CREAREA RESURSELOR LINGVISTICE CU AJUTORUL UNUI LIMBAJ SPECIALIZAT

STEFAN DIACONESCU[1]

[1]*SOFTWIN, Bucuresti, România*

sdiaconescu@softwin.ro

Rezumat

Comunicarea de fata se refera la o metoda ce permite tratarea relativ unitara a mai multor capitole din lingvistica prin intermediul unui limbaj de reprezentare a cunostintelor lingvistice numit GRAALAN (Grammar Abstract Language). Acest limbaj ofera unui lingvist posibilitatea descrierii eficiente a cunostintelor lingvistice privind o limba naturala precum si corespondenta între doua limbi naturale.

1. Introducere

Exista numeroase si fructuoase încercari de uniformizare a reprezentarii cunostintelor lingvistice. O asemenea uniformizare ar oferi un avantaj foarte mare în dezvoltarea unor studii, statistici si, în cele din urma, aplicatii lingvistice care sa poata trata într-un mod asemanator diverse limbi naturale sau sa poata compara (stabili corespondente) într-un mod unitar între diverse limbi naturale. Din pacate diversele capitole lingvistice au suferit abordari întrucâtva independente, cum ar fi subcategorizarea (EAGLES, 1996b) adnotarea (EAGLES, 1996b), lexiconul (EAGLES, 1993), etc. astfel încât este uneori dificil de aplicat tratamente unitare.

Comunicarea de fata prezinta un limbaj de reprezentare a cunostintelor lingvistice numit GRAALAN (Grammar Abstract Language). Acest limbaj permite unui lingvist descrierea eficienta a cunostintelor lingvistice privind o limba naturala precum si corespondentele între doua limbi naturale.

2. Caracteristicile generale ale GRAALAN

Din punct de vedere teoretic, GRAALAN se bazeaza în special pe urmatoarele notiuni: gramatici generative de dependente (GDG - Generative Dependency Grammar) (Diaconescu, 2002), arbori de dependente (DT - Dependency Tree) (Diaconescu, 2002) si arbori atribut - valoare (AVT - Attribute Value Tree) (Diaconescu, 2005).

Pornind de la aceste notiuni, GRAALAN poate descrie diverse capitole lingvistice conforme cu gramaticile conventionale ale limbilor naturale: alfabetul, despartirea în silabe, morfologia, sintaxa, regulile de flexiune, formele de flexiune, lexiconul, corespondente lexicale între doua limbi (inclusiv între expresii multicuvânt MWE - Multiword Expression), corespondente morfologice, corespondente sintactice.

GRAALAN este în esenta un limbaj descriptiv care permite însa eventual si legatura cu anumite subprograme de tip procedural scrise în alte limbaje de programare.

În principiu, descrierile GRAALAN vor putea fi convertite printr-un compilator adecvat în formatul XML care este mai adecvat exploatarii ulterioare prin diverse programe.

3. Descrierea alfabetului

În GRAALAN se pot preciza pentru o anumita limba: alfabetul fonetic utilizat în descrierea limbii (care poate fi un subset al IPA (International Phonetic Alphabet) (IPA, 2005)), alfabetul normal si caracterele speciale.

În afara de acestea se mai pot defini: i) grupe de caractere (diftongi, triftongi, etc.), transcrise cu caractere normale (eventual speciale) dar si fonetice; ii) clase alfabetice (de exemplu clasa vocalelor, clasa consoanelor, etc.)

Caracterele folosite în GRAALAN se considera codificate în UNICODE (ISO, 1992).

4. Descrierea despartirii în silabe

În GRAALAN sunt considerate trei tipuri de despartire în silabe: i) Despartirea eufonica a cuvintelor scrise cu alfabetul normal si respectând modul de pronuntie; ii) Despartirea fonetica a cuvintelor scrise cu alfabetul fonetic si respectând de asemenea modul de pronuntie; iii) Despartire morfologica - analoga cu despartirea eufonica însa respectând si restrictii ce tin cont de structura morfematica a cuvântului.

Primele doua tipuri au reguli specifice. Ulitmul tip nu are reguli speciale deoarece ea actioneaza ca o despartire eufonica cu restrictiile suplimentare privind morfemele obtinute din consultarea lexiconului.

5. Descrierea morfologiei

În GRAALAN, morfologia unei limbi (mai exact ansamblul categoriilor lexicale si al valorilor lor), se reprezinta sub forma unui arbore atribut valoare (AVT) (Diaconescu, 2005) în care nodurile de tip atribut corespund categoriilor lexicale iar nodurile de tip valoare corespund valorilor categoriilor lexicale. În plus, cele doua tipuri de noduri mai au atasate diverse alte tipuri de informatii: numele, abrevieri, (eventual) atasamente procedurale, etc.

În sectiunea corespunzatoare morfologiei se poate indica de asemenea daca anumitor situatii de flexiune distincte le corespund forme flexionate identice.

6. Descrierea lexiconului

Lexiconul GRAALAN este un ansamblu de intrari de diverse tipuri: i) Morfeme (radacini, prefixe, sufixe, prefixoide, sufixoide, etc.); ii) Cuvinte care la rândul lor pot fi: intrari principale de tip lema, intrari suplimentare (care însotesc o lema), intrari principale care nu sunt însa leme; iii) MWE-urile carora li se indica si structura sub

forma unui arbore de dependente (DT); iv) Structuri morfologice analitice sau analitico-sintetice (forme flexionate formate din mai multe cuvinte) analoge MWE-urilor; v) Structuri sintactice de asemenea analoge MWE-urilor.

În functie de tipul lor, intrarile în lexicon mai pot avea asociate si alte tipuri de informatii: semantice, etimologice, morfologice, etc.

Lexiconul în general nu este scris direct în GRAALAN ci se creeaza cu ajutorul unui instrument specializat.

7. Descrierea regulilor de flexiune

Intrarea din lexicon care se poate flexiona (lema de exemplu) identifica o regula compusa de flexiune aflata în sectiunea GRAALAN a regulilor de flexiune. Regula compusa este o lista de reguli de baza. O regula de baza este de fapt un arbore atribut valoare care indica mai multe situatii de flexiune, câte una pentru fiecare frunza a sa. Fiecare situatie de flexiune (deci frunza) are asociata una sau mai multe reguli de flexiune elementare. O regula de flexiune elementara contine: i) O conditie de aplicare a regulii; ii) O secventa de transformari care trebuie facute asupra lemei (sau asupra altei forme de flexiune) pentru a obtine forma de flexiune curenta exprimata în alfabetul normal; iii) Analog cu (ii) pentru alfabetul fonetic; iv) În cazul formelor analitico-sintetice - o caracterizare sub forma unui AVT pentru fiecare cuvânt component si relatiile care se afla între diversele cuvinte componente.

Pe baza regulilor de flexiune aplicate intrarilor din lexicon se pot obtine formele din sectiunea GRAALAN a formelor de flexiune.

8. Descrierea formelor de flexiune

Sectiunea GRAALAN corespunzatoare formelor de flexiune contine câte o intrare pentru fiecare forma de flexiune. O intrare contine: i) Forma de flexiune în alfabet normal si fonetic; ii) Identificarea în lexicon a intrarii careia îi corespunde forma respectiva de flexiune; iii) Caracterizarea formei de flexiune sub forma unui ansamblu de categorii lexicale cu valorile lor (AVT); iv) Despartirea în silabe.

Formele de flexiune nu sunt scrise în general direct în GRAALAN ci se creeaza cu ajutorul unui instrument specializat.

9. Descrierea sintaxei

Sintaxa se descrie în GRAALAN sub forma unei liste de reguli sintactice etichetate (care respecta principiile gramaticilor de dependente generative (Diaconescu, 2002)).

O regula are un membru stâng care contine un neterminal însotit de un AVT format din categorii lexicale si/sau sintactice) si un membrul drept care contine unul sau mai multi alternanti. Un alternant este format din trei subsectiuni:

a) Subsectiunea sintactica care contine o secventa de NTPA: Neteminali, Terminali, Pseudo terminali, Actiuni (subprograme procedurale). Neterminalii si terminalii au

acceptiunea obisnuita. Pseudoterminalii sunt neterminali care, daca ar avea reguli care sa îi descrie, acestea ar contine direct terminali din lexicon. Actiunile sunt subprograme procedurale care ar putea fi utilizate în anumite tratamente specifice daca este cazul. Fiecare NTPA contine un nume, un AVT format din categorii lexicale si/sau sintactice, modul de legare (relationare) cu alti NTPA.

b) Subsectiunea de dependente unde se descriu relatiile de dependenta între NTPA-uri ale alternantului. Relatiile de dependenta pot fi de tip de regenta / subordonare sau de tip coordonare.

c) Subsectiunea de acord care descrie acordul între NTPA-urile alternantului sub forma unor conditii complexe.

Descrierea sintaxei în GRAALAN este reversibila adica poate fi folosita si în procesul de analiza sintactica prin care se genereaza din textul de suprafata un arbore de dependente ca forma de adâncime, si în procesul de generare din arborele de dependente a textului de suprafata.

10. Descrierea corespondentelor bilingve

Sectiunea GRAALAN privitoare la corespondentele bilingve descrie corespondente între urmatoarele tipuri de elemente apartinând la doua limbi diferite:

a) Corespondente între MWE-uri care sunt reprezentate în lexicon sub forma unor arbori de dependente se exprima prin echivalarea între expresia sursa si expresia tinta corespunzatoare dar si prin regulile de transformare care indica modul în care extensiile expresiei sursa din instante reale sunt preluate de expresia tinta.

b) Corespondente între cuvinte. Este un caz particular al corespondentei între MWE-uri în care expresiile echivalate au câte un singur cuvânt.

c) Corespondente între structuri sintactice. Este un caz particular al corespondentei între MWE-uri în care cele doua expresii pot avea drept caracterizari de noduri nu numai categorii lexicale (cu valorile lor) ci si categorii sintactice (cu valorile lor).

d) Corespondente între structuri morfologice. Este un caz particular al corespondentei între MWE-uri în care cel putin expresia sursa corespunde unei forme flexionate analitico-sintetice.

e) Corespondente între subarbori morfologici. Este o corespondenta între diverse seturi de categorii lexicale (cu valorile lor) organizate sub forma unor AVT-uri.

Informatiile din sectiunea de corespondente bilingve GRAALAN se pot folosi în aplicatii de generare a unor dictionare sau în aplicatii de traducere automata.

11. Concluzii

Descrierile de cunostinte lingvistice pot fi formulate direct în GRAALAN sau, în anumite cazuri (cum ar fi de exemplu pentru formele flexionate sau pentru lexicon) pot fi create cu ajutorul unor instrumente (programe) speciale care genereaza text GRAALAN. Textul GRAALAN obtinut pe o cale sau pe alta se compileaza cu un compilator adecvat care traduce textul GRAALAN în XML, creindu-se astfel o Baza de cunostinte lingvistice XML creata prin intermediul GRAALAN va putea fi exploatata într-un mod unitar pentru diverse studii sau pentru elaborarea de aplicatii informatice.

Deoarece textul GRAALAN se realizeaza în mai multe transe, o componenta speciala GRAALAN Link va determina legaturile între aceste transe si compatibilitatea lor.

Un compilator GRAALAN este în curs de implementare si unele cunostinte lingvistice privind limba româna au fost deja scrise în GRAALAN.

Referinte bibliografice

Diaconescu, S. (2002). Natural Language Understanding Using Generative Dependency Grammar, în Max Bramer, Alun Preece and Frans Coenen (Eds), Proceedings of ES2002, Cambridge UK, Springer, pp.439-452.

Diaconescu, S. (2003). Natural Language Agreement Description for Reversible Grammars, în Tamás D. Gedeon, Lance Chun Che Fung (Eds.), Proceedings of AI 2003, Perth, Australia, pp. 161-172.

Diaconescu, S. (2004) Multiword Expression Translation Using Generative Dependency Grammar, în Proceedings of ESTAL 2004 - ESPAÑA for NATURAL LANGUAGE PROCESSING, Alicante, Spain.

Diaconescu, S. (2005). Some Properties of the Attribute Value Trees Used for Linguistic Knowledge Representation, Proceedings of IICAI-05, INDIA.

EAGLES (1996a). Recommendations for the Morphosyntactic Annotation of Corpora.

EAGLES (1996b). Preliminary Recommendations on Subcategorisation.

EAGLES (1993). Lexicon architecture Draft Report, EAG--LSG/IR--T1.1.

IPA (2005) International Phonetic Association (2005): Handbook of IPA.

ISO/IEC 10646 (1992). Information technology -- Universal Multiple-Octet Coded Character Set (UCS).