MORPHO-2: MEDIU DE PROIECTARE ŞI DEZVOLTARE A LEXICOANELOR MONOLINGVE

Ing. Cristian Dumitrescu

Institutul de Cercetări în Informatică

Rezumat

În lucrare se descrie sistemul MORPHO-2, sistem proiectat în scopul gestiunii lexicoanelor monolingve şi a manipulării on-line a proceselor lexicale. Serviciile asigurate de sistem permit modelarea şi manipularea informaţiilor de natură morfologică, sintactică şi semantică la nivelul lexiconului. La proiectarea şi dezvoltarea proceselor lexicale, se are în vedere realizarea unei distincţii nete între procesele de prelucrare şi cunoştinţele care le guvernează. În acest mod, se asigură independenţa între limbajele (cu caracter preponderent flexionar) prelucrate şi mediul de procesare.

1. Introducere

Cercetările şi rezultatele obţinute în domeniul prelucrării limbajului natural, cu precădere în dezvoltarea sistemelor de traducere automată, au demonstrat rolul deosebit de important al lexiconului. Abordarea iniţială, care privea lexiconul ca pe o bază de date lexicală gestionată cu ajutorul unui Sistem de Gestiune a Bazelor de Date (SGBD), este tot mai mult criticată [Domenig, 1986,1988]. Principalele argumente împotriva acestei soluţii pot fi rezumate astfel:

a) un SGBD este inutil de general pentru problema lexicală, oferind un nivel scăzut de abstractizare şi un timp de răspuns prea mare;

b) un SGBD nu poate gestiona procese morfologice, iar limbajul de definire al datelor nu oferă mijloace pentru exprimarea acestor procese.

În accepţiunea modernă, lexiconul încetează a mai fi privit ca o colecţie amorfă de date, cercetările în acest sens concretizîndu-se prin modele şi tehnici specifice agregate în medii procesuale, referite în literatura sub numele de sisteme de gestiune morfo-lexicală.

Sistemul descris în lucrarea de faţă permite manipularea lexicoanelor monolingve şi înglobarea proceselor lexicale la nivelul lexiconului.

Întrucît procesele morfologice se circumscriu din punct de vedere teoretic unei morfologii de tip flexionar ([Tufiş, 1989], [Tufiş, 1990a]), analiza şi sinteza cuvintelor iau în consideraţie numai terminaţiile gramaticale, iar lexicoanele manipulate de MORPHO-2 sînt orientate numai pe rădăcini sau leme. Serviciile asigurate de sistem pot fi clasificate în raport cu următoarele obiective: proiectarea modelului morfologic, construirea fondului lexical şi manipularea informaţiilor lexicale (analiza şi sinteza cuvintelor).

Pentru îndeplinirea acestor obiective sînt puse la dispoziţia utilizatorilor trei tipuri de interfeţe: interfaţa lexicologului, interfaţa lexicografului şi interfaţa procesorului ţintă.

2. Interfaţa Lexicologului

Pentru construirea unui model morfologic, lexicologul are la dispoziţie un mediu integrat de dezvoltare ce permite editarea, vizualizarea şi compilarea descrierii modelului morfologic.

Ca urmare a specificaţiilor furnizate de lexicolog, în final se obţine un raport lexicografic şi o reprezentare compilată a modelului morfologic.

Definirea unui model morfologic se face în mai multe etape, în care lexicologul specifică următoarele informaţii:

- părţi de vorbire, caracteristicile şi valorile lor, precum şi dependenţele dintre acestea;
- descrierile paradigmatice;
- specificările de caracteristici implicite ataşate fiecărei descrieri paradigmatice;
- corespondenţa lema - intrare din descrierea paradigmatică pentru fiecare descriere paradigmatică;
- paradigmele flexionare şi regulile de detecţie a rădăcinilor.

Vom detalia în continuare, pe rînd, fiecare etapă din definirea unui model.

Unei părţi de vorbire (SUBSTANTIV, VERB, PRONUME etc.) îi putem asocia mai multe caracteristici.

Ca exemple de caracteristici putem enumera MOD, TIMP, GEN, CAZ etc. Pentru caracteristica MOD vom avea valorile: INDICATIV, IMPERATIV, ..., GERUNZIU.

Întrucît, în faza actuală, recunoaşterea morfologică se face numai la nivel de cuvînt individual, caracteristicile şi valorile acestora vor fi limitate la o submulţime corespunzătoare.

O relaţie de dependenţă pune în corespondenţa unei caracteristici mai multe valori:

MOD → (INDICATIV, IMPERATIV, ..., GERUNZIU)

TIMP → (PREZENT, IMPERFECT, ..., VIITOR)

În mod practic pentru caracteristici şi valori se folosesc abrevieri de genul:

NR → (SG, PL)

PER → (1, 2, 3)

Pentru uniformitatea reprezentării putem asocia o astfel de relaţie şi părţilor de vorbire:

PV → (SB, ADJ, VB, ...)

În descrierea unui model morfologic, o relaţie de dependenţă este dată sub forma unei perechi de forma (caracteristică : valoare $^{+}$).

Numim descriere paradigmatică o descriere ierarhică construită din mai multe perechi simple de tipul (caracteristică : valoare).

În figura următoare este prezentată parţial, sub forma unui arbore incomplet, descrierea ierarhică a

categoriilor gramaticale din modelul morfologic pentru limba română. Prin traversarea completă a arborelui, se pot genera toate descrierile paradigmatice ale modelului.
Într-un astfel de arbore fiecare nod neterminal conține o singură specificare de caracteristică, sub forma (caracteristică : valoare $^{+}$). Nodurile terminale însă, pot conține una sau mai multe specificări de caracteristici.

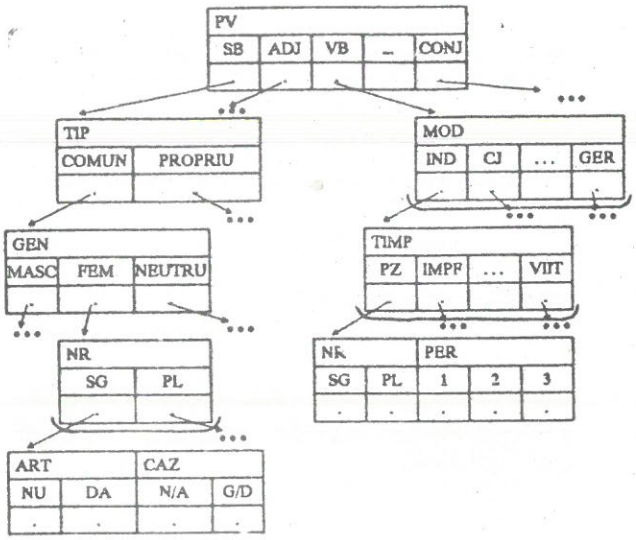

Fig.1. Descrierea ierarhică într-un model morfologic

Pentru fiecare valoare a caracteristicii dintr-un nod neterminal exista cîte un nod succesor. În funcție de criteriul de selecție a succesorilor, aplicat la traversarea unui nod neterminal, distingem noduri de tip CHOOSE sau de tip FOREACH.
Parcurgerea unui nod CHOOSE impune selecția unui singur succesor s_i, includerea perechii (caracteristică : valoare$_i$) în descrierea paradigmatică curentă și generarea în continuare a descrierii pentru subarborele dominat de s_i.
Prin traversarea celei mai lungi căi, ce pleacă din nodul rădăcină și care este formată numai din noduri CHOOSE se obține selectorul unei descrieri paradigmatice.
Pentru exemplul din figura anterioară expresiile următoare:

a) (PV = SB) AND (TIP = COMUN) AND (GEN = FEM)
b) (PV = VB)

sînt selectori.
Un selector astfel format reprezintă primul nivel în ierarhia descrierii paradigmatice.
Parcurgerea unui nod FOREACH impune selecția pe rînd a fiecărui succesor s_j (j = 1, ordin nod), includerea perechii (caracteristică: valoare$_j$) pe același nivel în ierarhia descrierii paradigmatice și generarea în continuare a descrierii pentru subarborele dominat de s_j. În figura un nod FOREACH este scos în evidență prin trasarea unei linii curbe peste arcele divergente din nod.
Descrierea atașată unui nod terminal este reprezentată de un scenariu de achiziție morfologică. Un astfel de scenariu se construiește pe baza produsului cartezian al mulțimilor obținute prin expandarea perechilor (caracteristică : valoare^{+}). Expandarea se face astfel încît pentru o pereche
(caract$_i$: val$_{i1}$, val$_{i2}$... val$_{ik}$)
să se obțină mulțimea:
{(caract$_i$: val$_{i1}$), (caract$_i$: val$_{i2}$), ... (caract$_i$: val$_{ik}$)}
Unei intrări dintr-un scenariu (numită în continuare slot) îi este atașată o descriere morfologică, ce corespunde de fapt unui punct în spațiul paradigmatic al descrierii.
Pentru specificațiile din Fig.1, în contextul generării descrierilor conform celor prezentate, se vor obține următoarele descrieri paradigmatice:

PV = SB & TIP = COMUN & GEN = FEM
NR = SG

ART	CAZ	WORD_FORM
NU	N/A	
NU	G/D	
DA	N/A	
DA	G/D	

NR = PL
...
...
PV = VB
MOD = IND
TIMP = PZ

NR	PER	WORD_FORM
SG	1	
SG	2	
SG	3	
PL	1	
PL	2	
PL	3	

TIMP = IMPF
...
TIMP = VIIT
...
MOD = CJ
...

Fig.2. Meniuri de achiziție morfologică

Ultima coloană din fiecare scenariu (numita WORD_FORM)este adăugată automat de către sistem și permite lexicografului inserarea formelor flexate ale paradigmei atașată cuvîntului curent.
Construirea unui model morfologic se face cu ajutorul unui limbaj adecvat [Dumitrescu, 1991a]. În acest limbaj există declarații speciale ce permit descrierea unui arbore de tipul celui descris.
Caracteristicile ultimei coloane dintr-un scenariu sînt specificate înaintea descrierii arborelui printr-o declarație de forma:

WORD_COLUMN = ("WORD_FORM", n)

unde n este lungimea cuvîntului maxim ce va fi introdus în lexicon.
În urma compilării modelului morfologic se poate obține un raport lexicografic, la solicitarea explicită a lexicologului, ce va conține printre altele și meniurile de achiziție morfologică.
După prezentarea completă a descrierilor paradigmatice, în dezvoltarea modelului morfologic urmează etapa în care se stabilesc specificările de caracteristici implicite atașate fiecărei descrieri.
Selectorilor de descrieri paradigmatice care acceptă specificări de caracteristici implicite li se pun în corespondență perechile (caracteristica: valoare$^+$) ce sînt considerate ca implicite.
În exemplul nostru este posibilă următoarea asociere:

(PV = VB) — (PER : 1 2 3)

și are următoarele semnificații:

a) perechea (PER: 1 2 3) se va atașa ca o proprietate a nodului ce conține perechea (MOD: IND CJ ... GER).

În cazul general proprietatea se atașează succesorului ultimului nod din calea ce definește selectorul.

b) proprietatea este globală subarborelui dominat de nod;

c) sloturile (generate la nivelul subarborelui) cărora le corespund puncte din spațiul descrierii paradigmatice ce nu conțin explicit caracteristica PER vor moșteni perechea (PER : 1 2 3).

d) pentru sloturile care conțin descrieri locale ale caracteristicii PER, proprietatea este inhibată.

În zona din modelul morfologic unde sînt descrise corespondentele (lema - intrare în descriere paradigmatică) sînt enumerate punctele din spațiile paradigmatice ale descrierilor ce caracterizează cîmpul lema din intrarea în lexicon.
Necesitatea cîmpului lema apare evidentă în contextul problemei traducerii automate. Lexiconul monolingv trebuie privit în această situație ca o componentă într-o organizare bilingvă a lexicoanelor (lexicon monolingv sursă, lexicon monolingv țintă și lexicon de transfer) [Tufiș, 1990c]. La nivelul lexiconului de transfer se manipulează doar leme, analiza cuvintelor și generarea formelor flexate fiind rezolvate la nivelul lexiconului sursă și respectiv ținta.
În ultima etapă din descrierea unui model morfologic, lexicologul are posibilitatea de a informa sistemul cum să construiască paradigmele flexionare și regulile de detecție a rădăcinilor.
Pentru fiecare descriere paradigmatică (identificată printr-un selector) lexicologul poate specifica mai multe mulțimi de terminații. Deoarece fiecare terminație dintr-o mulțime va corespunde unui slot din descrierea paradigmatică, vom numi o astfel de mulțime "familie paradigmatică de terminații".
O primă decizie luată de sistem, la interpretarea unei astfel de familii, se referă la memorarea terminațiilor. Terminațiile sînt organizate arborescent într-o structură *trie*. O astfel de structură este convenabilă identificării unei terminații, parcurgerea arborelui fiind coordonată de caractere preluate de la sfîrșitul cuvîntului.
Nodurile terminale (care nu sînt neapărat frunze) au asociate puncte din spațiul descrierilor paradigmatice ce caracterizează terminațiile definite de nodurile respective.
Tot dintr-o familie paradigmatică de terminații sistemul construiește paradigmele flexionare. O paradigmă flexionară este o familie paradigmatică îmbogățită cu descrierile morfologice corespunzătoare terminațiilor. Pentru limba română au fost identificate 136 de paradigme flexionare ([Tufiș, 1990a], [CPCOCLN,1982]).
Pe baza paradigmelor flexionare, sistemul determină regulile de detecție a rădăcinilor și de generare a cuvintelor.
O astfel de regulă are următoarea formă:
<inflexiune> ↔ [<paradigma_flexionară><număr_slot>]
și îi corespund următoarele interpretări:
a) dacă un cuvînt se termină cu <inflexiune> atunci
- rădăcina este ceea ce rămîne din cuvînt după ce s-a înlăturat inflexiunea <inflexiune>;
- rădăcina aparține paradigmei <paradigma_flexionară>;
- informațiile contextuale valabile pentru cuvîntul curent sînt date de <număr_slot>.

b) dacă un cuvînt aparține paradigmei <paradigma_flexionară> și cuvîntul este folosit în contextul dat de <număr_slot> atunci
- cuvîntul se obține din concatenarea rădăcinii date, cu inflexiunea <inflexiune>.

Interfața lexicografului este strict dependentă de specificațiile din interfața lexicologului, întrucît o mare parte din interfața de dialog a lexicografului este construită automat pe baza acestor specificații.
Dîndu-se o specificare compilată a modelului morfologic definit de lexicolog, MORPHO-2 poate genera un dialog care ghidează lexicograful, prin navigarea structurilor ierarhice, către identificarea descrierilor morfologice și a regulilor flexionare care trebuie activate la construirea unei noi intrări în lexicon.

3. Interfața Lexicografului

MORPHO-2 oferă lexicografului posibilitatea de a defini noi intrări în lexicon pe baza unui regim conversațional orientat pe lucrul cu ferestre.
O intrare în lexicon are următoarea structură:

(<lema>
  (<selector_descriere_paradigmatică>
   <paradigmă_flexionară>
    (<rădăcină><descriere_morfologică>)*
   (<descriere_sintactică><descriere_semantică>*)*
)*
)

În descrierea de mai sus cîmpurile <lema> și <selector_descriere_paradigmatică> au înțelesul

evident.
Lexicograful poate specifica identificatorul paradigmei flexionare de care aparţin rădăcinile curente, prin selecţia paradigmei corespunzătoare din cele care îi sînt propuse de către sistem. În acest sens lexicograful poate consulta raportul obţinut în urma compilării modelului morfologic sau poate activa o fereastră în care sistemul va pune la dispoziţie paradigmele posibile pentru lema şi partea de vorbire specificate.
Dacă nu exista o astfel de paradigmă, MORPHO-2 solicită lexicografului specificarea unei noi paradigme ce urmează a fi inserată în lexicon.
Cîmpul <rădăcina> poate conţine una sau mai multe rădăcini. Trebuie subliniat că în abordarea noastră ([Tufiş,1989], [Tufiş, 1990c]), rădăcina este definită ca partea cea mai lungă din cuvînt care este stabilă în raport cu o paradigmă flexionară.
Comenzile accesibile lexicografului în această fază permit două moduri de operare: inserarea de rădăcini fără controlul sistemului şi inserarea supervizată de sistem.
Primul mod de lucru este selectat în cazul cuvintelor regulate, cînd este suficientă specificarea unei singure rădăcini. Pe baza paradigmei flexionare şi a rădăcinii, sistemul poate genera toate formele flexionare.
Al doilea mod de lucru este activat în cazul cuvintelor neregulate, cînd sistemul, prin intermediul scenariilor de achiziţie morfologică, va solicita pe rînd rădăcinile neregulate. Pentru a uşura sarcina lexicografului, sloturile sînt deja completate cu terminaţiile corespunzătoare.
Înserarea rădăcinilor în lexicon se face astfel încît acestea să moştenească descrierile morfologice corespunzătoare sloturilor unde apar. Pentru rădăcinile care au ataşate mai multe descrieri morfologice, sistemul încearcă o compactare a acestora.
Cîmpul <descriere_sintactică> conţine referinţe către şabloane sintactice (modele de valenţă), care guvernează utilizarea corectă a cuvîntului în frază [Tufiş, 1991].
În continuare, pentru fiecare descrierea sintactică, lexicograful poate furniza una sau mai multe descrieri semantice. Cîmpul <descriere_semantică> conţine numele unei structuri cadru plasată într-o ierarhie de tipul generic-specific. Descrierile semantice propriu-zise sînt memorate într-un spaţiu de date separate de lexicon [Nirenburg,1987], iar gestiunea acestora este independentă de MORPHO-2.
În abordarea noastră, lexicografului i se creează posibilitatea de a specifica descriptorii semantici în colaborare cu proiectantul sistemului ţintă de prelucrare a limbajului natural, în conformitate cu categoriile utilizate de sistemul de prelucrare semantică. În acest mod, se asigură o independenţă între sistemul de gestiune a lexicoanelor, ca o unealtă lingvistică de bază şi un lexicon specific unei anumite aplicaţii.
Sistemul mai pune la dispoziţia lexicografului comenzi pentru ştergerea, modificarea unei intrări în lexicon şi listarea acestora, în conformitate cu diverse cereri ce vizează cîmpurile unei intrări [Dumitrescu,1991b].

4. Interfaţa Procesorului Ţintă

Procesorul Ţintă este beneficiarul proceselor lexicale executate de MORPHO-2. Analiza şi sinteza cuvintelor sînt servicii mediate de o interfaţă de proces.
În cazul analizei lexicale, dacă acestei interfeţe i se dă o secvenţă de cuvinte, atunci aceasta întoarce o secvenţă de atomi lexicali. Structura unui astfel de atom este prezentată mai jos.

```
(<rădăcină>
 (<lema>
  (<selector_descriere_paradigmatică>
   <descriere_morfologică>
    (<descriere_sintactică><descriere_semantică>*)*
  )*
 )
)
```

O descriere morfologică conţine atît informaţii contextuale, cît şi informaţii independente de context. Primele sînt obţinute din analiza terminaţiei, iar cele din urmă din intrarea din lexicon corespunzătoare rădăcinii. Informaţiile pentru celelalte cîmpuri din structura unui atom sînt de asemenea preluate din intrarea din lexicon corespunzătoare rădăcinii.
În raport cu rezultatul congruentei morfologice şi al regăsirii rădăcinii în lexicon, putem clasifica atomii lexicali ca fiind univoci, ambigui sau nedeterminaţi.
Atomii lexicali univoci pun în corespondenta unui cuvînt analizat o singură lema.
Atomii lexicali ambigui provin din cuvintele cărora li se pot pune în corespondenţă mai multe leme. În acest caz, unei rădăcini îi corespund, fie mai multe leme, fie cuvîntul, avînd mai multe segmentări posibile şi rădăcinile respective corespund unor leme diferite.
Asocierea unei rădăcini cu mai multe leme este posibilă fie datorită părţilor de vorbire diferite (ex: *vin* poate fi substantiv sau verb), fie datorită homografiilor aparente, generate de absenţa în grafia limbii române a marcatorilor prozodici (módele, modéle, tórturi, tortúri, acéle, ácele, modúl, módul etc.).
Interpretările posibile sînt astfel ordonate, încît prevalează cele care provin din rădăcini mai scurte (deci terminaţii mai lungi).
Pentru o rădăcină care corespunde unei leme şi are mai multe descrieri morfologice posibile pentru aceeaşi parte de vorbire, sistemul încearcă o compactare a acestor descrieri.
Atomii lexicali nedeterminaţi corespund cuvintelor pentru care nu există intrări în lexicon. Atomii generaţi în această situaţie au următoarea structură:

```
(UNKNOWN<cuvînt_necunoscut>
  (<rădăcină_posibilă><descriere morfologică>*)*)
```

Cuvîntului necunoscut îi sînt asociate toate segmentările legale, pentru fiecare dintre ele fiind inserată informaţia morfologică dedusă din terminaţiile identificate.

Sinteza lexicală este reversul analizei lexicale. Interfața de proces asigură transformarea unei seveņte de atomi lexicali într-un șir de cuvinte.
Spre deosebire de analiză, pentru sinteză MORPHO-2 solicită o descriere standard a atomilor lexicali sub forma:
(<identificator_intrare><descriere_morfologică>
<descriere_sintactică>)
unde <identificator_intrare> poate fi o lema, o rădăcină, sau o descriere semantică.
Trebuie să subliniem că, într-o fază premergătoare analizei și sintezei lexicale, pe baza unui protocol de comunicație, procesorul țintă poate configura structura atomilor lexicali în raport cu aplicația proiectată.

5. Implementare

Proiectul MORPHO, început în 1986, a avut ca prim rezultat realizarea unei versiuni prototip [Tufiș,1990b] operațională în momentul de față pe un calculator compatibil PDP-11. Versiunea a doua a sistemului, cea prezentată în lucrarea de față, este implementată în C pe un calculator personal de tip IBM-PC.
Tehnicile utilizate în implementare, de indexare a lexicoanelor cu ajutorul arborilor de tip B^+ virtual prefixați [Dumitrescu,1988] precum și de grupare a datelor în structuri "cluster", au dus la obținerea unui timp mediu de răspuns al proceselor lexicale, care să nu se schimbe semnificativ în raport cu creșterea dimensiunii lexiconului.

Bibliografie

-[CPCOCLN,1982]-*Cercetări privind comunicarea om-calculator în limbaj natural,* Raport final la contractul de cercetare ICI - Universitatea Al.I.Cuza-Iași, 1982.
- [Dumitrescu,1988] - DUMITRESCU, C.- *O metodă de memorare și adresare pe baza arborilor B+ virtual prefixați,* Raport tehnic, I.C.I., București, 1988.
- [Dumitrescu,1991a] - DUMITRESCU, C.- Limbaj de descriere a categoriilor gramaticale, Raport tehnic, București, 1991.
-[Dumitrescu,1991b]- DUMITRESCU,C.- *MORPHO-2 Manual de utilizare,* I.C.I., București, 1991.
- [Domenig,1986] - DOMENIG, M., SHANN, P. -*Towards a Dedicated Data Base Management System for Dictionaries.* In: Proceedings of COLING'86 Conference, Bonn, 1986, pp.91-96.
- [Domenig,1988] - DOMENIG, M.- Word Manager: *A System for the Definition, Access and Maintenance of Lexical Data Bases.* In: Proceedings of COLING'88, Budapest, 1988, pp.154-159.
- [Nirenburg,1987] - NIRENBURG, S., RASKIN, V. -*The Subworld Concept Lexicon and the Lexicon Management System.* In: Computational Linguistics, Vol.13, No.3-4, 1987, pp.270-289.
- [Tufiș,1987] - TUFIȘ, D., DUMITRESCU, C. -*O abordare paradigmatică a problematicii dicționarelor,* Simpozionul "Aplicațiile inteligenței artificiale", București, 1987.
- [Tufiș,1989] - TUFIȘ, D.- *It Would Be Much Easier if WENT Were GOED.* In: Proceedings of ECACL'89, Manchester, 1989, pp.145-152.
- [Tufiș,1990a] - TUFIȘ, D.- *Paradigmatic Morphology Learning.*; In: Computers and Artificial Intelligence, Vol.9, No.3, 1990, pp.273-290.
- [Tufiș,1990b] - TUFIȘ, D., DUMITRESCU, C.- *MORPHO - A Dictionary Management System;* In: Proceedings of the 13th International Seminar on DBMS, Mamaia, 1990.
- [Tufiș,1990c] - TUFIȘ, D., DUMITRESCU, C., POPESCU, O.- *Morfologia și dicționarul;* Raport tehnic, București, 1990.
- [Tufiș,1991] - TUFIȘ, D., DUMITRESCU, C., POPESCU, O.- *Teorii sintactice în lingvistica computațională,* Studiu, București, 1991.