1,IS2**).

Das Intersegment der ersten Stufe (**IS1**) ist die vom Ende des Stammes isolierte Gruppe von Buchstaben, die sich bei der Flexion des Wortes verändert. Wenn der Stamm (**KM**) unverändert bleibt, **KM=KMZ+0**, dann geht man von einem Nullintersegment aus.

Zu den Intersegmenten der zweiten Stufe (**IS2**) gehören die oben genannten Ableitungssuffixe, die zwischen dem Komplex Stammbasis+Intersegment der ersten Stufe (**KMZ+IS1**) und der Flexionsendung (**T**) der abgeleiteten Form stehen.

Intersegment ist also der Sammelbegriff für alle Segmente, die zwischen den Substantivendungen, Adjektivendungen, allen Typen von Verbalendungen bestimmter und unbestimmter Verbformen, den von Adjektiva paradigmatisch abgeleiteten Adverbialendungen, Endungen der Possessivadjektive, Numeralienendungen und der Stammbasis stehen.

Die Beschreibung des Endungssystems

Die paradigmatischen Endungssysteme unterscheiden sich im Tschechischen der Wortart nach voneinander. Wir wollen zunächst die einzelnen Begriffe definieren. Das Wort wird als eine Kette von Buchstaben zwischen den Leerstellen definiert. Wenn im Rahmen unserer Beschreibung von der Flexionsendung gesprochen wird, ist die Kette, die aus null-, ein-, zwei- oder drei Buchstaben besteht und vom Ende des Wortes getrennt wird, gemeint. Die tschechischen Endungen, die formal als Ketten aus 0-, ein-, zwei- oder drei Buchstaben definiert werden, bilden ein Inventar (eine Menge). Diese Menge wird in ein strukturiertes System von Untermengen aufgeteilt. Das Hauptkriterium dabei war, daß die Endungen in Mengen gruppiert sind, die die formbildende Charakteristik darstellen (Flexionsmuster, Paradigma).

Die formbildende Charakteristik jedes Flexionstyps zerfällt in das System der sogenannten Endungsuntermengen. Jedes klassische Paradigma wird in zwei Gruppen von Endungen geteilt: die Kernmenge (**KMJ**) und das System der Peripheriemengen (**KMP**).

Die Endungsmengen der tschechischen Namen

Die Kernmenge enthält die Endungen, die:

* keine Stammalternation verursachen:
    * keine stammfinale Alternation
    * keine Alternation der Finalgruppe des Stammes
    * keine vokalische Alternation der Wurzel
    * keine orthographische Alternation der Finale

* keine Endungsvariante haben, die begründet sind:
    * historisch
    * aufgrund orthographischer Regelung
    * aufgrund von Schwankung zwischen den einzelnen Flexionsmustern

Die Peripheriemengen enthalten gerade die Endungen, die den obengenannten Kriterien nicht entsprechen und entweder einen Typ von Alternation verursachen, oder je nach dem Flexionsmuster eine Endungsvariante haben.

* Die Endungen, die die Stammalternation verursachen:
    * stammfinale Alternation
    * Alternation der Finalgruppe des Stammes
    * vokalische Alternation der Wurzel
    * ortographische Alternation der Finale

* Die eine Endungsvariante haben, die begründet ist:
    * historisch
    * aufgrund ortographischen Regel
    * aufgrund von Schwankung zwischen den einzelnen Flexionsmustern

* Die speziellen Endungsmengen für die Fremdwortdeklination
* Endungsmengen transponierter Wortklassen (substantivierte Adjektive)

Die Endungsmengen der tschechischen Verben

Die Menge der tschechischen Verbalendungen zerfällt nach den klassischen Subparadigmen des Verbs in die folgenden Untermengen: Indikativ Präsens, Imperativ, Partizip Perfekt, Partizip Passiv, Transgressiv Präsens und Perfekt. Die letzte Untermenge stellt das System der Konditionalendungen des Hilfsverbs „být" („*sein*"). Jede Untermenge wird weiter nach folgenden Kriterien gegliedert:

* Im Rahmen eines Subparadigma vereint sich eine Gruppe von Endungen mit der Wurzel und einem stammbildenden Suffix, die andere Gruppe mit der Wurzel und einem anderen stammbildenden Suffix

* Das Subparadigma enthält parallele alternative Endungsinventare, dessen Distribution
  * historisch
  * aufgrund Lautstruktur der Wurzel bestimmt wird.

Die Beschreibung der einzelnen tschechischen Endungsmengen ist einheitlich und geöffnet. Die Mengen einer Wortart bzw. eines Subparadigma haben dieselbe formale Struktur. In der Beschreibung konnten neue Untermengen ergänzt werden. Die Beschreibung jeder Flexionsendung schließt auch die Informationen über die Wortart und die grammatischen Bedeutungen ein.

Beispiel der Beschreibung der verbalen Endungsmengen

Die Definition der Endungsmenge hat folgende Struktur:

=/ DER NAME DER DEFINIERTEN MENGE/
[WORTART,NUMERUS,GENUS*,STUFE*,MODUS,TEMPUS]
(ENDUNG,KASUS,GENUS,PERSON,STUFE)**

* Die Angabe von Genus und Stufe ist je nach der Wortart fakultativ.

** Die einzelnen grammatischen Bedeutungen sind je nach der Wortart und dem Typ der Flexion fakultativ.

Beispiel:

```
=W1A
  [Usi^]
    (š,B)
    (_,C)
  [Upi^]
    (me,A)
    (te,B)

=W1B
  [Usi^]
    (u,A)
    [Upi^]
    (ou,C)
```

U- Verbum
s,p - Singular, Plural
i - Indikativ
^ - Präsens /Futur
A,B,C - erste, zweite, dritte Person

Die Beschreibung der Flexionsmuster

Bei der Beschreibung des Formbildungsystems ist der Hauptbegriff das Paradigma. Das morphologische Paradigma ist die Gesamtheit der Wortformen, die das System seiner grammatischen Kategorien zum Ausdruck bringen. Bei den Substantiven, Adjektiven usw. stellt das Paradigma das Kasussystem des Singulars und Plurals dar. Bei Verben werden die einzelnen obengenannten Subparadigmen unterschieden, die entweder die Personalformen oder die partizipialen Genusformen ausdrücken.

Die Definitionen der neuen Flexionsmuster sind der Hauptbestandteil der algorithmischen Beschreibung der tschechischen Flexionsmorphologie bzw. der Wortableitung. Die algorithmische Beschreibung schließt, wie ich daherein oben erwähnt habe, die Definitionen der Endungsmengen ein. Die Muster werden dann mittels der Musterwörter definiert, die in drei Teile zerfallen: den relativ stabilen Teil des Wortes - Stambasis (**KMZ**), den variablen Teil des Wortes (**IS**) und die Endunsmengen (**MK**), die alle Endungen umfassen, die mit der Kombination **KMZ+IS** kombinierbar sind, damit ein korrektes tschechisches Wort entsteht.

Sl ⇨ KM+T
KM ⇨ KMZ+IS
IS ⇨ IS1+IS2

Die Modellbeschreibung ist eigentlich eine Formalregel für eine zulässige Kombination der einzelnen Segmente des Wortes.

+KMZ$_a$
<IS1$_a$> KM$_a$,...,KM$_x$
<IS2$_x$> KM$_b$,...,KM$_z$
+KMZ$_b$
<IS1$_{aa}$> KM$_{aa}$,...,KM$_{xx}$
<IS2$_{xx}$> KM$_{bb}$,...,KM$_{zz}$

Beispiel der Beschreibung der Musterdefinitionen

Das Beispiel zeigt die konkrete Form der algorithmischen Beschreibung des Verbalmusters (*hnát/ženu = treiben,ie,ie*).

Die Struktur weist, wie das Wort in drei Teile segmentiert wird, wie das variable Intersegment (**IS**) die Menge der Stämme im Wörterbuch reduziert.

Anstatt *(hna-, hnav-, hná-, hnan-, hnavš- žen-, žene-, žeň-, ženouc-,) 5+4=9* Varianten des Stammes (**KM**) gibt es nur *2* Varianten der Stambasis (**KMZ**).

+hn
<a> W3A
<av> W7
<á> W4C,W5A

<an> V13,PRT1,PRT2
<án> PRT3
<avš> PRMP
+že
<n> W1B,W6A
<ne> W1A
<ň> W2A
<nouc> PRMI

Die statistische Übersicht über die Wechsellbeziehungen zwischen den klassischen und den neuen Flexionsmustern

Die folgende Tabelle zeigt die Wechselbeziehungen zwischen den klassischen und den neuen Flexionsmustern. Die Leerstellen bei den klassischen Mustern der Pronomina und Numeralien drücken die allgemeine Vagheit bei der systematischen Erfassung ihrer Flexion in den klassischen Grammatiken aus. Die Ziffern, die die Situation bei den nichtflektierbaren Wortarten (Präpositionen, Konjunktionen) beschreiben, beruhen auf dem Versuch einer Klassifizierung der Rektion von Präpositionen und einer Differenzierung von Konjunktionen (koordinierende/subordinierende).

Wortart	Klassische Muster	Neue Muster
Substantiven	14	370
Adjektiven	2+2 =4	93
Pronomina	--	55
Numeralien	--	32
Verben	14	202
Adverbien	--	5
Präpositionen	--	8
Konjunktionen	--	2
Partikeln	--	1
Interjektionen	--	1
Zusammmen	--	769

Das Computer-Wörterbuch

Das Computer-Wörterbuch ist eigentlich eine Liste von Stammbasen der flektierbaren Wörter und eine Liste von nichtflektierbaren und unflektierbaren Wörtern. Jeder Stammbasis wird das entsprechende Flexionsmuster zugeordnet, das als zulässige Kombination von **IS+T** für die gegebene Stammbasis aufgefasst wird. Jedem unflektierbaren Wort wird ein Muster zugeordnet, das die einzelne mögliche Form des Wortes als alle möglichen Formen der entsprechenden

Wortart definiert. Jedem nichtflektierbaren Wort wird ein Muster zugeordnet, das Wortart und Typ definiert.

Diese zwei Listen stellen ein einheitliches Computer-Wörterbuch dar, das die automatische morphologische Analyse - die Identifizierung und die Generierung der Wortformen - ermöglicht. Die formale Beschreibung und das Computer-Wörterbuch bilden die Basis der obengenannten Software-Produkte.

Beispiel aus dem Stammbasis-Wörterbuch

hn-á-t, žen-u @V >hn,žn
_, do, na, nad, obe, ode, po, pod, pood, popo, povy, pro,
pře, přede, při, roze, s, se, u, v, vy, za,

Die Automatische morphologische Analyse

Wir wollen noch einmal zusammenfassen, wie die morphologische Analyse mit Hilfe des Computers durchgeführt wird. Das Wort **SI** wird für die geschriebene Sprache, mit der gearbeitet wird, als eine Buchstabenkette zwischen den Leerstellen definiert. Im ersten Schritt werden die nicht- und unflektierbaren Wörter durch Identifizierung der isolierten Ketten im Computer-Wörterbuch ausgegliedert. Wenn die isolierte Kette nicht gefunden und für ein un- oder nichtflektierbares Wort erklärt wird, geht die Analyse weiter. Die isolierte Kette wird schrittweise vom Ende in drei Segmente zerlegt: die Flexionsendung **T**, die formal als Kette von minimal null und maximal drei Buchstaben definiert wird, das **IS**, das formal als Kette minimal null und maximal fünf Buchstaben definiert wird und die Stammbasis **KMZ**, die den Rest des Wortes nach der Abtrennung von **IS+T** darstellt. Die Kombination der Segmente wird im Computer-Wörterbuch der Stammbasen und in der Tabelle der Flexionmuster identifiziert. Die Kombination **KMZ+IS+T**, für die im Computer-Stammbasiswörterbuch die Stammbasis vorhanden ist, der ein Muster zugeordnet wird, das die Kombination **IS+T** zuläßt, wird für die korrekte Wortform erklärt.

3. Die Verwendung in anderen slawischen Sprachen (Slowakisch, Russisch)

Im Zusammenhang mit der Lösung der einzelnen theoretischen und praktischen Probleme und infolge der konkreten praktischen Nachfrage nach linguistischer software, erhob sich die Frage nach einer eventuellen Applikation der Formalbeschreibung des Tschechischen auf die anderen slawischen Sprachen. Die einheitliche Beschreibung der verwandten Sprachen ermöglicht einerseits eine vereinfachte Lösung analogischer Probleme (inspirativ waren vor allem die Analogien im Verbalsystem), andererseits ergeben sich Anwendungen gerade auch bei der Lösung von Problemen, die die Unterschiede zwischen den Sprachen mit sich bringen (rythmisches Gesetz im Slowakischen, Akzentschemata im Russischen). Der große Vorzug einer einheitlichen Lösung liegt in einer

schnellen und einfachen Transformation der existierenden Softwareprodukte und in ihrer gegenseitige Kompatibilität.

4. Weitere Richtungen der Entwicklung

Die formale Beschreibung der tschechischen Flexionsmorphologie und ihre Applikation im Rahmen der Computer-Analyse von natürlicher Sprache hat ihren Wert auch als Bestandteil der Beschreibung höherer Sprachschichten.

1. Lexikographie

Das Computer-Wörterbuch bildet die Basis für den automatischen Lemmatisator, der auch als Bestandteil des tschechischen Computer-Thesaurus (Pala, Všianský 1993) verwendet wird. Es wird auch mit einer Verwendung für die verschiedenen Typen von Übersetzungswörterbüchern gerechnet. Das Computer-Wörterbuch der Stammbasis ist eigentlich ein morphologisches Wörterbuch, das als Grundlage für ein Derivationswörterbuch dienen kann.

2. Syntax

Die in dem morphologischen Analysator enthaltenen Informationen werden für die automatische syntaktische Analyse genutzt. Die möglichen Bedeutungen der Wortformen, die bei der Analyse ermittelt werden, gewinnen erst als die syntaktischen Funktionen Bestimmtheit. Auch die Konstruktion des Stichwortes der Wortart (die Klassifikation der unflektierbaren Wortarten) ist für die Zwecke der syntaktischen Analyse (Synthese) erstellt werden.

3. Wortbildung

Es ist geplant die Verwendung des heutigen Computer-Wörterbuches für die Erstellung eines Computerderivationswörterbuchs. Es wird aufgrund des Stammbasiswörterbuches ein morphematisches Wörterbuch gebildet sein. Jeder Kombination von Morphemen wird eine explizite Beschreibung zugeordnet, ein Derivationsmuster. Die Beschreibung der Derivationsmuster wird der Beschreibung der Flexionsmuster analoge. Zuerst wird die Liste der einzelnen Derivationssuffixen und ihrer möglichen Bedeutungen erstellt werden, dann werden die Derivationsmuster als Kombinationen aus Wurzel und einzelnen Derivationssuffixen definiert.

4. Semantik

Obwohl z.B. die Problematik der Mehrdeutigkeit bei der Lemmatisierung noch nicht gelöst ist, kann das Wörterbuch in Zukunft verschiedenen Experimenten auf dem Gebiet der Semantik dienen.

Automatischer Korrektor des Tschechischen (spelling-checker)

Auf der Grundlage dargestellten linguistischen Beschreibung arbeitet der automatische Fehlerkorrektor. Außer dem Fehlermeldungen könnten unbekannte korrekte Wörter in einem Benutzerwörterbuch gespeichert werden; es werden auch formal ähnliche Wortformen für eine schnellere Korrektur angeboten.

Automatischer Lemmatisator und morphologischer Analysator des Tschechischen, automatischer Tagger

Der automatische Lemmatisator und morphologische Analysator ist ein Programm, das dem Benutzer ermöglicht, aufgrund der oben beschriebenen linguistischen Basis in der interaktiven Ordnung zu jeder Wortform

1. das Lemma — die Grundform (Lemmatisation)
2. die möglichen grammatischen Bedeutungen der Wortform (Analyse)
3. alle zugelassene Wortformen (Synthese) zu finden.

Der automatische Lemmatisator wird auch zum automatischen „tagging" des Korpus benutzt. Das Tschechische Nationalkorpus als elektronisch gespeicherte, elektronisch verarbeitete und elektronisch zugängliche Gesamtheit tschechisch geschriebener oder gesprochener Texte, die bei der Untersuchung der Sprache, für ihre Beschreibung und für die Zusammensetzung verschiedener Wörterbücher als universale Quelle dient, wird jetzt im Institut des tschechischen Nationalkorpus der FF UK erarbeitet. Eine Reihe von Arbeiten, die mit der Verarbeitung des Computerwörterbuches zusammenhing, wurde im Rahmen des Grants 405/93/2018 (Das Korpus der tschechisch geschriebenen Texte) durchgeführt.

5. Zusammenfassung der theoretischen und praktischen Ergebnisse

1. Erste Version der Formalbeschreibung der tschechischen Morphologie und des Computer-Wörterbuches (20 000 Lemmata), die die linguistische Basis des automatischen morpho-syntaktischen Analysators KLARA (Osolsobě, Pala, Franc 1987) bildete.
2. Erweiterte und präzisierte Version der Formalbeschreibung der tschechischen Morphologie; komplettes Stichwortverzeichnis des Wörterbuches der tschechischen Literatursprache (SSJČ). Jeder Grundform (Lemma) wird die Information über die Wortart und den Flexionstyp zugeordnet. Diese Version wurde als linguistische Basis des automatischen Korrektors (Spelling-checker) und des automatischen morphologischen Analysators (Franc, Osolsobě, 1990) verwendet. (Der Spelling-checker wurde in den verbreiteten Textprocessor T602 implementiert.)
3. Aufgrund der Erfahrungen mit dem Tschechischen wurde eine analogie Beschreibung und ein automatischer Korrektor für die slowakische Morphologie entwickelt (Franc, Osolsobě, 1990). Der Spelling-checker wurde in den Editor T602 implementiert.
4. Aufgrund der Bearbeitung verschiedener Texte (Fachliteratur, Publizistik) wurde eine Liste von unbekannten Wörtern für eine Erweiterung des Wörter-

buches erstellt. So entstand die neue erweiterte Version des Computer-Wörterbuches und die neue korrigierte Version der Formalbeschreibung. Die neuen Versionen des automatischen Korrektors wurden im WP, Microsoft - Word implementiert. Die algorithmische Beschreibung wurde zur Basis für den automatischen Lemmatisator und dem morphologischen Analysator (Osolsobě, Ševeček, 1993–1995).

5. Aufgrund der Erfahrungen mit dem Tschechischen wurde eine analoge Beschreibung für die deutsche Morphologie entwickelt (Osolsobě, Ševeček, 1993).
6. Aufgrund der Erfahrungen mit dem Tschechischen wurde eine analoge Beschreibung für die französische Morphologie entwickelt (Osolsobě, Ševeček, 1994).
7. Weitere Arbeit an der Erweiterung und formalen Bearbeitungen des Computerwörterbuches. Das Programm für die automatisierte Transkription der Zahlwörter in Ziffern und umgekehrt. (Osolsobě, Ševeček 1994).
8. Aufgrund der Erfahrungen mit dem Tschechischen wurde eine analoge Beschreibung für die russische Morphologie entwickelt (Osolsobě, Ševeček, 1995).
9. Entwurf einer integrierten Beschreibung von geschriebenen und gesprochenen Formen der tschechischen Morphologie erfaßt das Programm für die automatische Lemmatisation und Analyse der geschriebenen und gesprochenen morphologischen Formen (Osolsobě, Ševeček, 1995).

FORMÁLNÍ POPIS ČESKÉ MORFOLOGIE A JEHO VYUŽITÍ V OBLASTI NLP

Článek shrnuje výsledky v oblasti počítačového zpracování přirozeného jazyka (Natural Language Processing - NLP), jichž bylo dosaženo v rámci výzkumu Kabinetu počítačové lingvistiky FFMU, Brno.

Kap. 2. představuje teoretická východiska a metody použité pro algoritmický popis české formální morfologie.

Kap. 3. informuje o konkrétním využití počítačového modelu české flexe pro další slovanské jazyky.

Kap. 4. se zamýšlí nad dalšími možnostmi rozvoje a přesahu při algoritmickém popisu vyšších rovin přirozeného jazyka.

Kap. 5. uvádí přehled jednotlivých dílčích výsledků v oblasti budování lingvistického software v letech 1990–1995.

LITERATURA

BOGURAEV, B., BRISCOE, T. (1990): Computational Lexicography for Natural Language Processing, Longman, London and New York.
ČERMÁK, F. (1990): Syntagmatika a paradigmatika českého slova I., II., UK, Praha.
ČERMÁK, F., BLATNÁ, R. (1995): Manuál lexikografie, Nakladatelství H&H, Praha.
HAVRÁNEK, B., JEDLIČKA, A. (1981): Česká mluvnice, SPN, Praha.
KOMÁREK, M. (1987): Ke dvěma koncepcím stavby jednoduchých slovesných tvarů v češtině, Acta Universitatis Olomucensis, Studia Bohemica IV, SPN, Praha.
MATERNA, P., PALA, K., ZLATUŠKA, J. (1989): Logická analýza přirozeného jazyka, Academia, Praha.
MISTRÍK, J. (1976): Retrográdny slovník slovenčiny, Univerzita Komenského v Bratislave, Bratislava.
OSOLSOBĚ,K., PALA,K., FRANC,S. (1987): Česká morfologie a syntax v PROLOGU, sb. semináře SOFSEM 1987, VUSEIAR, Bratislava.
HALASOVÁ-OSOLSOBĚ, K. (1990): Algoritmický popis české formální morfologie substantiv a adjektiv, SPFFBU, A 37–38, 1989–90,s.83–97.
OSOLSOBĚ,K. , PALA, K. (1990): Czech Stem Dictionary for IBM PC XT/AT, Conference on Computer Lexicography, Balatonfüred, September.
OSOLSOBĚ, K. (1991): Popis systému českých substantivních a slovesných vzorů, rukopis, Brno.
OSOLSOBĚ,K. , PALA, K. (1993): Czech Stem Dictionary, SPFFBU, A 41, 1993, s. 70–83, Brno.
OSOLSOBĚ,K. (1994): Česká formální morfologie na počítači, aneb jak se počítač učil časovat česká pravidelná a nepravidelná slovesa, In: Přednášky a besedy z XXVII. běhu LŠSS, Brno, s. 16–31.
OSOLSOBĚ,K. (1995): Automatické rozpoznávání a generování českých určitých číslovek a od nich odvozených číselných pojmenování na počítači, SPFFBU A 43, 31–48, Brno.
PALA, K., VŠIANSKÝ, J. (1994): Slovník českých synonym, Nakladatelství Lidové noviny, Praha.
PETR, J., kol., Mluvnice češtiny I.,II. (1986): Academia, Praha.
PŘÍRUČNÍ MLUVNICE ČEŠTINY (1995), editoři: Karlík, P., Nekula, M., Rusínová, Z., Nakladatelství Lidové noviny, Praha.
ROMPORTL, S. (1961) Návrh principu automatického šifrování a dešifrace gramatických příznaků českého slovesa při překládání z češtiny do češtiny, In: SbVUT, Brno.
ROMPORTL, S. (1970): Struktura gramatické složky slovesných tvarů určitých v češtině, Academia, Praha.
SGALL, P. (1967): Generativní popis jazyka a česká deklinace. ČSAV, Praha.
SLOVNÍK SPISOVNÉHO JAZYKA ČESKÉHO 1–8 (1989): Academia, Praha.
ŠONKOVÁ, J. (1995): Morfologie mluvené češtiny, kandidátská práce, Praha.
ZALIZNJAK, A. A. (1977): Grammatičeskij slovar' russkogo jazyka, Izdatel'stvo Russkij jazyk, Moskva.