

EDITOR DE BAZE DE CUNOȘTINȚE

Cristian Popescu

*Institutul de Cercetări în Inteligența Artificială
cristian@racai.ro*

Rezumat: Acest articol a fost structurat în șase secțiuni. Prima secțiune cuprinde o introducere în problematica teoriei psiholingviștilor și apariția WordNet-ului. Este urmată de prezentarea tipurilor de relații din WordNet. Partea a treia cuprinde modalitățile de structurare a datelor de intrare pentru aplicație. În secțiunea a patra, se prezintă modalitățile de transpunere a cunoștințelor în afișarea 3D. Prezentarea aplicației se face în partea a cincea. Această secțiune este urmată de câteva concluzii și referințe bibliografice.

Cuvinte cheie: concepte, tipuri de relații între concepte, modalități de reprezentare a cunoștințelor, modalități de afișare 3D.

1. Introducere

De multe ori, comunicarea este definită generic ca limbaj. Însușirea tipurilor de limbaj cunoscute nu se va încheia niciodată, descoperind mereu noi caracteristici ce pot să deschidă un tip aparte de comunicare. Aceasta denotă diversitatea modalităților de a transmite celor din jur informații dintre cele mai variate.

Dezvoltarea abilității individuale își are la bază limbajul, ca una dintre premisele dezvoltării cunoașterii. Tehnologia informației tinde să transforme societatea într-un „sat global” [2], ajutat de progresul vizibil din telecomunicații, precum și de calculatoare care reduc distanțele, informează aproape instantaneu, fiind factorii principali în comunicație. Viteza de procesare și capacitatea de stocare a actualelor calculatoare cresc cu o viteză de nebanuit, creând o infrastructură solidă în progresul semnificativ tehnologic al procesării limbajelor naturale, în interconectivitate, multimedia și teleservicii [2].

Pe lângă lingvisticile teoretice și aplicate, procesarea limbajului include un număr de discipline întretesute din câmpul de tehnologia informației. Ele împart împreună un scop principal de procesare atât a scrisului, cât și a limbajului uman vorbit. În contextul informatizării societății, limbajul natural devine subiectul investigației tehnice și tehnologice.

Comunicarea înlesnește transmiterea informației între indivizi, caracterul informației fiind perceput corespunzător în funcție de limbajul folosit. Folosirea unui limbaj general cunoscut face ca toate sensurile informației transmise să fie percepute corespunzător. Pentru ca informația transmisă să poată să își atingă scopul, adică să fie deplin înțeleasă, limbajul trebuie procesat. Metodele sunt diferite, de la tehnologiile familiare gen procesare de cuvinte, stocarea și salvarea de text, până la domeniile mai ambițioase precum traduceri automate, recunoașterea vorbirii și recunoașterea scrisului etc. Aceste tehnologii variază în complexitate și sunt, de aceea, în diferite stadii de progres și exploatare comercială.

Ingineria limbajului face parte din centrul tehnologiei informației și tehnologia informației va face parte, la rândul ei, din industria cheie în secolul 21. Superautostrăzile informației, care sunt imaginate în ziua de astăzi, vor transporta, în curând, cantități infinite de date în format digital, imagini, sunete, tabele, figuri, calcule, și protocoale de proces.

Resursele limbajului reprezintă materialul brut al tuturor tehnologiilor limbajelor. Cu cât sunt mai bune, cu atât este mai scumpă crearea lor.

În cadrul comunicării directe, interumane, prin limbaj articulat, sensul și amploarea informației sunt date nu numai de elementul de bază – cuvânt, ci și de ton, mimică, gesturi, context lexical, context informațional, situațional etc. Este știut faptul că cel mai simplu cuvânt afirmativ „yes” („da” în limba engleză) își are multiple sensuri în funcție de ton, mimică, situație etc.

Având în vedere sensurile multiple ale unui element de bază „cuvânt”, în funcție de context, calitatea aplicațiilor tehnologiei limbajului constă, în special, din totalitatea și capacitatea datelor limbajului cu care se creează acele aplicații. De exemplu, pentru aplicații tradiționale precum „corectorul de ortografie”, se folosesc date sub formă brută, forme ale cuvintelor (incluzând variații ortografice) și analize morfo - sintactice.

Procesarea limbajului, în acest context, devine o inginerie din ce în ce mai complicată, atunci când scopul este de a face o traducere liberă, ce trebuie să țină cont de sensuri și nuanțe.

Procedurile standard alfabetice, pentru organizarea informației lexicale, aduc împreună cuvinte care sunt scrise la fel și cuvinte diferite, cu înțeles similar sau apropiat. Acestea sunt puse într-o ordine întâmplătoare într-o listă alfabetică, unde găsirea lucrurilor poate fi obositoare și consumatoare de timp.

Mulți oameni care doresc să caute ceva într-un dicționar decid, în ultimă instanță, să nu se mai deranjeze cu el, deoarece găsirea informației le-ar întrerupe munca și disturba șirul de idei.

Un motiv important pentru orientarea către dicționare on-line – baze de date lexicale, care pot fi citite de către calculatoare – constă în capacitatea calculatoarelor de a căuta într-o listă organizată alfabetic mult mai rapid decât pot oamenii. O intrare într-un dicționar poate fi accesată în același moment în care cuvântul căutat este selectat sau tipărit. Mai mult chiar, deoarece dicționarele sunt tipărite din unități de stocare, citite de către calculatoare, este destul de ușor ca acele baze de date să fie convertite în orice format dorit.

Folosind calculatoarele în acest scop, se constată că este destul de inefficient a utiliza aceste mașini puternice și complexe doar ca să întoarcă paginile foarte rapid. Provocarea este ce putem face ca să le utilizăm mai bine. WordNet-ul este o propunere pentru o mai bună combinație a informației lexicografice tradiționale și computația de mare viteză actuală.

Secolul XX a fost martorul evoluției *psiholisticii*, un domeniu interdisciplinar al cercetării, interesat de baza cognitivă a competentelor lingvistice. Pe parcursul evoluării teoriilor lingvistice de-a lungul câtorva decade, lingviștii au ajuns în mare măsură la concluzia că un lexicon trebuie să conțină în ordine, pe lângă componentele: fonologică, sintactică și lexicală, și înțelesul mesajului lingvistic. Aceste propuneri au fost încorporate în teoria psiholingviștilor.

Începând cu studierea asocierii cuvintelor, la mijlocul secolului, și continuând cu munca experimentelor sofisticate, în ultimii 20 de ani, psiholingviștii au descoperit multe proprietăți sincronice ale lexiconului mental, care pot fi folosite în lexicografie.

Un grup de psihologi și lingviști de la Universitatea Princeton, în anul 1985, au pornit un proiect cu scopul de a realiza o bază de date lexicală de-a lungul liniilor trasate de aceste investigații [3]. Ideea inițială era de a furniza un ajutor în folosirea căutării dicționarelor conceptuale, mai degrabă decât alfabetic. Pe parcursul desfășurării proiectului, totuși, s-a simțit nevoia unei clarificări mai bune a principiilor și scopurilor care stăteau la baza proiectului. WordNet este rezultatul. Se poate spune că WordNet-ul este un dicționar bazat pe principiile psiholingviștilor.

Cum ar fi trebuit ca teoriile psiholingviștilor să ghideze acest proiect, nu a fost niciodată prea clar. Majoritatea domeniului de interes pentru psiholexicologie a fost să lucreze împreună cu mostre relativ mici din lexiconul Englez, de multe ori concentrându-se asupra substantivelor în defavoarea celorlalte părți de vorbire. De prea adesea ori, o ipoteză interesantă este pusă în față, sunt considerate 50 sau 100 de cuvinte care o ilustrează și extensia către restul lexiconului este lăsată ca un exercițiu cititorului [3].

Un motiv pentru crearea WordNet-ului a fost expunerea acestor teorii întregului domeniu al unui vocabular obișnuit. WordNet în prezent conține aproximativ 95600 de forme diferite de cuvinte (51500 cuvinte simple și 44100 colocații) organizate în 70100 de înțelesuri de cuvinte sau seturi de sinonime și doar cele mai robuste ipoteze au supraviețuit [3].

În starea actuală, WordNet-ul conține numai substantive, verbe, adjective, și adverbe. Mulțimea relativ mică de cuvinte funcționale în limba Engleză este ca și inexistentă pe baza presupunerii că vor fi mai mult ca sigur stocate separat, ca parte sintactică a limbajului [3].

2. Tipuri de relații în cadrul WordNet-ului

Relațiile care au fost analizate în cadrul WordNet-ului sunt sinonimia, antonimia, hiponimia și meronimia. În continuare, se va oferi o definiție pentru aceste tipuri de relații.

Făcând referire la o definiție (de obicei, atribuită lui Leibniz) două expresii sunt sinonime dacă substituția uneia pentru cealaltă niciodată nu schimbă valoarea adevărului unei propoziții în care a avut loc substituția. Prin această definiție, sinonimele adevărate sunt rare, dacă ele există cumva cu adevărat.

O versiune mai simplistă a acestei definiții ar face sinonimia relativă la un context: două expresii sunt sinonime într-un context lingvistic C dacă substituția uneia pentru alta în C nu alterează valoarea adevărului. De exemplu, substituția „scândură” pentru „bancă” va modifica destul de puțin valoarea de adevăr într-un context despre lemnărie, totuși sunt alte contexte pentru „bancă” în care substituția ar fi inacceptabilă.

Definirea sinonimiei în termeni de valori de adevăr se pare că face ca sinonimia să pară ca o problemă discretă: două cuvinte fie sunt sinonime sau nu. Dar cum unii filozofi ar putea argumenta și majoritatea psihologilor ar fi de acord fără a considera nici o alternativă, sinonimia este considerată cel mai bine ca un sfârșit al unui continuu de-a lungul căruia similaritatea înțelesului poate fi gradată. Este probabil cazul în care cuvintele similare semantic pot fi interschimbate în mai multe contexte decât cuvintele nesimilare semantic.

O altă relație familiară este antonimia care s-a dovedit să fie surprinzător de dificil de definit. Un antonim al unui cuvânt x este câteodată $\text{not-}x$, dar nu întotdeauna. De exemplu, „bogat” și „sărac” sunt antonime, dar să spui că cineva „nu este bogat” nu presupune că este automat „sărac”. Mulți oameni se consideră nici bogați, nici săraci. Antonimia care pare să fie doar o simplă relație simetrică, este destul de complexă. Totuși, vorbitorii nativi ai unei limbi pot destul de ușor să recunoască antonime când le întâlnesc.

Antonimia este o relație lexicală între formele cuvântului, nu o relație semantică între înțelesurile cuvântului. De exemplu, înțelesul {răsare, ascede} și {cade, coboară} pot să fie opuse conceptual, dar nu sunt antonime; {răsare, coboară} sunt antonime la fel ca și {ascede, cade}, dar majoritatea oamenilor ezită și se uită bănuitori când sunt întrebați dacă „răsare”, „coboară” sau „ascede” și „cade” sunt antonime. Aceste fapte fac aparent nevoia pentru distingerea între relațiile semantice dintre formele cuvântului și relațiile semantice dintre înțelesurile cuvântului.

Antonimele furnizează un principiu de organizare centrală pentru adjective și adverbe în WordNet, și complicațiile care apar din faptul că antonimia este o relație semantică între cuvinte, sunt mai bine discutate în acest context.

Spre deosebire de sinonimie și antonimie care sunt relații lexicale între formele cuvântului, hiponimia/hipernimia este o relație semantică între înțelesurile cuvântului.

De exemplu „arțar” este un hiponim al lui „pom” și „pom” este un hiponim al lui „plantă”. Multă atenție a fost acordată hiponimiei/hipernimiei (denumită și subordonare/superordonare, subset/superset sau relație ISA).

Tipurile de reprezentări ierarhice sunt folosite pe scară largă în construcția sistemului de regăsire a informațiilor („*information retrieval system*”), în care sunt denumite sisteme cu moștenire: un hiponim moștenește toate calitățile a mai multor concepte generice și adaugă cel puțin o caracteristică pentru a face distincția față de hipernim și față de alte hiponime ale hipernimului respectiv. Această convenție furnizează principiul de organizare centrală a substantivelor în WordNet.

Sinonimia, antonimia, și hiponimia sunt relații familiare. Ele se aplică pe scară largă în tot lexiconul, iar oamenii nu consideră necesară o pregătire specială în lingvistică pentru a le putea înțelege.

O altă relație care folosește aceste avantaje – relația semantică – este relația parte-întreg (sau HAS-A), cunoscută semanticilor lexicale sub denumirea de meronimie/holonimie. Un concept reprezentat de synsetul $\{x, \xi, \dots\}$ este o meronimie a conceptului reprezentat de synsetul $\{y, \psi, \dots\}$ dacă vorbitorii nativi ai unei limbi, acceptă structuri de propoziții din aceste bucăți precum: „Un y are un x (ca o parte)” sau „Un x este parte a lui y ”.

Relația de meronimie este tranzitivă (cu calificări) și asimetrică, și poate fi folosită la construirea unei părți în mod ierarhic (cu câteva rezervări, ținând cont că o meronimie poate avea multe holonime).

Aceasta și alte relații similare servesc organizării lexiconului mental. Ele pot fi reprezentate în WordNet prin grupări părintești sau prin săgeți de direcție (pointeri), arce numerotate, de la un synset către altul. Aceste relații reprezintă asociații care formează o rețea complexă; locul unde este situat un cuvânt în această rețea reprezintă un pas important în aflarea înțelesului cuvântului.

3. Reprezentarea formatului datelor de intrare

Folosind MFC și OpenGL, a fost dezvoltată o aplicație de vizualizare a datelor și a tipurilor de relații din cadrul WordNet-ului. De asemenea, aplicația permite și crearea de relații noi, precum și de dicționare lexicale gen WordNet. Aplicația permite atât primirea unei baze de date din fișiere, cât și construirea ei treptată de către utilizator. Baza de date din fișier(e) este structurată în format RDF[4] sau XML[1]. Ca date de intrare în cadrul aplicației, au fost folosite fișierele de definire a WordNet-ului. Acestea sunt:

- *wordnet_glossary-20010201.rdf* care conține definirea și descrierea conceptelor;
- *wordnet_hyponyms-20010201.rdf* care conține definirea legăturilor de tip hiponim între conceptele din fișierul *wordnet_glossary-20010201.rdf*;
- *wordnet_similar-20010201.rdf* care conține definirea legăturilor de tip similar între conceptele din fișierul *wordnet_glossary-20010201.rdf*;
- *wordnet_nouns-20010201.rdf* care conține mulțimea cuvintelor din dicționarul Englez care aparțin conceptelor definite în fișierul *wordnet_glossary-20010201.rdf*, precum și părțile de vorbire sub care pot apare.

Mai există un fișier *wordnet-20000620.rdf*s (informația în acest fișier este structurată în format RDF-SCHEMA), care conține definirea structurilor care apar în cadrul WordNet-ului.

Modelul structurării bazei de date în fișiere sub format RDF a fost preluat în cadrul aplicației. Se observă că pentru noduri avem un fișier separat. Pentru fiecare tip de legătură avem câte un fișier. Preluând acest model, la încărcarea bazei de date, se va selecta întâi fișierul care conține nodurile structurate în format RDF și apoi fișierele care vor conține legăturile dintre noduri.

Salvarea bazei de date se va face asemănător politicii structurării WordNet-ului în fișiere RDF/XML[8]: întâi se alege un fișier în care vor fi salvate conceptele din baza de date, împreună cu definirea și descrierea lor. Apoi pentru fiecare tip de legătură se va selecta câte un fișier în care vor fi salvate legăturile de acel tip ale fiecărui concept cu celelalte concepte din baza de cunoștințe. Fiecare structură gen (concept, mulțimea legăturilor cu celelalte concepte), din cadrul fișierelor care conțin legăturile structurate în format RDF [6] (la XML diferența apare prin nespecificarea tag - urilor caracteristice RDF-ului) este definită în felul următor:

```
<rdf:Description rdf:about="&a;
    indice_concept_de_la_in_cadrul_bazei_de_date">
  <b:TipLegatura rdf:resource="&a;
    indice_concept1_catre_in_cadrul_bazei_de_date"/>
  <b:TipLegatura rdf:resource="&a;
    indice_concept2_catre_in_cadrul_bazei_de_date"/>
  .....
  <b:TipLegatura rdf:resource="&a;
    indice_conceptn_catre_in_cadrul_bazei_de_date"/>
</rdf:Description>
```

Formatele XML[5] și RDF sunt foarte des folosite datorită ușurinței cu care se pot structura informațiile într-un fișier care poate fi accesibil pe Internet, și nu numai. Atât RDF-ul, cât și XML-ul [9] se bazează pe un model matematic. Acest model matematic furnizează un mecanism pentru gruparea seturilor de blocuri de metadate.

Dorindu-se transformarea societății umane într-un „sat global” [2], informația trebuie să fie accesibilă oricărei persoane care aplică pentru obținerea ei. Din acest motiv, orice bază de date trebuie structurată într-o formă ușor accesibilă, de oriunde, prin intermediul Internetului, și nu numai.

Relațiile de hiponimie și hiperonimie pot fi vizualizate separat, în cadrul aplicației, datorită caracterului lor special. Dacă avem două concepte A și B și o relație de hiponimie de la A la B, atunci hiperonimia ar fi relația de la B la A. Prin relația de hiponimie de la A la B, conceptul A este inferior (subordonat) conceptului B. Hiperonimia ar fi relația de la B la A, deci conceptul B este superior conceptului A.

Vom reprezenta mai jos relația de hiperonimie și de hiponimie între concepte. Vom lua conceptele din WordNet ca exemplu. S-a considerat reprezentarea conceptelor sub forma unor sfere, iar legăturile dintre concepte, sub forma arcelor care unesc sferele. Astfel, întreaga bază de cunoștințe pe care o reprezintă WordNet-ul poate fi reprezentată sub forma unei tabelă hash sau a unui graf [10].

Dăm următorul exemplu de relații de tip hiperonimie și hiponimie între concepte (figura 1):

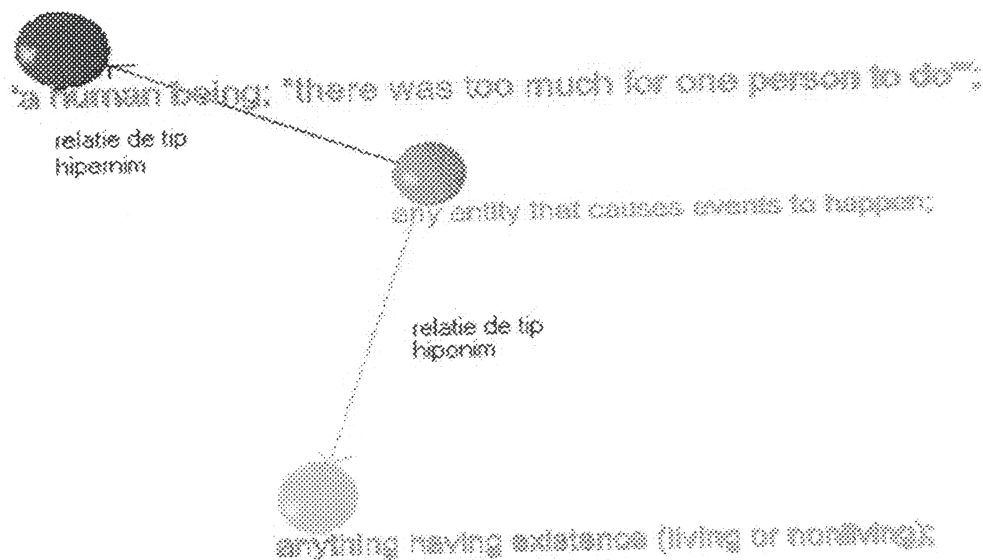


Figura 1. Reprezentarea relației de hiperonimie și de hiponimie

Sfera reprezintă conceptul, iar textul de sub sferă reprezintă descrierea conceptului respectiv. Săgețile care ies din sferă reprezintă legăturile pe care le are conceptul principal cu celelalte concepte, în cadrul bazei de date. În dreptul fiecărei legături, s-a precizat și tipul ei. S-a ales acest exemplu pentru a se putea înțelege mai bine ce reprezintă relațiile de hiponimie și hiperonimie între concepte.

În fișierele care conțin baza de date WordNet în format RDF/XML, apare doar relația de hiponimie. Relația de hiperonimie nu există datorită legăturii dintre relații: dacă A este hiponim pentru B, atunci B este hipernim pentru A. S-a considerat că este necesară introducerea relației de hiperonimie doar atunci când utilizatorul o dorește. Astfel, utilizatorul își poate alege dintre relațiile din baza de date, relația de hiponimie pentru a se genera relația inversă între concepte, de hiperonimie.

4. Transpunerea cunoștințelor în modul 3D

Pentru a putea reprezenta mult mai bine cunoștințele din baza de date, s-a folosit, în cadrul aplicației, interfața OpenGL [7], care este o interfață software pentru reprezentarea diferitelor obiecte. Această interfață consistă din aproximativ 150 de comenzi distincte, cu ajutorul cărora se pot specifica obiecte și operații necesare pentru a furniza utilizatorului aplicații realizate 3D.

OpenGL a fost proiectat ca o interfață independentă de tehnologia hardware. Acest lucru îi oferă posibilitatea portabilității pe diferite platforme hardware. OpenGL nu furnizează comenzi de nivel înalt pentru descrierea modelelor obiectelor tridimensionale (precum părți ale corpului, avioane). Cu ajutorul OpenGL-ului trebuie construit modelul dorit, pornind de la un set mic, de primitive geometrice, precum puncte, linii, și poligoane.

O bibliotecă sofisticată, care furnizează aceste utilitare, a fost construită în vârful OpenGL-ului. Biblioteca de Utilitare OpenGL (OpenGL Utility Library – GLU) furnizează multe dintre utilitarele de modelare, precum suprafețele cvadrice și curbele NURBS. GLU este o parte standard, a oricărei implementări OpenGL.

Fiecare concept din cadrul bazei de date va fi reprezentat sub forma unei sfere. Fiecare legătură între concepte va fi, atunci, reprezentată printr-un arc care va lega sferele.

Pentru o mai bună înțelegere a imaginii de pe ecran, utilizatorului i se pune la dispoziție posibilitatea de a naviga printre conceptele de pe ecran. Astfel utilizatorul poate roti imaginea după cele trei axe: x, y și z (figura 2). Deoarece s-a încercat ca observatorul să se miște în jurul imaginii obiectelor de pe ecran, atunci, la folosirea cursorilor, se observa pseudomișcarea obiectelor față de observator, în sens contrar.

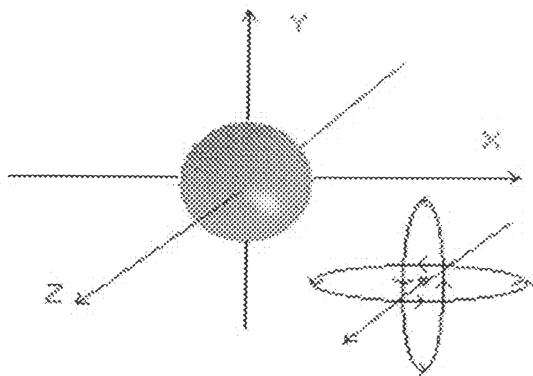


Figura 2. Mișcarea după axe x, y și z

La afișare, se va reprezenta mai întâi conceptul principal. Apoi se reprezintă conceptele cu care conceptul principal are legături, fiecare concept va fi reprezentat la rândul lui prin intermediul unei sfere. Fiecare concept este reprezentat la distanța raza sferei, față de conceptul principal. Acesta din urmă, conceptul principal, se află reprezentat în centrul sferei.

Sub fiecare concept reprezentat, este trecută descrierea lui. Inițial s-a considerat că trebuie reprezentat fiecare concept împreună cu indicele conceptului din baza de cunoștințe. În final, s-a ajuns la concluzia că acest lucru nu era deloc edificator pentru caracterizarea conceptului, astfel, utilizatorul trebuia de fiecare dată să țină minte sau să caute descrierea conceptului cu identificatorul respectiv. Astfel, este mai bine ca fiecare concept, când este reprezentat, să fie pus în pereche cu descrierea lui.

Informațiile cu privire la fiecare concept se pot afla selectând conceptul din imaginea 3D. Astfel, apare o fereastră în cadrul aplicației, care nu modifică derularea acesteia, în care apar informații cu privire la conceptul selectat la acel moment (figura 3).

Pentru generarea punctelor pe o sferă, s-a folosit următorul algoritm. Întâi, se generează un număr finit de puncte pe sferă, iar apoi sunt dispuse sferile (conceptele) în acele puncte prestabilite. Pentru că algoritmul este destul de laborios, mulțimea punctelor posibile de pe sferă, care au posibilitatea să fie ocupate în viitor de sfere este finită. Mulțimea punctelor este ținută sub forma unui vector. Această mulțime a punctelor este generată la început, o dată cu pornirea aplicației. De fiecare dată când se selectează un concept nou pentru a fi reprezentat împreună cu conceptele, din baza de cunoștințe, cu care are legături, se ocupă triplete (x, y, z) din mulțimea de puncte disponibile pe sferă, pentru a fi ocupate.

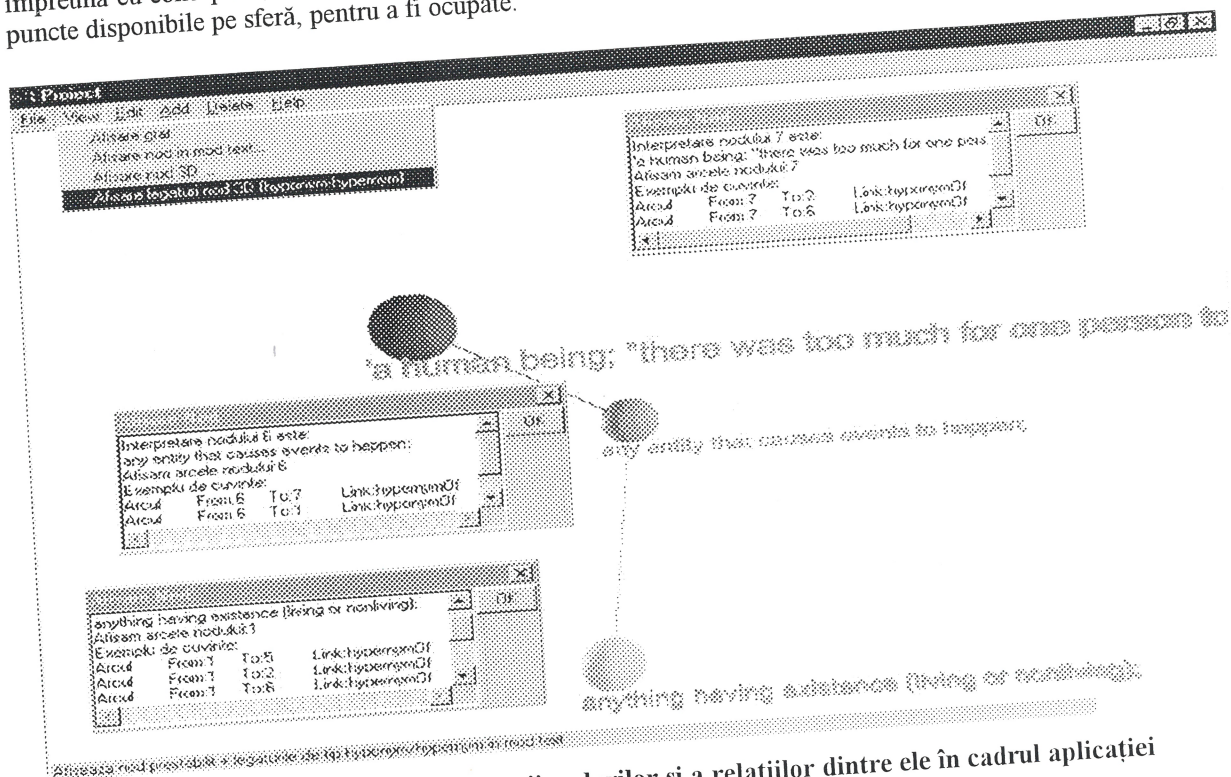


Figura 3. Prezentarea vizualizării nodurilor și a relațiilor dintre ele în cadrul aplicației

5. Prezentarea aplicației

Această aplicație a fost dezvoltată în Visual C. Rulează pe platforme Windows sau compatibile Windows. Această aplicație are o interfață prietenoasă. Bara de meniu este structurată după modelul Windows. Acest aspect oferă utilizatorului o ușurință în folosirea funcțiilor din bara de meniuri, datorită familiarității cu mediul Windows.

Meniul este structurat pe șase grupuri de funcții: File, View, Edit, Add, Delete, Help. Fiecare dintre acest grup conține mai multe funcții cu proprietăți asemănătoare. Funcțiile sunt grupate după parametrii pe care-i primesc și după rezultatul obținut. Există o fereastră de vizualizare a conceptelor. Fereastra conține imagini 3D. În această fereastră de vizualizare, se pot observa particularitățile diferitelor concepte, legăturile lor cu conceptele din baza de cunoștințe.

Conceptele sunt reprezentate în această fereastră, prin intermediul unor sfere. Fiecare sferă dispune de anumite particularități, în primul rând prin conceptul pe care-l reprezintă.

Scena este privită de către utilizator ca fiind un observator în afara acesteia. Utilizatorul poate manevra scena astfel încât să observe o mai bună conturare a conceptelor și a relațiilor dintre ele. El poate roti scena, după axele x și y; poate să se apropie de scenă sau să se îndepărteze.

Obiectele care apar în scena 3D pot să fie selectate, această selectare făcându-se cu ajutorul mouse-ului. În caz că acolo în imaginea 3D există un obiect, aplicația trece la identificarea lui, prin intermediul unui algoritm prezentat în [7]. Pe ecran, apare o fereastră cu descrierea și toate legăturile conceptului identificat prin acel obiect (în cazul nostru o sferă).

Această aplicație dispune și de un mediu de ajutor și informații cu privire la modalitățile de folosire ale aplicației (modul Help) pentru ca aplicația să fie accesibilă oricărui utilizator. În acest mediu, sunt descrise toate funcțiile din meniu, precum și modul lor de folosire și ce rezultate produc.

Aplicația poate primi ca date de intrare fișiere în format RDF sau XML. Aceste fișiere sunt pasate și din ele se extrag date/informații cu ajutorul cărora este construită baza de date. Această bază de date este ținută sub forma unui graf, asemănător unui tabel hash. Avem o listă de noduri, iar fiecare concept are la rândul lui o listă de arce, legături către celelalte noduri din listă. Parcurgerea în baza de cunoștințe se realizează la fel ca și parcurgerea într-o listă. Pentru salvarea unui concept cu toate legăturile pe care le are cu celelalte concepte din baza de cunoștințe, este, în schimb, folosită structura de graf și se realizează o parcurgere în adâncime a grafului. Nodurile sunt salvate până se atinge adâncimea de recurență maximă. Fiecare nod din această structură este format din mai multe câmpuri: identificatorul nodului în cadrul bazei de cunoștințe, identificatorul nodului în baza de cunoștințe de unde a fost extrasă informația (fișier RDF/XML), descrierea conceptului care are asociat acel nod din baza de cunoștințe, precum și o listă de arce pe care le are cu celelalte noduri din baza de cunoștințe.

Fiecare arc este identificat prin numărul arcului din cadrul listei de arce asociată fiecărui nod, o descriere a arcului respectiv, precum și o legătură la nodul cu care avem legătura. Descrierea arcului respectiv reprezintă tipul de legătură pe care o are conceptul cu nodul către care este arcul respectiv.

Cuvintele asociate fiecărui concept sunt introduse într-o structură de arbore binar de căutare. Fiecare nod din acest arbore are asociate următoarele atribute: descrierea cuvântului, indicele nodului care reprezintă conceptul din cadrul bazei de cunoștințe, căruia aparține cuvântul. Se poate observa că un cuvânt poate avea mai multe înțelesuri, deci poate aparține mai multor concepte. S-a considerat că, dacă un cuvânt X are înțelesurile A și B, atunci avem două cuvinte (X cu înțelesul A) și (X cu înțelesul B), având două intrări în arborele de cuvinte. S-a ajuns la concluzia aceasta deoarece este soluția optimă. Atunci când păstrăm cuvintele în arbore, pentru a căuta toți indicii nodurilor care reprezintă concepte din baza de cunoștințe și care sunt în pereche cu un singur cuvânt X sau toate cuvintele care aparțin aceluiași concept din baza de cunoștințe.

Această aplicație poate primi atât o bază de cunoștințe în format RDF sau XML, din care să-și extragă cunoștințele sau se poate crea cu ea o nouă bază de cunoștințe. Astfel, aplicația dispune de instrumente pentru a crea concepte noi, care să fie adăugate la baza de cunoștințe. De asemenea, se pot adăuga și legături între aceste concepte. Apoi, baza de cunoștințe nou creată poate fi salvată în fișiere RDF sau XML, ca să fie păstrată pe viitor, în caz că se dorește refolosirea ei.

Aplicația permite concatenarea mai multor baze de cunoștințe într-una singură, sau salvarea bazei de cunoștințe în mai multe segmente. Acest ultim lucru poate fi realizat prin salvarea numai a anumitor concepte și a legăturilor lor cu celelalte concepte din baza de cunoștințe, până la o anumită adâncime.

Așa cum se pot adăuga elemente noi în baza de cunoștințe, aplicația dispune și de posibilitatea eliminării elementelor din baza de cunoștințe. Astfel, se pot elimina concepte/noduri din baza de cunoștințe, întreaga bază de cunoștințe (care este încărcată în memorie) sau doar legături între conceptele din cadrul bazei de cunoștințe. Eliminarea legăturilor dintre conceptele din cadrul bazei de cunoștințe este foarte complexă:

- se poate elimina o legătură precizată pe care o are un anumit concept;
- se pot elimina toate legăturile pe care le are un concept (nod) sursă cu un alt concept(nod) destinație, prestabilit;
- șterge un anumit tip de legătură din baza de date, eliminând acel tip legătură de la fiecare nod/concept din cadrul bazei de date cu celelalte concepte din cadrul bazei de date;
- șterge toate legăturile de un anumit tip pe care le are un concept (nod) sursă cu alte concept(noduri) destinație, prestabilit.

Aplicația permite mai multe tipuri de vizualizare a informației din cadrul bazei de date:

- vizualizare în mod text;
- vizualizare în mod grafic.

În cadrul vizualizării în mod text, utilizatorul poate afla informații atât cu privire modul de ierarhizare a conceptelor în cadrul bazei de date, cât și informații cu privire la fiecare concept din cadrul acesteia. Utilizatorul poate căuta un concept în două moduri: ori prin căutarea în cadrul bazei de date a indicelui conceptului respectiv sau prin căutarea în baza de cuvinte a unui cuvânt care aparține conceptului respectiv, identificarea în cadrul bazei de date a indicelui asociat conceptului și apoi afișarea în mod text a informațiilor referitoare la nodul respectiv.

În cadrul vizualizării în mod 3D, utilizatorului i se pun la dispoziție mai multe modalități de vizualizare a conceptului cerut. Observatorul poate vizualiza toate legăturile pe care acel concept le are cu restul conceptelor

din cadrul bazei de date, numai a unei anumite legături sau numai a legăturilor de tip hiponim sau hipernim dintre conceptul respectiv și restul conceptelor.

În cadrul vizualizării 3D, conceptele sunt reprezentate prin sfere, iar legăturile dintre concepte sunt reprezentate prin linii care unesc sferele. Fiecare sferă are un text asociat sub ea, care reprezintă descrierea conceptului respectiv. Acest lucru ușurează înțelegerea conceptului și a tipurilor de legături pe care le are conceptul respectiv cu celelalte concepte din cadrul bazei de cunoștințe. Fiecare concept care apare în fereastra de afișare imagine, poate fi identificat și are asociată o fereastră cu descrierea conceptului, precum și tipurile de legături pe care le are conceptul respectiv cu restul conceptelor.

6. Concluzii

Tendința tehnologiei actuale, a transferului de informații, a comunicației, este de a transforma societatea într-un „sat global” [2]. Progresul tehnologic precum și accesul la informație face ca granițele fizice dintre diferitele națiuni să dispară. Tendința este să se realizeze o comunitate globală. Limbajul uman reprezintă încă o modalitate de a delimita vorbitorii unei limbi. Totuși, cu ajutorul calculatoarelor și datorită puterii lor de procesare a informațiilor, această barieră tinde să dispară.

Comunicația a reprezentat cheia secolului XX și va juca un rol foarte important și în secolul XXI. Datorită comunicației, serviciilor de comunicație, oamenii au schimbat idei în mod liber. Comunicația și informația au dus la căderea regimurilor totalitariste din Europa de Est. Toată lumea din ziua de astăzi dorește să fie informată. În fiecare zi suntem bombardați cu informații care sosesc de pe tot Pământul. Limba Engleză a devenit un standard pentru comunicația din ziua de astăzi. Internetul a devenit un mediu unde poți să te informezi și să schimbi idei cu persoane care se află la mii de kilometri depărtare.

Comunicația pe Internet, la ora actuală, se face aproape instantaneu. Se încearcă transformarea calculatorului dintr-un instrument folosit, într-un partener pentru om, mai ales pentru a-i lărgi orizonturile și serviciile pe care le poate oferi. Din acest motiv se încearcă o „traducere” a limbajului uman, într-o „formă matematică” astfel încât să poată fi procesat mai bine de către calculator. Cuvintele nu mai sunt grupate ca în dicționarele standard, în ordine alfabetică. Se încearcă o grupare a lor în jurul unor concepte, găsirea unei structurări a limbajului uman, mai apropiată de matematică, deci de „limbajul” folosit de către calculatoare.

Aplicația descrisă în această lucrare preia ideea pe care este constituit WordNet-ul. Această aplicație este de fapt un editor de baze de cunoștințe de tip WordNet. Aplicația poate fi utilizată ca un instrument de învățare, de observare, folosind baza de cunoștințe WordNet; precum și ca un instrument de dezvoltare a bazelor de cunoștințe. Putând să primească la intrare și să ofere la ieșire fișiere în format RDF sau XML, această aplicație poate fi utilizată în crearea unor baze de cunoștințe, care apoi să fie folosite pe Internet.

Bibliografie

1. **ST. LAURENT, S.:** XML: A Primer, IDG Books Worldwide, Inc., 1998.
2. **TUFIȘ, D., P. ANDERSEN:** Recent Advances in Romanian Language Technology, Editura Academiei Române, București, 1997.
3. **MILLER, G. A., R. BECKWITH, C. FELLBAUM, D. GROSS, K. MILLER:** Introduction to WordNet: An On-line Lexical Database, 1993.
4. * * *: W3C RDF Model and Syntax Specification (Working Draft) web: <http://www.w3.org/TR/WD-rdf-syntax/>
5. * * *: World Wide Web Consortium Recommendation, Extensible Markup Language (XML) 1.0, web: <http://www.w3.org/TR/REC-xml>
6. * * *: W3C Metadata Activity, web: <http://www.w3.org/Metadata/>
7. * * *: Waite group Press, Open GL Super Bible, 1996, Macmillan Computer Publishing
8. **SMITH, N. E.:** Practical Guide to SGML/XML Filters, Wordware Publishing, Inc., 1998.
9. **LIGHT, R.:** Presenting XML, 1997, Macmillan Computer Publishing.
10. **CRISTEA, V., I. ATHANASIU, E. KALISZ, V. IORGA:** Tehnici de Programare, Editura Teora, București, 1998.