

**UNIVERSITATEA „ALEXANDRU IOAN CUZA”  
IAȘI  
FACULTATEA DE INFORMATICĂ**

**LUCRARE DE LICENȚĂ**

**Temporalitate și referențialitate utilizând teoria nervurilor**

**Îndrumător științific:  
prof. dr. Dan Cristea  
asist. drd. Corina Forăscu**

**Student:  
Alistar Elvis**

**Iași  
Iunie 2008**

## Abstract

Multe aplicații pentru procesarea limbajului natural, cum ar fi extragerea de informații (IE – *Information Extraction*), sisteme Întrebare-Răspuns (QA – *Question-Answering*), detectarea și urmărirea subiectelor principale (TDT – *Topic Detection and Tracking*), ar avea performanțe crescute dacă ar exista posibilitatea de a poziționa cu acuratețe evenimente în timp, fie relativ la celelalte evenimente, fie în mod absolut prin intermediul timpului calendaristic. În ultimii ani cercările în domeniul recunoașterii, extragerii și prelucrării informației temporale au cunoscut o dezvoltare remarcabilă ([Mani *et al.*, 2005a] pentru o colecție a celor mai frecvent citate articole).

Teoria nervurilor [Cristea, Ide și Romary, 1998] reprezintă o nouă abordare în privința parsării și prelucrării discursului, care vine să completeze și să îmbunătățească teorii și metode deja existente, cum ar fi teoria centralității și teoria structurii retorice.

Această lucrare descrie o analiză a modului în care extragerea informațiilor temporale din text poate fi îmbinată cu teoria nervurilor. Am creat un corpus de articole extrase din Wall Street Journal, care au fost adnotate automat pentru temporalitate și nervuri. Am arătat că această adnotare este incompletă și conține inconsistențe. Am demonstrat că teoria nervurilor aduce îmbunătățiri semnificative unei astfel de adnotări temporale, venind în sprijinul cercetătorilor care doresc să obțină rezultate excelente prin preprocesarea automată a textului. Am adus îmbunătățiri instrumentului utilizat pentru adnotarea automată a temporalității, obținând adnotări cu o acuratețe de peste 92%. Evaluarea întregului proces de adnotare s-a realizat utilizând rezultatele obținute în urma adnotării manuale a unor texte din corpusul propus. Am dezvoltat o aplicație care, folosind teoria nervurilor, să determine legăturile temporale dintre evenimentele unui text. Am demonstrat astfel că, utilizând teoria nervurilor, pot fi găsite legături temporale, între evenimentele unui text, pe care sistemele actuale de adnotare automată sau chiar adnotatorii umani nu le pot găsi.

Sistemul descris utilizează marcatori temporali specifici din textele în limbaj natural, precum și proprietăți ale discursului date de coerența și coeziunea sa. Sistemul poate fi îmbunătățit prin scrierea unui program care să poată ordona în timp relațiile între evenimente găsite cu ajutorul teoriei nervurilor.

## Cuprins

Abstract.....	2
Cuprins.....	3
1. Introducere .....	5
1.1. Motivație .....	6
1.2. Obiective .....	6
1.3. Conținutul lucrării .....	7
2. Teoria nervurilor și temporalitate.....	8
2.1. Elemente introductive.....	8
2.2. Teoria nervurilor .....	9
2.2.1. Originea teoriei.....	9
2.2.2. Descrierea teoriei .....	11
2.3. Temporalitate .....	15
2.3.1. Istoric .....	15
2.3.2. TimeML .....	16
2.3.2.1 Expresii temporale .....	17
2.3.2.2. Tagul EVENT.....	20
2.3.2.3. Tagurile de legături LINK .....	24
2.3.2.3.1. Legături temporale: TLINK.....	24
2.3.2.3.2. Legături de subordonare: SLINK.....	26
2.3.2.3.3 Legături aspectuale: ALINK.....	27
2.3.2.4. Tagul MAKEINSTANCE .....	22
2.3.2.5. Tagul SIGNAL .....	23
3. Corpusul de texte .....	28
3.1. Obținerea nervurilor .....	29
3.2. Obținerea adnotării pentru temporalitate.....	30
3.3. Obținerea corpusului final .....	32
4. Analiza temporalității în relație cu teoria nervurilor.....	36
4.1. Probleme în procesul adnotării.....	37
4.2. Marcarea tagului SIGNAL .....	39
4.3. Închiderea tranzitivă a temporalității .....	41
4.4. Distanța medie între legăturile temporale .....	45
5. Concluzii .....	47
5.1. Contribuții .....	47
5.2. Probleme nerezolvate.....	47
Bibliografie .....	49
Anexa 1.....	51

## Index figuri și tabele

<b>Fig. 1.</b> Calcularea expresiilor nervură.....	13
<b>Fig. 2.</b> Reprezentarea nervurilor pe arborele de parsare .....	14
<b>Fig. 3.</b> Reprezentare succintă a procesului de obținere a corpusului de text. ....	29
<b>Fig. 4.</b> Arhitectura utilitarului pentru adnotarea automată a temporalității, TARSQI.....	31
<b>Fig. 5.</b> Captură de ecran a spațiului de lucru din Tango.....	43
<b>Fig. 6.</b> Cele 13 relații de bază din algebra lui Allen .....	43
<b>Tabel 1.</b> Relații RST împreună cu sensul lor pentru un nucleu sau un satelit.....	10
<b>Tabel 2.</b> Statistici obținute pe corpus fără LinkMerger.....	34
<b>Tabel 3.</b> Statistici obținute pe corpus cu LinkMerger .....	34
<b>Tabel 4.</b> Statistici privind distribuția TLINKurilor în funcție de atributul relType.....	35
<b>Tabel 5.</b> Comparatie între o adnotare TimeBank și una WSJ .....	38
<b>Tabel 6.</b> Paralelă între numărul de taguri SIGNAL pentru .....	40
<b>Tabel 7.</b> Maparea relațiilor din TimeML la algebra lui Allen.....	44
<b>Tabel 8.</b> Analiza închiderii temporalității .....	45
<b>Tabel 9.</b> Distanța medie între legăturile temporale .....	45
<b>Tabel 10.</b> Distanța între legături pentru documente de mărimi diferite din corpusul WSJ.....	46

## 1. Introducere

Articolele de știri prezintă, de obicei, întâmplări care se dezvoltă de-a lungul timpului. Evenimentele și momentele în timp când s-au produs acestea sunt introduse rând pe rând, iar cititorul înțelege care este ordinea corectă în care s-au desfășurat lucrurile. Întrebări simple, cum ar fi: „Când au început Jocurile Olimpice din Beijing?” pot primi răspuns doar dacă sunt disponibile informații despre evenimente și relațiile temporale dintre acestea. Un document trebuie adnotat manual sau automat pentru a oferi aceste informații.

În ultimii ani s-au făcut cercetări majore în ceea ce privește extragerea evenimentelor, extragerea expresiilor temporale și în privința ancorării și ordonării acestora unele față de altele. Un pas înainte, important în acest domeniu, îl constituie crearea limbajului TimeML [Ingria și Pustejovsky, 2002], care permite analiza detaliată a temporalității.

În ultimii 30 de ani s-au făcut multe cercetări pentru a înțelege ce caracteristici are un text considerat a fi un discurs [Saurí *et al.*, 2006]. Aceste studii s-au axat în mare parte pe structura de discurs și pe relațiile care există între structura discursului și referențialitate. Grosz și Sidner în Teoria Stărilor Atenționale (AST – *Attentional State Theory*) [Grosz și Sidner, 1986] propun o structură segmentală recursivă a discursului, care se bazează pe o reprezentare de tip arbore (rezultatul considerării a două relații între segmentele de discurs: dominanță și satisfacție-precedență). În Teoria Structurii Retorice (RST – *Rhetorical Structure Theory*) a lui Mann și Thompson [Mann și Thompson, 1988] accentul se mută înspre performanța retorică: în ce moduri poate un scriitor (orator) să convingă un cititor (ascultător) să accepte intențiile comunicate. Discursul este reprezentat ca un arbore unde nodurile terminale sunt clauze sau structuri elementare de discurs, nodurile de pe nivele intermediare reprezintă relații (retorice) între fragmente de text, iar coordonarea și subordonarea elementelor componente este similară cu cea a structurilor sintactice. O altă teorie importantă care trebuie luată în considerare este Teoria Centrelor (CT – *Centering Theory*) [Grosz, Joshi și Weinstein, 1995]. Aceasta oferă o explicație convingătoare asupra a ceea ce face un discurs să fie coerent.

Folosind noțiunea de nuclearitate din RST, Teoria Nervurilor [Cristea, Ide și Romary, 1998] descoperă o structură „ascunsă” în arborele de discurs numită *nervură*, care permite determinarea *domeniului de accesibilitate referențială* pentru fiecare unitate de discurs. Teoria nervurilor oferă o explicație care integrează punctele comune ale celor trei teorii prezentate mai sus, corectând în același timp câteva presupuneri AST cu privire la domeniile de accesibilitate și generalizând Teoria Centrelor de la un discurs local la unul global.

## 1.1. Motivație

Pentru a studia evenimentele dintr-un discurs în relație cu ordonarea lor în timp avem nevoie de o adnotare completă și consistentă a textului. Dezvoltarea limbajului TimeML permite o adnotare parțială pentru temporalitate a textelor. Teoria Nervurilor prezintă caracteristici care promit să îmbunătățească acest tip de adnotare. Motivația principală a acestei lucrări se bazează pe o argumentație în patru puncte:

1. O adnotare temporală explicită este necesară în aplicațiile de procesare a limbajului natural cum ar fi sisteme întrebare-răspuns sau sisteme automate de rezumare a textului;
2. Adnotările temporale automate dezvoltate până în prezent au o acuratețe mult perfectibilă;
3. Trebuie să ne bazăm pe adnotarea manuală, dar aceasta este dificilă și nu ne putem aștepta ca rezultatele obținute să fie complete și consistente;
4. Soluția este să observăm cum putem îmbunătăți cât mai mult adnotarea automată pentru a reduce din timpul și munca necesare unui adnotator uman pentru a obține rezultatele dorite. Adnotarea manuală se acceptă doar pentru crearea de resurse pe baza cărora să se construiască apoi instrumentele automate.

## 1.2. Obiective

Adnotarea temporalității face parte din aria mai largă a interpretării temporale a limbajului natural. În acest context, adnotarea temporală reprezintă o încercare de a captura informațiile temporale din texte. Așa cum a fost menționat mai înainte, această sarcină este dificil de realizat [Pustejovsky *et al.*, 2002], nu numai datorită densității și a complexității, dar și datorită lipsei de claritate la anumite nivele. De exemplu, când ne gândim la adnotarea temporală apar următoarele întrebări:

1. Care sunt evenimentele ce ar trebui selectate dintr-un text?
2. Cât de precise trebuie și pot să fie relațiile temporale între evenimente?
3. Ce relații temporale din toate cele care sunt posibile ar trebui adnotate?

Din fericire, timpul este un domeniu bine structurat și o mașină poate ajuta adnotatorul uman să îndeplinească mai bine sarcina adnotării temporale a unui text.

Această lucrare propune analiza temporalității în relație cu Teoria Nervurilor urmărind o serie de pași pentru a demonstra validitatea câtorva presupuneri [după crearea unui corpus de texte (articole de ziar) adnotate atât pentru temporalitate, cât și pentru nervuri]:

- un discurs are o structură bine definită, iar relațiile temporale pot fi studiate în relație cu această structură;

- nervurile pot corecta erorile apărute la adnotarea automată pentru temporalitate a unui text;

- nervurile pot identifica relații temporale între evenimentele care nu au fost descoperite la adnotarea manuală sau automată a unui text;

- închiderea tranzitivă a relațiilor temporale poate asigura consistența adnotării, iar în corelare cu teoria nervurilor poate asigura chiar completitudinea.

Pentru obținerea corpusului pe care a fost realizat studiul descris în această teză a fost utilizat instrumentul de adnotare automată pentru temporalitate TARSQI [Mani *et al.*, 2005a]. Tagul SIGNAL (descriș în secțiunea 2.3.2.5.) este o componentă importantă a limbajului TimeML, dar TARSQI nu marchează acest tag. Am creat un program automat care să adauge fișierelor existente în corpus și marcatorul SIGNAL. A fost utilizat, de asemenea, un modul care să calculeze nervurile pentru un text adnotat pentru RST.

### **1.3. Conținutul lucrării**

Lucrarea este structurată în patru părți. Capitolul 2 prezintă fundamentele teoretice ale Teoriei Nervurilor, iar apoi detaliază limbajul TimeML, folosit exclusiv în adnotările automate realizate pe corpusul propus (185 de articole selectate din publicația Wall Street Journal). Capitolul 3 prezintă motivația alegerii corpusului amintit împreună cu toți pașii care au dus la transformarea în forma lui actuală, formă care conține atât adnotări pentru temporalitate, cât și pentru nervuri. Capitolul 4 descrie detaliat principiile, metodele utilizate și programele implementate pentru a atinge scopurile propuse în lucrare. Capitolul 5 conține rezultate, concluzii, contribuțiile autorului, probleme deschise și posibile moduri de a continua cercetarea în acest domeniu.

## 2. Teoria nervurilor și temporalitate

### 2.1. Elemente introductive

Discursul este definit ca orice mesaj (text sau comunicare verbală) ce este interpretat și înțeles de un om sau de un sistem automat.

Din definiție se observă deja o primă proprietate importantă a unui discurs, și anume **coerența**. Un discurs coerent se compune din elemente strâns legate (și armonizate) între ele. Nici un text nu este coerent decât dacă există și un înțeles în spatele lui. Această condiție esențială a discursului este punctul de plecare pentru cercetare: dacă un text are semnificație, putem presupune că el trebuie să aibă o anumită **structură**, un anumit mod de construcție ce îl face inteligibil, ce îi dă o semnificație mai bogată decât cea a simplei alăturări întâmplătoare de cuvinte și propoziții. Un text este structurat în cuvinte, propoziții, fraze, paragrafe sau alte unități textuale. Coerența este reprezentată în termeni de relații între segmente de text, cum ar fi *elaborarea, cauza sau explicarea* [Mani, 2001]. Pentru a ilustra proprietatea de coerență, considerăm textul: *Ionel a ăzut și și -a spart ochelarii*. Evenimentul *a ăzut* este *cauza* evenimentului *și-a spart*, deoarece a creat condițiile necesare pentru producerea celui din urmă.

O altă presupunere esențială asupra discursului este aceea că există relații între elementele componente ale discursului, relații ce dau discursului proprietatea de **coeziune** și au o contribuție semnificativă la coerența textului. Coeziunea reprezintă calitatea unui discurs (text) de a fi bine format în sensul unității lui interne, făcându-l să „se lege”. Propozițiile se completează ușor una pe cealaltă în cadrul discursului. Există relații interpropoziționale potrivite și marcate fie explicit, fie implicit. Pentru a exemplifica proprietatea de coeziune, considerăm mesajul de pe un indicator rutier: *Reduceți viteza! Ea e cauza multor accidente*.

Coeziunea se realizează prin pronumele din propoziția a doua (*Ea*) care referă un element introdus în prima propoziție (*viteza*).

## 2.2. Teoria nervurilor

### 2.2.1. Originea teoriei

Plecând de la ideile prezentate anterior, Mann și Thompson [Mann și Thompson, 1988] elaborează și descriu Teoria Structurilor Retorice (Rhetorical Structure Theory – RST). Această teorie a devenit una dintre cele mai populare printre lingviști, fiind fie acceptată ca atare, fie folosită ca punct de plecare pentru teorii ulterioare.

Ideea centrală a RST este noțiunea de **relație retorică** ce leagă două fragmente continue și adiacente de text. **Unitatea elementară de discurs**, ce se găsește la nivelul cel mai de jos al reprezentării structurii RST este identificată ca fiind o clauză/propoziție ce cuprinde o predicatie. Relațiile leagă aceste unități într-o structură arborescentă, ce are ca frunze unități elementare de discurs și ca noduri interioare grupuri de mai multe unități elementare adiacente în discurs.

RST identifică două tipuri mari de relații retorice: paratactice și hipotactice. O relație este *paratactică*, sau echinucleară, dacă leagă doi sau mai mulți constituenți egali ca importanță și *hipotactică* dacă leagă constituenți ce nu sunt egali ca importanță. Între constituenții uniți de relațiile hipotactice există întotdeauna unul singur mai important, numit **nucleu**, ceilalți fiind numiți **sateliți**. La relațiile paratactice, prin convenție se consideră că toți constituenții sunt nucleari. Aceste relații sunt clasificate în 27 de tipuri ce diferă prin legătura semantică dintre fragmentele legate și de semnificația individuală a constituenților. În **Tabelul 1** prezentăm câteva din relațiile RST împreună cu semnificația lor pentru un constituent care este nucleu sau pentru unul care este satelit.

Pentru fraza: **1. Angajații trebuie să completeze un nou formular de beneficiar al asigurării pe viață 2. ori de câte ori există o schimbare în statutul marital.**, între constituenții **1.** și **2.** există o relație de tip „Condition”. Nucleul este reprezentat de partea **1.**, în timp ce partea **2.** reprezintă satelitul. Această relație este hipotactică, în termenii definiți mai sus. În exemplul următor (dintr-o rețetă culinară), cele două propoziții sunt în relația RST Sequence una față de cealaltă și ambele propoziții reprezintă nucleii: **1. Cojiți mărul, 2. apoi tăiați-l felii.** Acesta este un exemplu de relație paratactică sau echinucleară.

Numele relației	Nucleu	Satelit
ANTITHESIS	Idei aprobate de autor	Idei dezaprobată de autor
BACKGROUND	Text al cărui înțeles este clarificat	Text care ușurează înțelegerea
CIRCUMSTANCE	Text care exprimă evenimente sau idei care apar în contextul interpretativ	Un context interpretativ al unei situații sau a unui timp
CONCESSION	Situație afirmată de autor	Situația aparent inconsistentă, dar de asemenea afirmată de autor
CONDITION	Situație a cărei apariție rezultă din apariția unei situații condiționale	Situație condițională
ELABORATION	Informație de bază	Informație adițională
ENABLEMENT	O acțiune	Informație care intenționează să ajute cititorul în a face o acțiune
EVALUATION	O situație	Un comentariu care evaluează situația
EVIDENCE	O afirmație	Informație care crește încrederea cititorului în acea afirmație
INTERPRETATION	O situație	O interpretare a situației
RESTATEMENT	O situație	O reformulare a situației
SUMMARY	Text	Un sumar al textului

**Tabel 1.** Relații RST împreună cu exemplificarea lor pentru un nucleu sau un satelit

Deși RST permite formalizarea relațiilor dintre unitățile discursului și modul în care contribuie ele la semnificația și forma discursului, nu precizează nimic referitor la coerența și structura locală, din interiorul acestor unități elementare, și nici nu explică de ce unele texte sunt mai ușor de interpretat decât altele, fie de un analizator uman, fie de unul automat.

Inițial apărută ca idee încă din 1981, *Teoria Centrelor* (CT) a fost definită ca atare în 1995 [Grosz, Joshi și Weinstein, 1995] și a dat prima descriere funcțională a coerenței la nivel de unități elementare de discurs. Principalul scop al CT este să explice de ce unele texte sunt mai greu de interpretat decât altele.

Fie exemplul :

- a. *George a jucat șah cu Victor.*
- b. *El a câștigat repede, apoi Victor a plecat să joace fotbal.*
- c. *El era un șahist talentat.*

Acest text poate fi înțeles cu ușurință, nefiind probleme în a identifica pronumele “el” din ultima propoziție ca refindu-se la George.

- a. *George a jucat șah cu Victor.*
- b. *El a câștigat repede, apoi Victor a plecat să joace fotbal.*
- c. *El a dat un gol.*

În acest exemplu avem o dificultate în a identifica persoana referită de pronumele din a treia propoziție. Putem recunoaște pe “el” ca fiind Victor doar pentru că acțiunea realizată se leagă de acțiunea sa din a doua propoziție.

CT presupune discursul împărțit în unități. Ce înseamnă **unitate de discurs** nu este definit riguros în teorie. Autorii utilizează termenul *utterance* (exprimare), în toate exemplele acestea fiind fraze, dar putem considera aceeași unitate ca și în cazul RST, respectiv o propoziție, uneori o clauză.

Expresiile referențiale cuprinse într-o unitate realizează **centre**. Un centru este o entitate semantică, spre deosebire de o expresie referențială care este o entitate lexicală.

CT explică această dificultate prin schimbarea **centrului principal** de la propoziția a doua la a treia. Centrul unei propoziții este identificat ca fiind entitatea principală a unei unități de discurs, în general cea care are și rol de subiect și apare la începutul propoziției. Schimbarea centrului principal implică o dificultate sporită la înțelegerea textului.

### 2.2.2. Descrierea teoriei

Teoria nervurilor (VT – *Veins Theory*) este un model de interpretare globală a discursului. Împrumutând din RST noțiunile de nuclearitate și relații, dar ignorând numele relațiilor. Teoria nervurilor dezvăluie o structură „ascunsă” în arborele de discurs, numită nervură (sau venă), care permite determinarea domeniilor de accesibilitate evocativă (DEA - *Domain of Evocative Accessibility*) pentru fiecare unitate de discurs, ca fiind acel spațiu al discursului unde toți anaforii, aparținând unității de discurs, își găsesc un antecedent.

Teoria nervurilor calculează, cu ajutorul structurilor retorice (RST), șiruri de unități de discurs, numite nervuri, din care putem determina mai departe domenii de accesibilitate pentru fiecare unitate de discurs. Urmând Teoria Structurilor Retorice, considerăm unitățile de bază ale unui discurs ca fiind fragmente de text care nu se suprapun, de obicei reduse la o propoziție și incluzând un singur predicat; și presupunem că între unități individuale sau grupuri de astfel de unități se păstrează diverse relații retorice, coezive și coerente.

Dan Cristea, Nancy Ide și Laurent Romary [1998] propun o generalizare a Teoriei Centrelor de la nivel local la nivelul global al discursului. Astfel, în vreme ce CT se ocupă de problema referențialității între unități de discurs adiacente și situate în același fragment al discursului (referințe locale), VT ia în considerație relațiile dintre structurile globale ale

discursului și rezoluția anaforei, identificând domenii de accesibilitate ale referințelor pentru fiecare unitate de discurs peste structura arborescentă a discursului.

VT are la bază următoarele principii, similar RST:

- Structura unui discurs poate fi reprezentată printr-un arbore, care în cazul VT este binar;
- Un nod terminal (frunză) din acel arbore reprezintă o unitate elementară a discursului, considerată a fi o propoziție (clauză);
- Un nod intermediar din arbore reprezintă o mulțime de unități elementare adiacente ce formează un fragment continuu de discurs care are o structură proprie;
- Nodurile arborelui sunt polarizate: ele pot fi nuclee sau sateliți în funcție de importanța lor relativ la semnificația discursului;
- VT nu identifică tipuri de relații între nodurile arborelui, spre deosebire de cele 27 identificate de RST.

VT introduce o serie de noțiuni importante:

**Expresia „head”** a unui nod este lista ordonată (în ordinea apariției în discurs) a celor mai importante unități din fragmentul de discurs corespunzător nodului. Aceasta se calculează “bottom-up” în felul următor:

- „head”-ul unui nod terminal este eticheta sa (a unității elementare respective);
- „head”-ul unui nod neterminal este concatenarea „head”-urilor nodurilor fii nucleare.

Expresia „head” proiectează unitățile importante în arbore până la nivelul la care ele ajung să facă parte dintr-un satelit sau până la rădăcina arborelui.

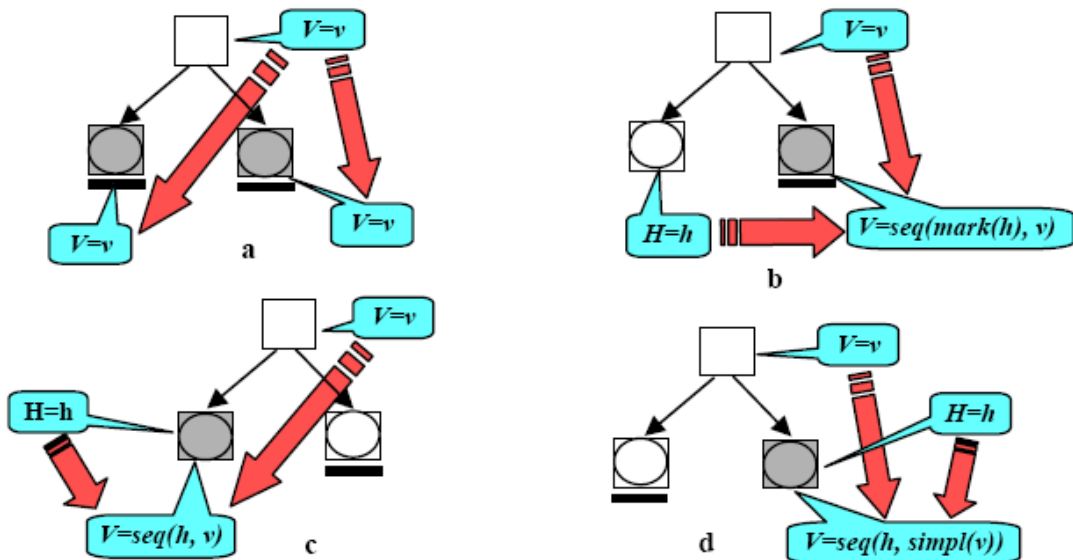
**Expresia „nervură”** (nervura) unui nod reprezintă lista ordonată (în ordinea apariției în discurs) a unităților elementare ce sunt necesare pentru a înțelege semnificația fragmentului de discurs acoperit de nod în contextul întregului discurs. „Nervurile” se calculează top-down în felul următor:

- expresia „nervură” a rădăcinii este aceeași cu expresia “head” a rădăcinii;
- expresia „nervură” a unui nod nuclear fără frate satelit la stânga este aceeași cu expresia “nervură” a nodului părinte;
- expresia „nervură” a unui nod nuclear cu frate satelit la stânga este concatenarea expresiei „nervură” a nodului părinte cu unitățile marcate din „head”-ul fratelui;
- expresia „nervură” a unui fiu satelit stâng este concatenarea „nervurii” părintelui cu expresia „head” a nodului respectiv;

- expresia „nervură” a unui fiu satelit drept este concatenarea „nervurii” părintelui, din care sunt eliminate unitățile marcate, cu expresia „head” a nodului respectiv.

Pentru a defini expresiile „nervură” utilizăm următoarele notații:

- fiecare nod terminal (nod frunză, unitate de discurs) are atașată o etichetă;
- $mark(\alpha)$  este o funcție care primește un șir de simboluri  $\alpha$  și întoarce fiecare simbol din  $\alpha$  marcat într-un anumit fel (de ex. între paranteze drepte);
- $unmark(\alpha)$  este funcția invers pentru  $mark()$ . Elimină toți marcatorii atașați simbolurilor din expresia  $\alpha$ . (ex.  $unmark(\alpha \cdot mark(\beta) \cdot \gamma) = \alpha \cdot \beta \cdot \gamma$ );
- $simpl(x)$  este o funcție care elimină toate simbolurile marcate din argumentul său, dacă acestea există, de ex.  $simpl(mark(\alpha)) = \emptyset$ , șirul vid, și  $simpl(\alpha \cdot mark(\beta) \cdot \gamma) = \alpha \cdot \gamma$ ;
- $seq(x, y)$  este o funcție de secvențiere care primește ca parametri două șiruri disjuncte de noduri terminale etichetate,  $x$  și  $y$ , și returnează acea permutare a lui  $x$  concatenat cu  $y$  dată de citirea de la stânga la dreapta a secvenței de etichete din  $x$  și  $y$  de pe frontiera terminală a arborelui. Funcția păstrează marcajele, dacă acestea există și  $seq(\emptyset, \beta) = \beta$ ;  $seq(\alpha, seq(\beta)) = seq(seq(\alpha), \beta) = seq(\alpha, \beta)$ ;
- $H(n)$  și  $V(n)$  sunt notațiile pentru expresiile „head” și „nervură” pentru un nod  $n$ ;
- $pref(u, \alpha)$  păstrează prefixul expresiei  $\alpha$  pâna la simbolul  $u$  inclusiv.



**Fig. 1.** Calcularea expresiilor nervură. Nodul pentru care se aplică calculul este reprezentat cu gri; nodurile nucleu sunt subliniate [Cristea, 2005]

Un exemplu de calcul al acestor expresii și de reprezentare a arborelui cu “nervuri” marcate:

1. Când l-a auzit pe George în camera alăturată
2. Victor l-a chemat
3. ca să-i ceară ajutorul.
4. Însă Victor îl deranjase pe George
5. și acesta se întoarse în camera sa.
6. Deși George îl refuzase categoric,
7. Victor încă mai spera să îl ajute.

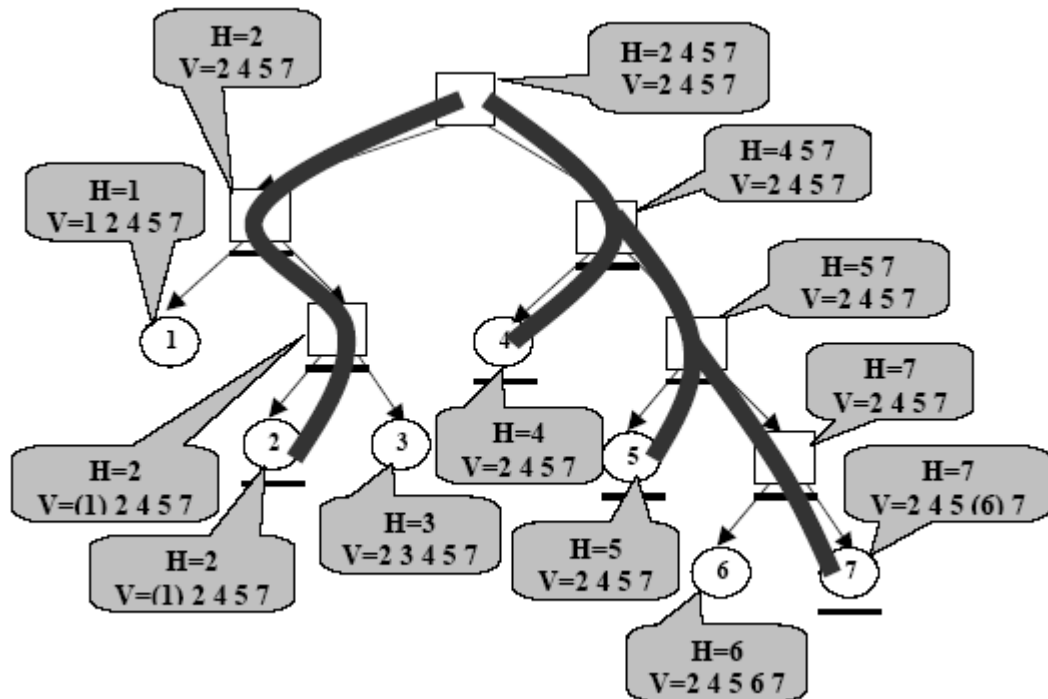


Fig. 2. Reprezentarea nervurilor pe arborele de parsare

Cu „H” este notată expresia „head” iar cu „V” expresia „venă” a unui nod. Cu linii îngroșate sunt marcate pe arbore liniile principale de argumentație în text, așa cum sunt ele deduse din expresiile „venă” calculate. Frunzele arborelui reprezentat în **Figura 2** reprezintă unitățile elementare din care este format textul oferit drept exemplu. Expresiile “head” pentru acestea sunt formate din eticheta corespunzătoare unității pe care o conțin. Pentru nodul rădăcină, expresia “head” ( $H = 2\ 4\ 5\ 7$ ) reprezintă concatenarea tuturor “head”-urilor din nodurile fii satelit (cele subliniate). Nodul frunză cu eticheta **1** este fiu satelit stâng pentru nodul părinte. Expresia “venă” ( $V = 1\ 2\ 4\ 5\ 7$ ) a acestuia este obținută prin concatenarea expresiei “venă” a nodului părinte ( $V = 2\ 4\ 5\ 7$ ) cu expresia “head” ( $H = 1$ ) corespunzătoare nodului satelit.

O altă noțiune introdusă de VT este aceea de **domeniu de accesibilitate evocativă** (DEA) al unui nod terminal și reprezintă o listă de unități elementare de discurs, ordonate în

ordinea apariției lor, în care este cel mai probabil să fie găsiți antecedentii anaforici ai entităților semantice din unitatea elementară desemnată de nod. DEA se calculează pentru unitatea „u” ca fiind prefixul venei unității „u” luat până la apariția unității „u” (toate unitățile apar în expresia „venă” a lor).

## 2.3. Temporalitate

### 2.3.1. Istoric

Recunoașterea automată a expresiilor temporale și a evenimentelor în limbajul natural a devenit recent un domeniu de cercetare intensivă în lingvistica computațională și Inteligență Artificială. Importanța informației temporale în sistemele de tip Întrebare-Răspuns a devenit mult mai evidentă pe măsură ce aceste sisteme tind să depășească bariera înțeleșului la nivel de cuvânt. Cercetarea în acest domeniu s-a axat inițial pe un corpus de articole din ziare și este descrisă pe larg de James Pustejovsky [Pustejovsky *et al.*, 2005a] și Inderjeet Mani [Mani *et al.*, 2005b].

Articolele din presă descriu evenimente cu diverse moduri de corelare a acestora în timp. Așa cum se întâmplă, totuși, mare parte din informația temporală este subînțeleasă într-un astfel de text. Localizarea temporală a evenimentelor este rareori explicită și multe expresii temporale sunt vagi. Un prim pas crucial în extragerea informațiilor temporale a fost capacitatea de a identifica ce evenimente sunt descrise în text și de a explicita când au avut loc aceste evenimente.

Întrebări precum cele enumerate mai jos pot primi cu ușurință răspuns din partea oamenilor după citirea unui articol de ziar, în schimb, sistemele automate pot oferi doar răspunsuri limitate:

- 1) Este Merkel actualul cancelar al Germaniei?
- 2) Ce s-a întâmplat pe plan politic în Rusia în ultima săptămână?
- 3) Când a avut loc fuziunea între Banca ING și Banca Țiriac?

Recunoașterea „cuvintelor cheie” specifice temporării (de ex.: *actualul, ultima săptămână, când*) reprezintă în mod clar o necesitate pentru înțelegerea și oferirea de răspunsuri acestor întrebări. În primul rând, aspecte temporale ale proprietăților entităților (de ex.: proprietatea de a fi cancelar al Germaniei) trebuie reprezentate în mod adecvat. În al doilea rând, trebuie avut în vedere extragerea descrierilor evenimentelor împreună cu amprenta lor temporală. Veridicitate a evenimentelor trebuie verificată de asemenea (de ex.:

evenimente actuale vs. evenimente probabile ). După cum se observă din aceste trei întrebări oferite ca exemplu, extragerea și procesarea automată informației despre evenimente și expresii temporale ridică noi probleme în cadrul cercetării actuale.

Cercetările în acest domeniu au dus la apariția inițială a schemelor de adnotare temporală TIMEX și TIMEX2 [Ferro *et al.*, 2001]. Mai apoi, în contextul a trei workshop-uri și proiecte AQUAINT, a fost definit standardul de adnotare temporală TimeML<sup>1</sup>.

### 2.3.2. TimeML

TimeML este un limbaj robust de specificare pentru expresii temporale și evenimente în limbajul natural. Spre deosebire de majoritatea încercărilor anterioare de specificare a timpului și evenimentelor, TimeML separă reprezentarea lor de dependențele de ordonare și ancorare care există în text. Mai jos sunt detaliate caracteristicile care evidențiază TimeML față de încercările anterioare de adnotare temporală, așa cum apar descrise de Pustejovsky [Pustejovsky *et al.*, 2005b]:

1. Extinde atributele de adnotare din TIMEX2.
2. Introduce **Funcții Temporale** ce permit expresii specificate intențional: *cu trei ani în urmă, luna trecută*.
3. Identifică semnale care determină interpretarea expresiilor și legăturilor temporale: *în timpul, la, înainte, după, în același timp*.
4. Identifică toate clasele de expresii eveniment:
  - (a) Verbe conjugate: *a plecat, a fost capturat, va demisiona*
  - (b) Adjective ce reprezintă evenimente statice: *scufundat, împotmolit*
  - (c) Substantive pentru evenimente: *Operațiune Militară*
5. Crează legături între evenimente și expresii temporale:
  - (a) Ancorate: *Ion a plecat luni*.
  - (b) Ordonate: *Petrecerea a avut loc după miezul nopții*.
  - (c) Incluse: *Ion a zis că Maria a plecat*.

Marcajele definite în TimeML au ca scop facilitarea ~~derivate~~ unelte și reprezentării care cer referințe la informații senzitive din punct de vedere temporal (de ex.: sisteme întrebare-răspuns, interogări în cadrul unor servicii web, rezumare de text). Pentru aceasta au fost incluse în TimeML patru structuri de date majore [Ingria și Pustejovsky, 2002]: EVENT, TIMEX3, SIGNAL și LINK. Tagul EVENT marchează toate evenimentele

---

<sup>1</sup> Informații suplimentare disponibile la adresa <http://timeml.org>

temporale. Tagul TIMEX3 este utilizat pentru a captura toate expresiile temporale. Cuvinte funcționale, precum *la*, *de la* sau *după*, sunt capturate de tagul SIGNAL. Toate relațiile între celelalte taguri sunt reprezentate cu taguri tip LINK: TLINK (*Time Link*), ALINK (*Aspectual Link*) și SLINK (*Subordinating Link*). În lucrarea de față vom lucra în mod special cu taguri de timpul TLINK și SIGNAL.

### 2.3.2.1 Expresii temporale

La baza oricărei scheme create pentru a oferi informații temporale există o metodă de a reprezenta expresii temporale specifice, cum ar fi *astăzi* sau *2006*. TimeML modelează acest tip de expresii cu tagul TIMEX3. Sunt patru tipuri de expresii temporale capturate în TIMEX3: TIME, DATE, DURATION și SET.

O expresie care primește tipul TIME este una care face referire la un timp al unei zile, chiar și într-un mod greu de definit. Pentru fiecare expresie temporală este calculat un grad de granularitate. Cel mai simplu mod de a deosebi tipul TIME de tipul DATE este să privim la granularitate. Dacă aceasta este mai mică decât o zi, atunci expresia este de tipul TIME. Exemple de expresii care intră în această categorie, o expresie fiind adnotată:

*George a plecat **târziu noaptea trecută**  
la 9 și 10 minute  
la 5 a.m., vineri, 20 octombrie*

```
<TIMEX3 tid="t1" type="TIME" value="T05:00" temporalFunction="TRUE">
  5:00 a.m.
</TIMEX3>,
<TIMEX3 tid="t2" type="DATE" anchorID="t3">
  vineri
</TIMEX3>,
<TIMEX3 tid="t3" type="DATE" value="xxxx-11-22">
  20 octombrie
</TIMEX3>
```

Atributul anchorID din a doua expresie temporală exprimă faptul că *vineri* face referire la data marcată de expresia temporală cu ID-ul t3. În valoarea atributului value din ultima expresie temporală "xxxx" marchează anul, în care s-au petrecut evenimentele ce fac referire la această dată, ca fiind necunoscut.

Orice expresie care face referire la o dată calendaristică primește tipul DATE. Pentru a evita confuzia ce se poate crea între tipul TIME și tipul DATE folosim testul granularității, amintit mai sus. Iată câteva exemple din această categorie:

*George a plecat **vineri, 1 iulie 1998**  
ieri  
în vara anului 1996*

```
<TIMEX3 tid="t1" type="DATE" value="2004-11-22">
  22 noiembrie 2004
```

</TIMEX3>

O expresie este de tipul DURATION și descrie un interval specific de timp.

Câteva exemple:

*George a stat **2 luni** în Boston.*

**48 de ore**

**3 săptămâni**

<TIMEX3 tid="t1" type="DURATION" value="P4D">

patru zile

</TIMEX3>

În sfârșit, tipul SET este utilizat pentru expresii care descriu o mulțime de timpi care se repetă cu regularitate:

*George înoată o dată la două săptămâni.*

**de două ori pe lună.**

<TIMEX3 tid="t1" type="SET" value="P1W" quant="EACH" freq="3D">

3 zile pe săptămână

</TIMEX3>

Valorile atributelor din exemplul de mai sus exprimă complet expresia temporală marcată: 3 zile ("3D" = 3 Days) pentru fiecare ("EACH") perioadă de o săptămână ("P1W" = Period 1 Week). Atributele marcatorului TIMEX3 pot avea foarte multe valori, acestea fiind definite în standardul TIDES [Ferro *et al.*, 2001].

Forma BNF<sup>2</sup> a tagului TIMEX3:

```

attributes:: = tid type (value | valueFromFunction)
[functionInDocument] [beginPoint] [endPoint] [quant] [freq]
[temporalFunction] [mod][anchorTimeID]
tid :: = ID
{tid :: = TimeID
TimeID :: = t<integer>}
type :: = 'DATE' | 'TIME' | 'DURATION' | 'SET'
value :: = CDATA
{value:: =
duration|dateTime|time|date|gYearMonth|gYear|gMonthDay|gDay|gMonth}
valueFromFunction:: = IDREF
{valueFromFunction:: = TemporalFunctionID
functionInDocument:: = 'CREATION_TIME' | 'EXPIRATION_TIME' |
'MODIFICATION_TIME' | 'PUBLICATION_TIME' | 'RELEASE_TIME' |
'RECEPTION_TIME' | 'NONE'
beginPoint :: = IDREF {beginPoint :: = TimeID}
endPoint :: = IDREF {endPoint :: = TimeID}
quant :: = CDATA
freq :: = CDATA
temporalFunction :: = 'true' | 'false'
    
```

---

<sup>2</sup> În informatică forma Backus-Naur (BNF) este o metasintaxă utilizată pentru a exprima gramatici independente de context: mai exact, o modalitate de a descrie limbaje formale.

```
mod:: = 'BEFORE' | 'AFTER' | 'ON_OR_BEFORE' | 'ON_OR_AFTER' |  
'LESS_THAN' | 'MORE_THAN' | 'EQUAL_OR_LESS' | 'EQUAL_OR_MORE' |  
'START' | 'MID' | 'END' | 'APPROX'
```

```
anchorTimeID :: = IDREF {anchorTimeID :: = TimeID}
```

1) **tid**: atribut obligatoriu, ID-ul expresiei temporale; fiecare expresie TIMEX3 trebuie să fie identificată printr-un ID unic. Acesta este asignat automat de instrumentul de adnotare.

2) **type**: atribut obligatoriu (descrie pe larg mai sus).

3) **value**: atribut obligatoriu; este echivalentul atributului VAL definit de TIMEX2.

4) **mod**: atribut opțional; echivalentul atributului MOD definit pentru marcajul TIMEX2. Valorile sale sunt cele prezentate în cadrul TIMEX2.

5) Atributele **beginPoint** și **endpoint** sunt folosite atunci când o durată este ancorată de o altă expresie temporală:

```
<TIMEX3 tid = "t6" type = "DURATION" value = "P2W" beginPoint = "t61"  
endPoint = "t62">two weeks</TIMEX3>  
<SIGNAL sid = "s1">from</SIGNAL>  
<TIMEX3 tid = "t61" type = "DATE" value = "2003-06-07">June 7,  
2003</TIMEX3>  
<TIMEX3 tid = "t62" type = "DATE" value = "2003-06-21"  
temporalFunction = "true" anchorTimeID = "t6"/>
```

6) Atributul **quant** cuantifică expresiile de tip SET, iar atributul **freq** conține un întreg și o granularitate a timpului care reprezintă frecvența cu care expresia temporală reapare regulat.

7) **temporalFunction** - atribut binar (false/true) care specifică necesitatea ca valoarea expresiei temporale să fie determinată folosind funcții temporale.

8) **anchorTimeID**: atribut opțional; introduce ID-ul unei expresii temporale la care este ancorat TIMEX3-ul curent. Valoarea lui este întotdeauna un timeID. Ancorele temporale sunt din afara spațiului marcajului TIMEX3. Atributul **anchorTimeID** apare cu **temporalFunction="true"**.

9) **valueFromFunction**: acest atribut nu este relevant pentru scopurile adnotării manuale. Adnotatorul uman ar trebui să-l ignore.

10) **functionInDocument**: acest atribut indică funcția pe care o are un TIMEX3 în cadrul unui document. Se disting câteva momente ce marchează etapele majore din viața unui reportaj de știri. Acestea sunt prezentate în continuare împreună cu valoarea pe care acest atribut o va lua în fiecare caz:

- ✓ momentul la care a fost creat - „CREATION\_TIME”;

- ✓ momentul la care a fost modificat - „MODIFICATION\_TIME”;
- ✓ momentul la care a fost publicat - „PUBLICATION\_TIME”;
- ✓ momentul la care el poate fi expediat (dacă nu imediat) - „RELEASE\_TIME”;
- ✓ momentul la care este primit de client - „RECEPTION\_TIME”;
- ✓ momentul la care reportajul expiră (dacă acesta există) - „EXPIRATION\_TIME”.

În cazul în care expresia adnotată nu îndeplinește în document nici una din funcțiile prezentate mai sus valoarea sa va fi „NONE”.

### 2.3.2.2. Tagul EVENT

Evenimentele sunt descrise prin tagul EVENT, imediat corelat cu tagul MAKEINSTANCE.

Se consideră evenimente acei termeni ce descriu situații care se întâmplă sau apar și predicate care descriu situații sau circumstanțe în care un fapt devine sau rămâne adevărat. Evenimentele pot fi punctuale sau pot să dureze o anumită perioadă de timp. Ele sunt exprimate prin:

- verbe cu sau fără timp: *We are **waiting** for him.*,
- substantivizări (nume de evenimente): *Several **demonstrations** have taken place in the last week in Manilla.*,
- adjective: *A volcano, **dormant** for two centuries, ...*
- predicate nominale: *There is no reason why we would not be **prepared**.*,
- expresii prepoziționale: *All people **on board** of the aeroplane died.*

Forma BNF a tagului EVENT este:

```
attributes ::= eid class
  eid ::= e<integer>
  class ::= REPORTING | PERCEPTION | ASPECTUAL | I_ACTION |
I_STATE | STATE | OCCURRENCE
```

Atributele marcajului EVENT sunt:

- 1) **eid**: atribut obligatoriu, ID-ul evenimentului – se asignează automat de instrumentul de adnotare de fiecare dată când este introdus un marcaj EVENT.
- 2) **class**: atribut obligatoriu; fiecare eveniment aparține uneia din clasele date mai jos. Verbele pot fi ambigue relativ la clasa din care fac parte. Dacă un verb apare într-un exemplu

ca făcând parte dintr-o anumită clasă, nu înseamnă că fiecare apariție a aceluși verb exprimă un eveniment din aceeași clasă.

Valorile posibile ale atributului **class**:

- **REPORTING**: evenimentele din această clasă descriu acțiunea unei persoane sau a unei organizații care declară ceva, narează sau informează despre un eveniment, etc.

Exemple: *a spune, a raporta, a relata, a povesti, a explica, a declara, etc.*

- **PERCEPTION**: această clasă include evenimente ce implică percepția fizică a unui alt eveniment.

Exemple: *a vedea, a privi, a ochi, a cerceta cu privirea, a auzi, a asculta, etc.*

- **ASPECTUAL**: evenimentele din această clasă surprind diferitele fațete ale istoriei unui eveniment:

- Inițierea: *a începe, a porni, a lansa, a iniția, a produce, etc.*

- Reinițierea: *a restarta, a reîncepe, a reiniția, etc.*

- Terminarea: *a opri, a anula, a sfârși, a termina, etc.*

- Punctul culminant: *sfârșit, completare, etc.*

- Continuarea: *a continua, a menține, a merge înainte, a înainta, a merge mai departe, a susține, a persista, a persevera, etc.*

- **I\_ACTION**: un eveniment din această clasă desemnează o acțiune dorită sau intenționată care introduce un eveniment explicit reprezentat în text.

O listă reprezentativă (dar nu exhaustivă) de evenimente de tip **I\_ACTION** (**INTENSIONAL\_ACTION**) conține evenimente ca: *a încerca, a depune eforturi, a cerceta, a investiga, a se uita la, a amâna, a evita, a preveni, a anula, a împiedica, a cere, a ordona, a determina, a convinge, cerere, aruga, a condamna, a îndemna, a autoriza, a promite, a oferi, a propune, a fi de acord, a decide, a jura, a numi, numirea, a alege.*

Exemplu: *Microsoft încearcă să monopolizeze piața sistemelor de operare..* Evenimentul din clasa **I\_ACTION** este *încercă*, în timp ce evenimentul explicit reprezentat în text, la care acesta dinainte face referire, este *să monopolizeze*.

- **I\_STATE**: evenimentele din această clasă sunt similare cu cele din clasa precedentă și se referă la lumi alternative sau posibile.

Următoarea listă de evenimente de clasă I\_STATE este reprezentativă, nu exhaustivă: *a crede, a gândi, a suspecta, a imagina, a se îndoii, a simți, a considera, a fi posibil, a fi sigur, a dori, a place, dorință, a cere, a tânji, a pofti, a vrea, a spera, a aștepta, a aspira, a plănuși, a se teme, a urî, a se înspăimânta de, a-și face griji, a fi speriat, a avea nevoie, a cere, a necesita, a fi gata, a fi nerăbdător, a fi pregătit, a fi capabil, a nu fi capabil.*

- **STATE:** evenimentele din această clasă descriu circumstanțe în care ceva devine sau rămâne adevărat:

- Stări care sunt identificabil schimbate pe parcursul documentului de marcat.

- Situații care sunt în relație directă cu o expresie temporală. Acest criteriu include toate situațiile legate la un TIMEX3 marcabile prin intermediul unui TLINK

- Situații care sunt introduse de un eveniment: I\_ACTION, I\_STATE sau REPORTING.

- Situații predicative a căror validitate depinde de momentul creării documentului.

- **OCCURRENCE:** această clasă include toate celelalte tipuri de evenimente care nu au fost încadrate în nici una din clasele anterioare.

### 2.3.2.3. Tagul MAKEINSTANCE

Bazat pe adnotarea evenimentelor, tagul MAKEINSTANCE indică instanțele unui eveniment, acestea fiind cele care participă în legăturile temporale. Acest tag se inserează în afara textului, pentru fiecare realizare sau instanță a unui eveniment, și își are originea în analize făcute pe corpusuri adnotate. Introducerea acestui tag este motivată de exemple precum *Ion a predat luni și marți.*, unde un singur verb (*a preda*) desemnează două instanțe diferite ale aceluiași eveniment. În acest caz vor trebui evidențiate două instanțe ale evenimentului marcat. Pe lângă posibilitatea de a instanția diferit evenimentele, tagul MAKEINSTANCE captează și alte informații, în general motivate lexical: timpul, aspectul, morfologia – pentru forme fără timp, polaritatea și modalitatea unei instanțe a evenimentului. Exemplul de mai jos [Pustejovsky *et al.*, 2005a] ilustrează și mai bine utilitatea folosirii acestui tag.

*John teaches on Monday but might not on Tuesday.*

O instanță a evenimentului *teaches* conține atît un operator de negare cît și unul modal, pe când cealaltă instanță - nu:

```
John <EVENT eid="e2" class="OCCURRENCE">teaches</EVENT> on  
<TIMEX3 tid="t1" type="DATE">Monday</TIMEX3> but might
```

```
<SIGNAL sid="s1">not</SIGNAL> on
<TIMEX3 tid="t2" type="DATE">Tuesday</TIMEX3>.
<MAKEINSTANCE eiid="ei1" eventID="e2" tense="PRESENT" aspect="NONE"/>
<MAKEINSTANCE eiid="ei2" eventID="e2" tense="PRESENT" aspect="NONE"
modality="MIGHT" polarity="NEG"/>
```

Forma BNF a tagului MAKEINSTANCE este:

```
attributes ::= eiid eventID tense aspect negation [modality]
[signalID] [cardinality]
eiid ::= ei<integer> //EventInstanceID
eventID ::= e<integer> //EventID
tense ::= 'PAST' | 'PRESENT' | 'FUTURE' | 'NONE'
aspect ::= 'PROGRESSIVE' | 'PERFECTIVE' | 'PERFECTIVE_PROGRESSIVE' | 'NONE'
negation ::= 'true' | 'false'
modality ::= CDATA
signalID ::= s<integer>
cardinality ::= <integer> | 'EVERY'
```

Atributele acestui tag sunt:

- 1) **eiid**: ID-ul marcajului de instanță, atribut obligatoriu ce se folosește în marcarea legăturilor;
- 2) **eventID**: ID-ul evenimentului pentru care a fost creat;
- 3) **tense**: timpul clauzei prin care este exprimat evenimentul;
- 4) **aspect**: în limba engleză există o categorie aparte pentru verbe care arată aspectul acestora. Aspectul este marcat prin combinații ale verbelor auxiliare (*be* sau *have*) și terminații ale verbului principal (*-ing* sau *-en/-ed*).
- 5) **signalID**: ID-ul *signal*-ului care arată cardinalitatea (numărul de instanțe);
- 6) **cardinality**: un întreg care reprezintă numărul de instanțe, atribut opțional, care este utilizat atunci când numărul de instanțe este mare.

#### 2.3.2.4. Tagul SIGNAL

Un *signal* este un element din text care face explicită relația dintre două entități (o expresie temporală și un eveniment sau două evenimente), indică faptul că evenimentul este determinat de un verb auxiliar modal, că este precedat de o negație sau că referă mai multe instanțe ale aceluiași eveniment.

În general un *signal* face parte din următoarele categorii:

- Prepoziții temporale: *la, în, pe, de pe, până pe, înainte, după, în timpul*, etc.;
- Conjunții temporale: *înainte, după, în timpul, cât timp, când* etc.;

Două signal-uri ce apar alăturate într-o propoziție sunt marcate separat doar dacă aparțin la tipuri diferite. Altfel sunt adnotate ca un singur SIGNAL:

*Ei vor investiga rolul pe care l-au avut Statele Unite*  
<SIGNAL sid="s2"> înainte, în timpul și după </SIGNAL> genocid.

Marcajul SIGNAL are un singur atribut care este obligatoriu: sid, id-ul unic al signalului. Acesta va fi asignat automat de instrumentul de adnotare de fiecare dată când un SIGNAL este marcat.

### 2.3.2.5. Tagurile de legături LINK

Marcajele de tip LINK codifică diferitele legături ce apar între elementele temporale ale unui document, specificând ordonarea și ancorarea în timp a instanțelor de evenimente, precum și relațiile de subordonare și cele aspectuale dintre aceste instanțe. Marcajele de legătură se inserează, ca și MAKEINSTANCE, în afara textului, tipul de legătură, dat de atributul relType, fiind fundamental în definirea acestor legături. Sunt definite trei tipuri de legături, prezentate în continuare.

#### 2.3.2.5.1. Legături temporale: TLINK

Un TLINK sau TemporalLink marchează o relație temporală de ancorare sau ordonare între două instanțe de evenimente sau între o instanță de eveniment și o expresie temporală.

În conformitate cu cele 13 relații ale lui Allen [Allen, 1984], în TimeML se definesc 13 tipuri de legături temporale (valorile posibile ale atributului relType), specificând dacă entitățile corelate sunt:

1. **SIMULTANEOUS** – entități temporale simultane sau temporar de nedistins în context;
2. **BEFORE** – o entitate înaintea celeilalte;

*Poliția a cercetat uciderile a 14 femei. În șase din aceste cazuri suspiecții au fost deja arestați.*

3. **AFTER** – o entitate după cealaltă. Aceasta este inversa relației precedente. Deci cele două evenimente marcate în exemplul anterior pot fi adnotate alternativ ca exprimând o relație de tip **AFTER**, dacă direcția este inversată.

- Modificatori temporali: *de două ori, de fiecare dată, etc.*;
- Expresii negative: *nu, nici unul, niciodată, nimeni, etc.*;
- Verbe auxiliare modale: *a putea, a trebui*;
- Prepoziții subordonatoare: *să*;
- Caractere speciale: „-” și „/”, în expresii temporale ce desemnează

4. **IMMEDIATELY\_BEFORE** – o entitate imediat înaintea celeilalte;

*Toți pasagerii au murit când avionul s-a prăbușit în munți.*

5. **IMMEDIATELY\_AFTER** – o entitate imediat după cealaltă;

6. **INCLUDES** – o entitate temporală este inclusă în cealaltă:

*El a ajuns în Iași joia trecută.*

7. **IS\_INCLUDED** – o entitate temporală o include pe cealaltă: inversa relației anterioare;

*Ion a predat în ziua de luni.*

```
Ion a <EVENT eid="e1" class="OCCURENCE">predat</EVENT>
<SIGNAL sid="s1"> în </SIGNAL>
<TIMEX3 tid="t1" type="DATE" value="XXXX-04-12"
temporalFunction="true"> ziua de luni </TIMEX3>.
<MAKEINSTANCE eiid="ei1" eventID="e1" tense="PAST" aspect="NONE"/>
<TLINK eventInstanceID="ei1" relatedToTime="t1" signalID="s1"
relType="IS_INCLUDED" />
```

8. **HOLDS** – pentru stări și evenimente ce persistă pentru o perioadă:

*El a fost director pentru 3 ani.;*

9. **BEGINNING** – o entitate e la începutul celeilalte:

*El e la sală de la 5 la 7.;*

10. **BEGUN\_BY** – o entitate este începută de cealaltă – inversa relației anterioare;

11. **ENDING** – o entitate e la sfârșitul celeilalte:

*El e la sală de la 5 la 7.;*

12. **ENDED\_BY** – inversa relației anterioare;

13. **IDENTITY** – pentru două evenimente simultane.

*John a călătorit spre Boston. În timpul călătoriei el a mâncat o gogoasă.*

În cazul adnotării manuale a unui text, decizia de a marca o relație temporală ca fiind AFTER sau IMMEDIATELY\_AFTER rămâne la latitudinea adnotatorului. Pentru un instrument care realizează adnotarea automată a textului este greu să decidă ce relație va marca în cazul amintit, iar cel mai adesea relația temporală va fi adnotată cu tipul AFTER.

Atributele tagului TLINK sunt descrise în BNF:

```
attributes :: = [lid] [origin] (eventInstanceID | timeID) [signalID]
(relatedtoEventInstance | relatedtoTime) relType
lid :: = ID
{lid :: = LinkID
LinkID :: = l<integer>}
origin :: = CDATA
eventInstanceID :: = ei<integer>
timeID :: = t<integer>
```

```

signalID ::= s<integer>
relatedToEventInstance ::= ei<integer>
relatedToTime ::= t<integer>
relType ::= 'BEFORE' | 'AFTER' | 'INCLUDES' | 'IS_INCLUDED' | 'DURING' |
'SIMULTANEOUS' | 'IAFTER' | 'IBEFORE' | 'IDENTITY' | 'BEGINS' | 'ENDS' |
'BEGUN_BY' | 'ENDED_BY'
    
```

Atributele includ ID-ul instanței sursei (`relatedToEventInstance`), al entității destinație (`eventInstanceID`), tipul relației (`relType`) și, dacă relația e semnalată de un *signal*, ID-ul acestuia (`signalID`).

### 2.3.2.5.2. Legături de subordonare: SLINK

Un SLINK sau SubordinatedLink va fi folosit pentru a marca relația de subordonare dintre două evenimente sau relația dintre un eveniment și un *signal*.

Un SLINK poate avea unul din următoarele tipuri:

**1. MODAL:** Această relație este introdusă de cele mai multe ori de un verb modal (*a putea, a trebui*), care va fi marcat ca un SIGNAL, dar și de evenimente care fac referință la o lume posibilă – mai ales I\_STATE-urile.

*Ion ar fi trebuit să cumpere niște vin.*

**2. FACTIVE:** Această relație este introdusă de verbe care exprimă o necesitate (sau o presupunere) a adevărului argumentelor lor, cum sunt: *a uita, a regreta, a reuși*.

*Ion a uitat că a fost în București anul trecut.*

**3. CONTRA\_FACTIVE:** Contrar relației anterioare, în acest caz evenimentul introduce o prezumpție despre neadevărul (neîndeplinirea) argumentelor lui: *a uita să, a nu fi capabil să* (la trecut), *a împiedica, a anula, a evita, a refuza* etc.

*Maria a uitat să cumpere vin.*

**4. EVIDENTIAL:** Acest tip de relație este introdusă de obicei de evenimente de clasă REPORTING sau PERCEPTION:

*Maria l-a văzut pe Ion cumpărând doar bere.*

**5. NEG\_EVIDENTIAL:** Această relație este introdusă de evenimente de clasă REPORTING și PERCEPTION cu o polaritate negativă:

*Ion a negat că a cumpărat doar bere.*

**6. NEGATIVE:** Un marcaj SLINK de acest tip va marca relația dintre o particulă negativă (marcată ca SIGNAL) și evenimentul pe care îl determină.

*Ion nu a uitat să cumpere vin.*

Pentru fiecare eveniment REPORTING sau PERCEPTION trebuie introdus un marcaj SLINK exprimând relația dintre acestea și evenimentele subordonate lor.

În mod similar, pentru fiecare I\_ACTION sau I\_STATE este introdus un SLINK ce exprimă relația între evenimentul intenționat și evenimentul subordonat lui.

Atributele tagului SLINK sunt incluse în BNF-ul acestuia:

```
attributes ::= [lid] [origin] [eventInstanceID] [signalID]
subordinatedEventInstance relType
lid ::= ID
{lid ::= LinkID
LinkID ::= l<integer>}
origin ::= CDATA
eventInstanceID ::= ei<integer>
signalID ::= s<integer>
subordinatedEventInstance ::= ei<integer>
relType ::= 'MODAL' | 'NEGATIVE' | 'EVIDENTIAL' | 'NEG_EVIDENTIAL' |
'FACTIVE' | 'COUNTER_FACTIVE'
```

### 2.3.2.5.3 Legături aspectuale: ALINK

Un ALINK sau AspectualLink marchează relația dintre un eveniment aspectual și evenimentul pe care îl determină. Exemple de relații aspectuale ce trebuie marcate:

#### 1. Inițierea: *John a început să citească.*

```
John a <EVENT eid="e1" class="ASPECTUAL">început</EVENT> să <EVENT
eid="e2" class="OCCURENCE">citească</EVENT>.
<MAKEINSTANCE eiid="ei1" eventID="e1" tense="PAST" />
<MAKEINSTANCE eiid="ei2" eventID="e2" tense="PRESENT" /> <ALINK
eventInstanceID="ei1" relatedToEvent="e2" relType="INITIATES" />
```

#### 2. Culminarea: *John a terminat de citit.*

#### 3. Terminarea: *John s-a oprit din vorbit.*

#### 4. Continuarea: *John a continuat să vorbească.*

#### 5. Reinițierea: *John a reînceput să vorbească.*

Atributele tagului ALINK sunt:

```
attributes ::= [lid] eventInstanceID [signalID]
relatedToEventInstance relType [syntax]
lid ::= ID
{lid ::= LinkID
LinkID ::= l<integer>}
eventInstanceID ::= ID
{eventInstanceID ::= EventInstanceID}
signalID ::= IDREF
{signalID ::= SignalID}
relatedToEventInstance ::= IDREF
{relatedToEventInstance ::= EventInstanceID}
relType ::= 'INITIATES' | 'CULMINATES' | 'TERMINATES'
| 'CONTINUES' | 'REINITIATES'
```

syntax ::= CDATA

### 3. Corpusul de texte

**Adnotarea de corpusuri** reprezintă un instrument folosit în cercetarea lingvistică bazată pe date. Tradițional, un *corpus* face referire la un ansamblu de date în limbaj natural (de ex.: *text scris, discursuri rostite, etc.*), utilizat drept suport pentru cercetare lingvistică. În zilele noastre, această definiție s-a schimbat și termenul *corpus* descrie un ansamblu de texte în format electronic care pot fi procesate de un calculator, utilizat ca parte a instrumentelor din domeniul procesării limbajului natural.

Pentru realizarea studiului propus în această lucrare am ales un corpus de text creat de Daniel Marcu [Marcu *et al.*, 1999]. Acest corpus este compus din 385 de articole în engleză americană din Wall Street Journal (WSJ), extrase din Penn Treebank [Marcus *et al.*, 1993] și adnotate pentru structura de discurs conform cu RST. Corpusul conține 176,383 de cuvinte, cu o medie de 458 de cuvinte/text și 574 de unități elementare de discurs/text. Fiecare unitate elementară de discurs (propoziție sau unitate mai mică) conține în medie 8 cuvinte.

Alegerea acestui corpus este motivată de faptul că oferă ușurință în calculul nervurilor, conține texte cu multiple expresii temporale și evenimente legate de acestea și textele au fost adnotate manual pentru RST, ceea ce conferă credibilitate rezultatelor obținute.

Pornind de la corpusul inițial (WSJ) următoarele etape au permis obținerea corpusului de lucru final (conform cu **Fig. 3**):

1. Obținerea adnotării pentru nervuri;
2. Obținerea adnotării pentru temporalitate;

3. Obținerea corpusului final prin operația de reuniune (*merge*) a celor două adnotări.

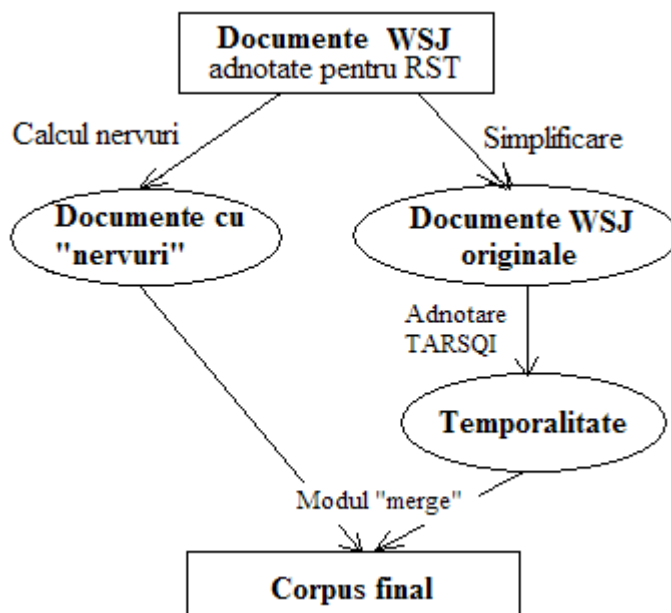


Fig. 3. Reprezentare succintă a procesului de obținere a corpusului de text.

### 3.1. Obținerea nervurilor

Utilizând formulele de calcul pentru „heads” și „nervuri” descrise în secțiunea 2.2.2, am utilizat un modul [Pistol, 2005] care primește la intrare un fișier adnotat RST și întoarce acest fișier la care au fost adăugate, pentru fiecare segment de text, informații despre „nervuri”.

De exemplu, secvența xml:

```

<seg id='2' nuc='yes' leaf='1' rel2par='span' >
  <w pos='JJ'>Federal</w>
  <w pos='NNS'>investigators</w>
  ....
</seg>
    
```

va deveni în urma aplicării formulelor de calcul:

```

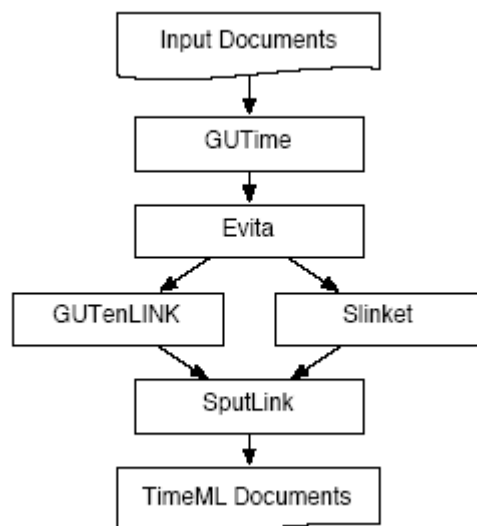
<seg ID='2' CONTINUE='' nuc='yes' h='2' vein='2,4,13' >
  <w pos='JJ'>Federal</w>
  <w pos='NNS'>investigators</w>
  ....
</seg>
    
```

### 3.2. Obținerea adnotării pentru temporalitate

Pentru acest pas, din fișierele obținute anterior au fost eliminate toate ~~ăch~~ rezultatul fiind textul inițial al articolelor din WSJ. Pentru adnotarea automată a acestor r texte cu ajutorul limbajului TimeML, am utilizat instrumentul de adnotare TARSQI [Mani *et al.*, 2005a].

Proiectul TARSQI (Temporal Awareness and Reasoning Systems for Question Interpretation) a fost creat pentru a îmbunătăți sistemele tip întrebare -răspuns astfel încât acestea să poată trata corespunzător întrebări despre evenimente și entități din articolele de ziar cu referire la plasarea acestora în timp. O adnotare manuală completă pentru TimeML nu este fezabilă datorită complexității mari și a numărului mare de documente care trebuiesc procesate. TARSQI poate fi utilizat ca instrument de sine sătător sau ca un ajutor pentru cei care realizează adnotarea manuală a textelor.

Sistemul este compus din mai multe module dezvoltate în Java, Perl, Python și Prolog și execuția în cascadă a fiecăruia modifică textul inițial și returnează adnotarea lui cu TimeML. La intrare, TARSQI are nevoie de text adnotat ~~părțile de~~ părțile de vorbire ale cuvintelor. Această adnotare a fost obținută cu ajutorul POS-tagger-ului TreeTagger [Schmid, 1994], dezvoltat de Universitatea din Stuttgart.



**Fig. 4.** Arhitectura utilitarului pentru adnotarea automată a temporalității, TARSQI

În cele ce urmează, voi oferi o scurtă descriere a modulelor ce au fost utilizate pentru adnotarea temporalității în corpusul ales. Astfel, tagger-ul **GUTime**<sup>3</sup>, dezvoltat la Georgetown University, extinde capabilitățile tagger -ului TempEx [Mani și Wilson, 2000] dezvoltat de MITRE, permițând recunoașterea duratei și a valorilor normalizate pentru expresii temporale, într-o formă standardizată. Acest modul prelucrează atât valori temporale absolute (de ex.: 2 Iunie 2008), cât și valori relative (de ex.: *Vineri*), în urma unui număr de teste pe care le aplică contextului local. Marcatori lexicali precum *ieri*, *mâine*, *luna viitoare*, *săptămâna trecută*, sunt determinați pe baza calculării direcției și magnitudinii față de un timp referință, care de obicei este data la care a fost publicat documentul.

**Evita** (Events in Text Analyzer) este un instrument pentru recunoașterea evenimentelor care are două utilizări de bază: recunoașterea robustă a evenimentelor și analiza unor indicii gramaticale, cum ar fi timpul și aspectul (de ex: *aspect continuu*).

**GUTenLINK** parsează rezultatul obținut în urma aplicării modulelor descrise anterior pe documentul inițial și adaugă tag-uri de tipul TLINK pe baza unor reguli sintactice și lexicale dezvoltate manual. GUTenLINK folosește reguli prestabilite pentru ordonarea evenimentelor.

**Slinket** (SLINK Events in Text) este un parser de recunoaștere a legăturilor de subordonare SLINK din TimeML, implementat în Python, bazat pe identificatorul de evenimente Evita, deci implicit pe informație morfo-sintactică. Pentru un eveniment identificat, folosind reguli lexicale și sintactice, parserul îi atribuie un grad de certitudine asupra factualității sale, specificând dacă evenimentul este factiv, contra-factiv, evidențial sau modal.

**SputLink** este o componentă de închidere temporală care ia relații temporale cunoscute din text și derivă noi relații implicate de acestea, de fapt, făcând explicit ceea ce era implicit. O astfel de componentă ajută la găsirea unor legături temporale globale, care nu ar fi putut fi determinate cu ajutorul altor metode.

O dată cu versiunea 1.2.1. a specificației TimeML în TARSQI a fost introdus componenta **S2T** (SLINK to TLINK). Scopul acesteia este să creeze noi legături temporale din legăturile de subordonare adnotate anterior. Adesea există relații temporale între evenimentele

---

<sup>3</sup> Informații suplimentare pot fi obținute vizitând adresa <http://timeml.org>

ce creează o legătură de subordonare care nu sunt capturate de celelalte componente. Din cele șase tipuri de relații SLINK, doar legăturile *factive*, *evidential* și *modal* sunt eligibile pentru crearea de noi legături temporale. S2T utilizează câteva reguli de creare a legăturilor temporale bazate pe informații legate de timp și aspect conținute în instanțele evenimentelor participante. S2T primește la intrare un document TimeML adnotat cu tagurile EVENT și SLINK și returnează noile taguri TLINK generate.

Mai jos putem observa un exemplu de adnotare temporală a secvența de text „Federal investigators have identified the problem in last July.”:

```
<s>
  <NG>
    <lex pos="JJ">Federal</lex>
    <lex pos="NNS">investigators</lex>
  </NG>
  <VG>
    <lex pos="VBP">have</lex>
    <lex pos="VBN">
      <EVENT eid="e1" class="OCCURRENCE">identified</EVENT>
    </lex>
    <MAKEINSTANCE eventID="e1" polarity="POS" pos="VERB"
      eiid="e1" tense="PRESENT" aspect="PERFECTIVE"/>
  </VG>
  <NG>
    <lex pos="DT">the</lex>
    <lex pos="NN">problem</lex>
  </NG>
  <lex pos="IN">in</lex>
  <NG>
    <TIMEX3 tid="t1" TYPE="DATE" VAL="200707">
      <lex pos="JJ">last</lex>
      <lex pos="NNP">July</lex>
    </TIMEX3>
</s>
```

### 3.3. Obținerea corpusului final

Corpusul final trebuie să conțină fișierele adnotate pentru „nervuri” la care se adaugă printr-o operație de „merge” informațiile temporale (tag-urile EVENT, MAKEINSTANCE, TIMEX3 și TLINK). Într-o primă fază au fost eliminate, din fișierele care conțin adnotările pentru temporalitate, toate tag-urile nespecifice acestui scop (tag-uri pentru cuvinte, leme, părți de vorbire, etc.). Pentru a putea realiza operația de „merge” între două fișiere tip xml, aplicația Java creează cere ca textele obținute în urma eliminării tuturor adnotărilor să fie identice. Textele originale WSJ au fost adnotate pentru părțile de vorbire utilizând POSagger -ul Qtag. Utilitarul TARSQI se bazează pe Tree Tagger. Qtag și TreeTagger procesează textul diferit,

astfel încât în momentul în care se elimină toate adnotările apar porțiuni de text care diferă (de ex: *hasn't și hasn't* sau *\$ 1, 000, 000 și \$1,000,000*). Pentru a elimina aceste neconcordanțe am utilizat o serie de expresii regulate. Există și cazuri de excepție când operațiunea de „merge” nu poate returna un rezultat satisfăcător în urma rulării automate, astfel încât într-o ultimă fază, de validare, fișierele din corpus au fost inspectate manual.

Mai jos prezentăm o secvență dintr-un fișier din corpusul final:

```
<Root>
  <rel ID="0" nuc="no" h="13" vein="13">
    <rel ID="1" nuc="yes" h="2" vein="13">
      <seg ID="2" CONTINUE="" nuc="yes" h="2" vein="13">
        <w pos="JJ">Federal </w>
        <w pos="NNS">investigators </w>
        <w pos="HV">have </w>
        <w pos="VBN">
          <EVENT eid="e1" class="OCCURRENCE">identified</EVENT>
          <MAKEINSTANCE eventID="e1" polarity="POS" pos="VERB" eiid="ei1"
            tense="PRESENT" aspect="PERFECTIVE"/>
        </w>
        <w pos="DT"> the </w>
        <w pos="NN">problem </w>
        <w pos="IN">in </w>
        <TIMEX3 tid="t1" TYPE="DATE" VAL="200707">
          <w pos="OD">last </w>
          <w pos="NN">July`s </w>
          <w pos="VB">
            <EVENT eid="e2" class="OCCURRENCE">crash</EVENT>
            <MAKEINSTANCE eventID="e2" polarity="POS" pos="NOUN"
              eiid="ei2" tense="NONE" aspect="NONE"/>
          </w>
        </TIMEX3>
        ...
      </seg>
    </rel>
  </rel>
  <TLINK relatedToTime="t3" lid="197" relType="BEFORE" eventInstanceID="ei19"
    origin="CLASSIFIER 0.999950"/>
  <TLINK lid="186" relatedToEventInstance="ei5" relType="BEFORE" eventInstanceID="ei4"
    origin="CLASSIFIER 0.998944"/>
</Root>
```

La crearea corpusului propus de Daniel Marcu [Marcu *et al.*, 1999] au participat mai mulți experți, iar adnotarea manuală s-a realizat eșantionat, pe o perioadă îndelungată de timp (aprilie 2000 – ianuarie 2001). Aproximativ un sfert din textele corpusului au fost dublu adnotate. Multiple îmbunătățiri au fost adăugate cu fiecare revizuire a adnotării, iar rezultatul final a fost de 97% acord între adnotatori. Performanțele instrumentului de adnotare automată TARSQI pot fi măsurate pentru fiecare modul în parte. Astfel, în analiza prezentată în [Mani *et al.*, 2005a] GUTime a obținut F-measure 0.85, EVITA are precizie 0.75, recall 0.87 și F-measure 0.8, iar GUTenLINK prezintă o precizie de 0.75. Deoarece la adnotarea automată a

corpusului pe care l-am propus, modulele amintite mai sus au fost rulate unul câte unul, în cascadă, la fiecare pas în adnotare s-au strecurat tot mai multe erori. Astfel, printr-un calcul intuitiv simplu - înmulțind preciziile fiecărui modul - , putem spune că precizia finală obținută este de 0.42. Nervurile au fost calculate pe baza timpului în segmente RST, utilizând formulele descrise în secțiunea 2.2.2.

În procesul de adnotare automată pentru temporalitate, după ce am rulat pe textele WSJ fiecare modul descris în secțiunea 3.2, am obținut o adnotare bogată în TLINK-uri. Rezultatele sunt prezentate în **Tabelul 2**. TARSQI conține un ultimul modul numit LinkMerger. Acesta citește un fișier din corpus, după care creează un graf fără muchii și separat o listă ordonată cu toate muchiile (acestea fiind definite de tag-uri de tip TLINK). Aceste muchii sunt adăugate în graf una câte una, rulând închiderea tranzitivă după fiecare adăugare pentru a verifica consistența noului graf obținut. După ce toate muchiile au fost adăugate, graful este redus și TLINK-urile rezultate sunt rescrise în fișierul inițial. În cadrul rulării TARSQI pe un fișier de intrare, sunt mai mult reguli care generează TLINK-uri, astfel că pot exista situații în care un TLINK să apară marcat de două ori în același fișier. La pasul de reducere al grafului se normalizează muchiile, se marchează inversele relațiilor existente și se elimină toate TLINK-urile duplicate. Este posibil ca datorită regulilor de generare a TLINK-urilor, aceleași două evenimente să apară adnotate de două ori, dar cu un tip de relații (așa cum sunt descrise în secțiunea 2.3.2.3.) diferite între ele. De exemplu, aceleași două evenimente pot fi adnotate ca fiind în relație de tipul BEFORE într-un TLINK și în relație IS\_INCLUDED în alt TLINK. Normalizarea muchiilor rezolvă, teoretic, această problemă. Teoretic, deoarece din rezultatele obținute practic (evidențiate în **Tabelul 3**), am observat că după rularea modului LinkMerger sunt eliminate și TLINK-uri bune, care ar fi adus informații în plus despre temporalitate. Astfel, în final, am decis că pentru scopul acestei lucrări este mai bine să alegem corpusul obținut din TARSQI fără LinkMerger. Vom arăta în acest capitol că informația suplimentară pe care o aduc nervurile poate substitui funcționalitatea acestui modul.

ALINK	SLINK	TLINK	TIMEX3	EVENT	MAKEINSTANCE
49	947	10999	910	6714	6714

**Tabel 2.** Statistici obținute pe corpus fără LinkMerger

ALINK	SLINK	TLINK	TIMEX3	EVENT	MAKEINSTANCE
49	947	6536	910	6714	6714

**Tabel 3.** Statistici obținute pe corpus cu LinkMerger

Prezentăm în tabelul de mai jos o distribuție a TLINK-urilor după tipul de relație dintre evenimentele și expresiile temporale pe care le reprezintă.

AFTER	1153
BEFORE	8096
BEGINS	8
BEGUN_BY	6
DURING	1
ENDED_BY	1
ENDS	0
IAFTER	0
IBEFORE	8
IDENTITY	186
INCLUDES	1187
IS_INCLUDED	305
SIMULTANEOUS	19
TOTAL	10970

**Tabel 4.** Statistici privind distribuția TLINK-urilor în funcție de tipul relației pe care le conțin

Diferența dintre totalul TLINK-urilor obținute în **Tabelul 4** comparativ cu totalul TLINK-urilor din **Tabelul 2** apare datorită faptului că la adnotarea automată unele TLINK-uri apar fără atributul „relType”, care definește tipul de relație. Acest lucru poate fi datorat unei erori din TARSQI sau, mai sigur, faptului că tipul de relație nu a putut fi determinată cu exactitate.

În capitolul următor, vom face o analiză detaliată a problemelor apărute la analiza corpusului și de asemenea, o analiză a îmbunătățirilor aduse adnotării automate a temporalității prin utilizarea nervurilor.

## 4. Analiza temporalității în relație cu teoria nervurilor

În cele ce urmează, vom investiga legătura dintre temporalitate și structura de discurs. Dacă există o astfel de legătură, o vom putea utiliza pentru a reduce efortul uman în cadrul adnotării manuale a relațiilor temporale, vom putea îmbunătăți adnotarea automată a unui text pentru relații temporale și vom putea îmbunătăți parsarea unui discurs.

Teoria nervurilor susține faptul că există o strânsă legătură între structura de discurs și referențialitate. Acest fapt a fost dovedit experimental [Cristea, 2003]. Mai mult decât atât, acest rezultat a fost utilizat pentru:

- a recupera mai ușor și mai sigur relații referențiale când structura nervurilor este cunoscută;
- a parsa discursul atunci când relațiile referențiale sunt cunoscute.

Vom încerca în cele ce urmează să vedem dacă, de asemenea, putem găsi o legătură între nervuri și relațiile temporale dintr-un text. Modul în care am definit nervurile intuiește că o astfel de legătură ar exista. Dacă demonstrăm experimental că acest lucru este adevărat, atunci putem să folosim acest rezultat pentru:

- a recupera mai ușor și mai sigur relații temporale când structura discursului este cunoscută;
- a parsa discursul, când relațiile temporale sunt cunoscute.

## 4.1. Probleme în procesul adnotării

Pentru a putea realiza o analiză cât mai detaliată a rezultatelor obținute, am utilizat un corpus gold format din 7 fișiere TimeBank, care apar și în corpusul descris în capitolul 3. TimeBank conține 183 de articole de știri adnotate manual cu standardul TimeML 1.2.

Am comparat în paralel aceste fișiere din TimeBank cu fișierele corespunzătoare care conțin adnotarea făcută de TARSQI. În **Tabelul 5** este prezentat un exemplu complet, iar apoi sunt prezentate și exemplificate punctual problemele întâlnite în toate fișierele.

Exemplul din **Tabelul 5** reprezintă analiza textului: *„A group of investors led by Giant Group Ltd. and its chairman, Burt Sugarman, said it filed with federal antitrust regulators for clearance to buy more than 50% of the stock of Rally`s Inc., a fast-food company based in Louisville, Ky. Rally`s operates and franchises about 160 fast-food restaurants throughout the U.S. The company went public earlier this month, offering 1,745,000 shares of common stock at \$15 a share. Giant has interests in cement making and newsprint. The investor group includes Restaurant Investment Partnership, a California general partnership, and three Rally`s directors: Mr. Sugarman, James M. Trotter III and William E. Trotter II. The group currently holds 3,027,330 Rally`s shares, or 45.2% of its common shares outstanding. Giant Group owned 22% of Rally`s shares before the initial public offering. A second group of three company directors, aligned with Rally`s founder James Patterson, also is seeking control of the fast-food chain. It is estimated that the Patterson group controls more than 40% of Rally`s stock. Rally officials weren`t available to comment late yesterday. For the year ended July 2, Rally had net income of \$2.4 million, or 34 cents a share, on revenue of \$52.9 million.”*

Corpus TimeBank			Corpus WSJ (TARSQI)		
Lema	Tip	Clasa	Lema	Tip	Clasa
			led	EVENT	OCCURRENCE
said	EVENT	REPORTING	said	EVENT	REPORTING
filed	EVENT	I_ACTION	filed	EVENT	OCCURRENCE
clearance	EVENT	I_ACTION			
buy	EVENT	OCCURRENCE	buy	EVENT	OCCURRENCE
			based	EVENT	OCCURRENCE
			Rally	EVENT	OCCURRENCE
			operates	EVENT	OCCURRENCE
			franchises	EVENT	OCCURRENCE
went	EVENT	OCCURRENCE	went	EVENT	OCCURRENCE
this month	TIMEX3	TIME	this month	TIMEX3	DATE
offering	EVENT	OCCURRENCE	offering	EVENT	I_ACTION
			has	EVENT	OCCURRENCE
			making	EVENT	OCCURRENCE
			includes	EVENT	OCCURRENCE
			Rally	EVENT	OCCURRENCE
			holds	EVENT	OCCURRENCE
			Rally	EVENT	OCCURRENCE
owned	EVENT	STATE	owned	EVENT	OCCURRENCE
offering	EVENT	OCCURRENCE			
			aligned	EVENT	OCCURRENCE
seeking	EVENT	I_ACTION	seeking	EVENT	I_ACTION
control	EVENT	STATE			
			estimated	EVENT	OCCURRENCE
controls	EVENT	STATE	controls	EVENT	OCCURRENCE
available	EVENT	STATE			
comment	EVENT	OCCURRENCE	comment	EVENT	OCCURRENCE
yesterday	TIMEX3	DATE	yesterday	TIMEX3	DATE
the year	TIMEX3	DURATION			
July 2	TIMEX3	DATE			
had	EVENT	OCCURRENCE	had	EVENT	OCCURRENCE

**Tabel 5.** Comparație între o adnotare TimeBank și una WSJ

Din cele 41 de evenimente și expresii temporale evidențiate de adnotatorii umani și de TARSQI, doar 7 au fost adnotate perfect identic. În alte 5 cazuri, lema care determină evenimentul a fost găsită ca fiind aceeași, dar clasa din care face parte a fost adnotată diferit. Aceasta este oarecum normal pentru că un instrument de adnotare automată nu poate să deducă raționamente specifice în legătură cu lemele pe care le găsește, așa că cea mai întâlnită clasă în cazul adnotării automate rămâne OCCURRENCE. Aceasta este clasa care desemnează faptul că evenimentul descris de lema corespunzătoare nu a putut fi inclus într-o altă clasă. În alte două situații, adnotatorul automat omite să marcheze două expresii temporale deosebit de importante (*the year*, *July 2*) și alte 4 evenimente. În schimb, TARSQI găsește în plus față de documentul TimeBank 12 evenimente. Dintre acestea 3 sunt cu siguranță greșite, și anume, cele care conțin lema *Rally*. În textul analizat, *Rally* este numele unei companii, dar în limba engleză *to rally* este de asemenea un verb. Cuvântul *Rally* apare de 8 ori în text, dar este

adnotat ca și eveniment doar de 3 ori. Introducerea în TARSQI a unui modul de detecție a entităților care conține nume proprii ar reduce numărul erorilor de acest gen.

Faptul că TARSQI marchează mult mai multe verbe ca fiind evenimente nu este greșit, dar conduce la obținerea unui număr de legături temporale mult mai mare față de o adnotare manuală paralelă, ceea ce face mai dificil procesul de evaluare al adnotării automate.

O altă situație neplăcută apărută în cazul adnotării automate se datorează POS - Tagger-ului utilizat. Astfel, datorită unor spații în plus introduse de acesta în primele faze ale adnotării, în final ajungem ca în anumite situații să avem o dată calendaristică adnotată ca două expresii temporale diferite. De exemplu, pentru data *Oct. 15, 1989*, un adnotator manual ar crea un `<TIMEX3 tid="t1"> Oct. 15, 1989 </TIMEX3>`, pe când TARSQI creează

```
<TIMEX3 tid="t1"> Oct. 15 </TIMEX3>  
<w pos=","></w>  
<TIMEX3 tid="t2"> 1989 </TIMEX3>.
```

Trebuie atrasă atenția asupra faptului că așa cum am prezentat în **Tabelul 4**, TARSQI marchează taguri TLINK, care definesc preponderent între evenimente relații de tipul BEFORE, AFTER, INCLUDES și IS\_INCLUDED. În analiza pe care am realizat-o am normalizat aceste relații, astfel încât evenimentul *e1* AFTER *e2*, a fost înlocuit cu *e2* BEFORE *e1*. În adnotarea manuală, relațiile de tipul BEFORE rămân preponderente, dar într-o proporție mult mai mică. Vom încerca să probăm faptul că o parte din aceste probleme pot fi rezolvate cu ajutorul nervurilor.

## 4.2. Marcarea tagului SIGNAL

TARSQI nu face adnotarea automată a tagului SIGNAL, însă acest tag face parte din TimeML și este foarte important datorită informațiilor suplimentare pe care le aduce asupra evenimentelor și a relațiilor temporale. O descriere detaliată a tagului SIGNAL este oferită în secțiunea 2.3.2.5.

Am creat un modul Java care să completeze instrumentul de adnotare prin găsirea și marcarea automată a tagului SIGNAL. Într-o primă fază am extras din corpusul TimeBank o listă cu toate cuvintele și expresiile care au fost marcate ca fiind SIGNAL de către adnotatorii umani. Lista completă poate fi studiată în Anexa 1 a acestei lucrări.

Din experimentele realizate, am observat că orice SIGNAL se găsește înainte de un eveniment sau expresie temporală, la maxim 4-5 cuvinte distanță de lema care definește acest

eveniment și atât SIGNAL-ul, cât și evenimentul sau expresia temporală fac parte din același segment (unitate elementară de discurs) RST.

Astfel, având textul marcat pentru RST și evenimentele marcate manual sau automat, într-o primă trecere se rețin într-o listă toate cuvintele care sunt posibile SIGNAL-uri împreună cu segmentul pe care se află. La o a doua trecere se verifică dacă, în același segment, după un cuvânt marcat la pasul anterior urmează un eveniment. Dacă da, acest cuvânt se marchează ca fiind SIGNAL și el primește ca atribut un signalID, care va fi incrementat cu fiecare SIGNAL nou adăugat. Celelalte cuvinte marcate care nu au un eveniment pe care să-l semnaleze pe același segment vor fi ignorate. Într-o ultimă fază, se parcurge lista de TLINK-uri și pentru acele legături temporale care conțin un eveniment precedat de un SIGNAL va fi adăugat atributul sigID, ce va conține ID-ul respectivului SIGNAL.

Pentru a evalua acuratețea metodei descrise mai sus, am utilizat ca referințe 7 documente din TimeBank adnotate automat. O comparație în paralel, pe fiecare document în parte și per total, față de adnotarea automată a tagului SIGNAL poate fi studiată în **Tabelul 6**.

Index fișier	Manual	Automat	Comune
0	36	41	16
1	4	7	3
2	1	1	0
3	13	16	11
4	5	5	3
5	1	3	1
6	9	5	4
Total	69	78	38

**Tabel 6.** Paralelă între numărul de taguri SIGNAL pentru adnotare manuală și automată

Evaluarea adnotării pentru SIGNAL s-a realizat utilizând formulele  $P=t_p/(t_p+f_p)$ ,  $R=t_p/(t_p+f_n)$  și  $F=2*(P*R)/(P+R)$ , unde P este precizia, R este scorul pentru Recall, F reprezintă valoarea F-measure, iar  $t_p$  înseamnă numărul de elemente *true positive*,  $f_p$  sunt elementele *false positive*, iar  $f_n$  sunt cele *false negative*. Se obțin, pe baza informațiilor prezentate mai sus,  $P=0.53$ ,  $R=0.55$ , iar  $F=0.53$ . Aceasta se datorează în primul rând faptului că în adnotarea tagului SIGNAL se ține cont de evenimente, iar TARSQI adnotă mai multe evenimente comparativ cu cele marcate de adnotatorii umani în TimeBank. Astfel că nu toate tagurile SIGNAL care apar în plus sunt greșite. Pentru o evaluare mai realistă asupra preciziei acestei metode de adnotare, am eliminat manual din fișierele adnotate de TARSQI evenimentele care nu au fost semnalate de adnotatorii umani. Acuratețea finală rezultată a fost de 92%.

Diferența care încă mai rămâne se datorează faptului că TARSQI nu adnotează unele evenimente pe care adnotatorii umani le-au considerat importante. Mai există situația în care există două posibile SIGNAL-uri înainte de un eveniment, iar instrumentul de adnotare automată îl alege pe cel mai apropiat de eveniment. De exemplu, pentru fragmentul de text *not yet seem*, adnotatorii umani au marcat *seem* ca eveniment și *not* ca SIGNAL. Atât *not*, cât și *yet* sunt posibile SIGNAL-uri care pot fi adnotate, însă automat, după algoritmul descris mai înainte, va fi selectat cel mai apropiat, deci în acest caz, cel greșit, adică *yet*.

O altă situație delicată care duce la pierderea acurateții este determinată de construcțiile din limba engleză care se termină în *n't*, cum ar fi *isn't*, *hasn't*, *weren't*, etc. În TimeBank, adnotatorii manuali au considerat *n't* ca fiind SIGNAL și au despărțit tipul de construcții amintit în părțile componente – *isn't* devine *is* și *n't*, ca părți de vorbire separate – marcând astfel *n't* ca SIGNAL. Din păcate, POSTagger -ul utilizat de TARSQI marchează aceste construcții ca fiind o singură parte de vorbire.

În cazul unor expresii temporale, cum ar fi *last year*, modulul de adnotare automată găsește corect *last* ca fiind SIGNAL și *year* ca fiind TIMEX. Însă în TimeBank astfel de construcții sunt marcate TIMEX ca întreg, fără ca *last* să fie considerat ca semnalând expresia temporală *year*.

Ținând cont de situațiile excepționale descrise anterior, putem spune că este posibilă o adnotare automată a tagului SIGNAL cu o precizie foarte bună, de peste 95% adoptând o adnotare automată verificată manual.

### 4.3. Închiderea tranzitivă a temporalității

O componentă de închidere temporală ajută la crearea unei adnotări care să fie completă și consistentă. Este nevoie de adnotarea explicită a temporalității pentru aplicațiile utilizate în sumarizare sau pentru sistemele întrebare-răspuns. Încă nu este posibilă crearea unei adnotări temporale de calitate foarte mare. Deci, trebuie să ne bazăm într-o anumită măsură pe adnotarea manuală. Adnotatorul uman poate observa rapid cum se relaționează anumite evenimente în timp, fără a avea nevoie neapărat de marcatori textuali expliți și clari. În schimb, calculatorul poate procesa date de dimensiuni foarte mari și poate aplica cu succes anumite reguli de detecție a relațiilor temporale. Închiderea temporală este un aspect deosebit de important care să vină în ajutorul efortului de adnotare. Închiderea temporală ia relații

temporale cunoscute din text și derivă relații noi din acestea, de fapt făcând explicit ceea ce era implicit.

Efortul de adnotare umară este dificil datorită densității mari a informației legată de evenimente și expresii temporale, a vitezei mici de marcare a acestora, a acordului mic între adnotatori și a dificultății de a evita introducerea unor inconsistențe.

Densitatea mare a informației apare ca urmare a faptului că setul de relații temporale posibile este proporțional cu numărul de evenimente și expresii temporale din text. Dacă un document are  $N$  evenimente și expresii temporale, atunci există  $N(N-1)/2$  relații temporale posibile. Un document TimeBank obișnuit conține în jur de 50 obiecte temporale, ceea ce implică 1225 de relații temporale posibile. Documente mai mari cu aproximativ 150 obiecte temporale (evenimente și expresii temporale) au peste 10.000 de relații posibile.

Adnotarea manuală a expresiilor temporale cere adnotatorului uman mai mult timp de gândire decât adnotarea, de exemplu, a părților de vorbire. Tagurile sintactice și semantice, cum ar fi tagul EVENT, pot fi adăugate într-o manieră strict liniară. Relațiile temporale sunt diferite deoarece necesită specificarea atributelor de perechi de obiecte, și e posibil ca obiectele implicate să nu fie apropiate unul față de celălalt în text. Adnotarea unui articol de ziar de lungime medie poate lua peste o jumătate de oră unui adnotator expert, iar adnotarea rezultată nu este completă. În medie un adnotator uman va marca 1-5% din toate relațiile temporale posibile. Dezacordul între adnotatori pe același text se datorează faptului că fiecare adnotează în medie 1-5% din relații, dar dat fiind spațiul foarte mare din care pot alege obiectele pe care să le marcheze, puține relații vor fi comune între cele marcate de aceștia.

Pentru o analiză mai detaliată a închiderii temporale pe corpusul WSJ și TimeBank am utilizat Tango [Pustejovski *et al.*, 2003]. Tango este un program cu interfață grafică ce aduce funcționalități pentru marcarea informațiilor temporale, pentru vizualizarea și aranjarea lor pe o axă a timpului și care incorporează un algoritm de închidere a temporalității, dezvoltat și detaliat de Marc Verhagen în [Verhagen, 2004]. O captură de ecran din Tango poate fi observată în **Figura 5**.

## Temporalitate și referențialitate utilizând teoria nervurilor

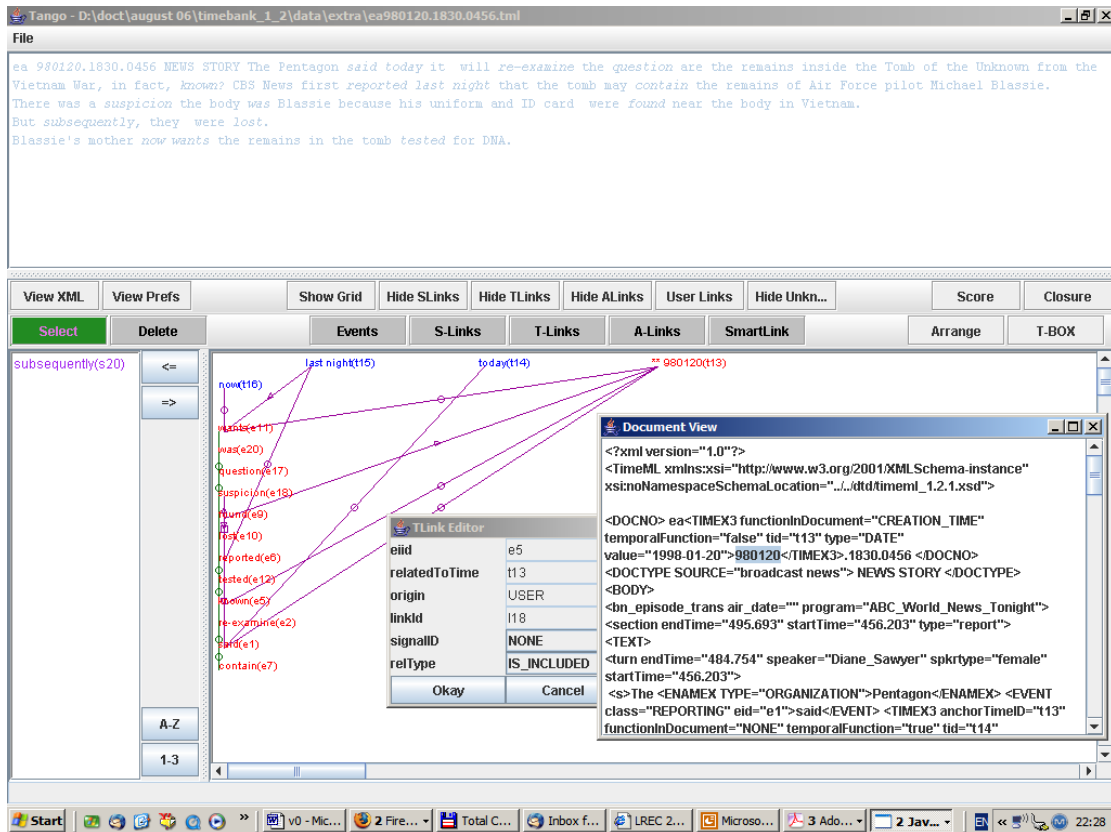


Fig. 5. Captură de ecran a spațiului de lucru din Tango

Algoritmul de închidere a temporalității este bazat pe calculul cu algebra intervalelor introdus de Allen [Allen, 1983]. Dezvoltarea acestui tip de calcul [Allen, 1984] a avut o influență majoră în domeniul cercetării temporalității. Există 13 relații temporale de bază între două intervale, așa cum este arătat în figura 5, care evidențiază 7 relații și 6 dintre inversele acestora.

Relație	Simbol	Inversa	Exemplu
X before Y	<	>	
X meets Y	m	mi	
X overlaps Y	o	oi	
X during Y	d	di	
X starts Y	s	si	
X finishes Y	f	fi	
X equal Y	=	=	

Fig. 6. Cele 13 relații de bază din algebra lui Allen

Fiecare interval poate fi reprezentat ca o pereche de puncte unde unul îl precede pe celălalt. De exemplu, intervalul A poate fi scris ca  $a_1 - a_2$ , unde  $a_1$  este punctul de început,  $a_2$  este punctul final și  $a_1 < a_2$ . Toate relațiile de bază prezentate mai sus pot fi rescrise utilizând relații de precedență și egalitate. De exemplu, *A before B* este echivalent cu  $a_2 < b_1$  și *A starts B* este echivalent cu  $a_1 = b_1 \wedge a_2 < b_2$  (unde „ $\wedge$ ” reprezintă operatorul logic „și”).

Se presupune că toate relațiile din TimeML pot fi mapate după relațiile lui Allen și după relații între puncte. O traducere a tuturor relațiilor din TimeML în acest mod este oferită în **Tabelul 7**.

Relatii TimeML	Relatii Allen	Relatii între puncte
A before B	<	$a_2 < b_1$
A after B	>	$a_1 < b_2$
A ibefore B	m	$a_2 = b_1$
A iafter B	mi	$b_2 = a_1$
A includes B	di	$a_1 < b_1 \wedge a_2 > b_2$
A is_included B	d	$a_1 > b_1 \wedge a_2 < b_2$
A identity B A simultaneous B A holds B A held_by B	=	$a_1 = b_1 \wedge a_2 = b_2$
A begins B	s	$a_1 = b_1 \wedge a_2 < b_2$
A begun_by B	si	$a_1 = b_1 \wedge a_2 > b_2$
A ends B	f	$a_1 > b_1 \wedge a_2 = b_2$
A ended_by B	fi	$b_1 > a_1 \wedge a_2 = b_2$

**Tabel 7.** Maparea relațiilor din TimeML la algebra lui Allen

Pentru a putea face o evaluare cât mai realistă a îmbunătățirilor aduse prin aplicarea închiderii tranzitive a temporalității pe corpusul propus, am trecut cele 7 documente comune cu TimeBank prin 2 faze de postprocesare. Într-o primă fază, am eliminat din toate fișierele TLINK-urile care conțineau legături ce nu puteau fi realizate pe nervuri. Acest lucru a fost realizat pe baza următorului raționament:

- fiecare eveniment se află pe un segment (definit de tagul *seg*);
- fiecare segment are atributele „head” și „vein”, care definesc nervura pe care se află;
- spunem despre două obiecte (evenimente sau expresii temporale) care definesc o legătură temporală (TLINK) că se găsesc pe nervuri dacă expresia „head” a segmentului în care se găsește primul obiect se regăsește în expresia „vein” a segmentului în care se găsește cel de-al doilea obiect.

În a doua fază, am eliminat din fișierele inițiale toate TLINK-urile găsite de TARSQI și am creat TLINK-uri pentru toate legăturile ce au putut fi determinate între evenimente și relații temporale doar pe nervuri. Rezultatele analizei pot fi studiate în **Tabelul 8**.

		Inițial	După faza I	Procentaj	După faza II	Procentaj
Număr TLINK-uri	Înainte de închidere	758	272	35.90%	817	107%
	După închidere	4350	365	8.40%	1434	33%

**Tabel 8.** Analiza închiderii temporalității

Vom explica mai detaliat rezultatele obținute în secțiunea următoare.

#### 4.4. Distanța medie între legăturile temporale

În această secțiune vom prezenta statistici care demonstrează că închiderea temporală adaugă adnotării TLINK-uri nelocale și că aceste legături erau în mare parte absente din adnotare înainte de închidere.

Adnotatorii care au marcat corpusul TimeBank pară să -au bazat pe strategii de adnotare similare legând evenimentele de alte evenimente și expresii temporale care erau în fragmentul de text cel mai apropiat. Rezultatele obținute pentru adnotarea automat demonstrează că o astfel de strategie a fost implementată și în algoritmi utilizați de TARSQI, marea majoritate a legăturilor temporale fiind realizate între evenimente sau expresii temporale consecutive ca ordine a apariție în text.

Textul corpusului propus este împărțit în segmente (de obicei o propoziție sau o frază). Dacă un TLINK conține un eveniment din același segment atunci distanța liniară între evenimente este 0; dacă evenimentele depășesc o limită de text (un segment), atunci distanța liniară este 1 și așa mai departe. Distanța medie pentru un document se calculează fiind suma tuturor distanțelor liniare obținute pentru fiecare legătură temporală împărțită la numărul legăturilor temporale. **Tabelul 9** conține distanțele medii între legături temporale pentru corpusul TimeBank și pentru corpusul WSJ, atât înainte de închiderea tranzitivă, cât și după aplicarea acesteia.

		TimeBank 1.1	După faza I	După faza II
Distanța medie	Adnotare inițială	2.42	7.88	39.84
	După închidere	6.89	37.34	89.09

**Tabel 9.** Distanța medie între legăturile temporale

Pentru TimeBank, după închiderea tranzitivă, distanța medie crește de la 2.42 la 6.89. Aceste valori evidențiază faptul că închiderea tranzitivă adaugă un întreg grup de legături nelocale care au fost omise sistematic de către adnotatori. Distanța medie între legături este, evident, direct proporțională cu mărimea documentului. După faza I de preprocesare, analizând distanța medie pe fișierele ce conțin legăturile temporale rămase după eliminarea celor care nu se găsesc pe nervuri, putem observa o diferență mare față de TimeBank. Acest lucru se datorează faptului că TARSQI adnotează mult mai multe evenimente, deci automat numărul posibil de relații între acestea crește simțitor. După faza II – în care păstrăm toate legăturile găsite pe nervuri – observăm că distanța medie obținută înainte de aplicarea închiderii tranzitive este chiar mai mare decât distanța medie obținută în faza I după închiderea tranzitivă. Acest rezultat demonstrează clar faptul că urmărind legăturile dintre evenimente și expresii temporale doar pe nervuri pot fi găsite legături la distanțe mari una de alta. Crearea de legături temporale urmărind nervurile este cea mai eficientă metodă de a găsi relații temporale greu de depistat utilizând adnotarea manuală sau oricare altă metodă de adnotare automată existentă.

În **Tabelul 10** poate fi observat faptul că distanța medie între legături crește proporțional cu mărimea documentului pe care se face analiza. Mărimea documentului este dată de numărul de evenimente și expresii temporale.

Nr. Obiecte temporale	Înainte închiderii	După închidere
8	1.6	1.6
9	2.06	2
21	2.36	6.03
25	3.88	4.43
27	8.87	11.39
41	6.35	9.07
144	14.68	54.55

**Tabel 10.** Distanța între legături pentru documente de mărimi diferite din corpul WSJ

## **5. Concluzii**

### **5.1. Contribuții**

Am descris în această lucrare o analiză a modului în care extragerea informațiilor temporale din text poate fi îmbinată cu teoria nervurilor. Am creat un corpus de articole extrase din Wall Street Journal adnotate automat pentru temporalitate și nervuri. Am arătat faptul că această adnotare este incompletă și conține inconsistențe. Am adus îmbunătățiri instrumentului utilizat pentru adnotarea automată a temporalității (TARSQI), obținând adnotări cu o acuratețe de peste 92% pentru tagul SIGNAL. Evaluarea întregului proces de adnotare s-a realizat utilizând rezultatele obținute în urma adnotării manuale a unor texte din corpusul propus (secvență de corpus gold din TimeBank). Am demonstrat că teoria nervurilor poate găsi legături temporale între evenimentele unui text pe care sistemele actuale de adnotare automată sau chiar adnotatorii umani nu le pot găsi.

### **5.2. Perspective de viitor**

Sistemul TARSQI utilizează un set de reguli complexe pentru adnotarea cât mai corectă și completă a temporalității. Relațiile temporale găsite cu ajutorul nervurilor, de către sistemul automat creat, au tipul de relație între evenimentele componente setat implicit pe „BEFORE”. Acest tip de relație este cel mai predominant în cadrul legăturilor temporale, dar nu este singurul. Acest lucru duce la generarea unei ordonări în timp parțial incorectă a evenimentelor găsite pe nervuri, dar nu afectează studiul propus în această lucrare. O îmbunătățire substanțială ce poate fi adusă acestui studiu o constituie scrierea unui program

care să implementeze un set de reguli pentru a genera cât mai precis tipul de relație temporală dintre două evenimente găsite cu ajutorul nervurilor.

Pentru a putea continua cu ușurință cercetările în acest domeniu ar putea fi adăugată programelor deja existente o interfață grafică intuitivă care să permită încărcarea unui text din corpus și apoi prelucrarea lui cu posibilitatea de a vedea textul împărțit pe nervuri și de a adăuga relații temporale între evenimentele existente pe aceste nervuri.

## Bibliografie

Allen J. F. – „**Maintaining Knowledge about Temporal Intervals**”, în *Communications of the ACM*, 26(11):832–843, 1983

Allen J. F. – „**Towards a General Theory of Action and Time**” în *Artificial Intelligence* 23: 123-154, 1984

Cristea D., Ide N., and Romary L. – „**Veins Theory: A Model of Global Discourse Cohesion and Coherence**” în *Proceedings of the 17th Coling and the 36th Annual Meeting of the ACL (COLINGACL'98)*. Montreal, CA, (pp. 281-85), 1998

Cristea D. – „**The relationship between discourse structure and referentiality in Veins Theory**”, în *W. Menzel and C. Vertan (Eds.): Natural Language Processing between Linguistic Inquiry and System Engineering*, „Al.I.Cuza” University Publishing House, Iași, 2003

Cristea D. – „**Motivations and Implications of Veins Theory**”, în *Natural Language Understanding and Cognitive Science, Proceedings of the 2nd International Workshop on Natural Language Understanding and Cognitive Science*, NLUCS, 2005

Ferro L., Mani I., Sundheim B., Wilson G. – „**TIDES Temporal Annotation Guidelines Draft - Version 1.02**”. MITRE Technical Report MTR 01W000004. McLean, Virginia, 2001

Grosz B., Joshi A., Weinstein S. – „**Centering: A Framework for Modeling the Local Coherence of Discourse**” în *Computational Linguistics*, 1995

Grosz, B.J., Sidner, C. – „**Attention, intentions, and the structure of discourse**” în *Computational Linguistics*, 12(3):175-204, 1986

Pustejovsky J., R. Gaizauskas, R. Sauri, A. Setzer, R. Ingria – „**Annotation Guideline to TimeML 1.0**.”, 2002, disponibilă la <http://timeml.org>

Mani I., Wilson G. – „**Processing of News**” în *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000)*. Pag. 69-76, 2000

Mani I. – „**Automatic Summarization**”, în *Natural Language Processing*, John Benjamins Publishing Company, 2001

Mani I., Verhagen M., Sauri R., Knippen R., S. B. Jang, Littman J., Rumshisky A., Phillips J., Pustejovsky (2005a) – „**Automating Temporal Annotation with TARSQI**”, 2005

Mani I., Pustejovsky J., Gaizauskas R. (2005b) – „**The Language of Time: A Reader**”. Oxford University Press, ISBN-13: 978-0-19-926853-5, 2005

Mann W., Thompson S. – „**Rhetorical Structure Theory: Toward a functional theory of text organisation**”, 1988

Marcu D., Amorrortu E., Romera M. – „**Experiments in constructing a corpus of discourse trees**” în *Proceedings of the ACL Workshop on Standards and Tools for Discourse Tagging*, 1999

Marcus M., Santorini B., and Marcinkiewicz M. – „**Building a large annotated corpus of English: the Penn Treebank**”, *Computational Linguistics* 19(2), 313-330, 1993

Pistol I. – „**Parsarea automată a discursului lingvistic**”, Lucrare de dizertație, Iași, Iunie 2005

Pustejovsky J., Belanger L., Castaño J., Gaizauskas R., Hanks P., Ingria B., Katz G., Radev D., Rumshisky A., Sanfilippo A., Sauri R., Setzer A., Sundheim B., Verhagen M. – „**NRRC Summer Workshop on Temporal and Event Recognition for QA Systems**”, 2002

Pustejovsky J., Ingria B., – „**TimeML Specification 1.0**”, 2002, <http://timeml.org>

Pustejovsky J., Mani I., Belanger L., van Guilder L., Knippen R., See A., Schwarz J., Verhagen M. – „**TANGO Final Report. Technical report**”, The MITRE Corporation, Bedford, Massachusetts, 2003

Pustejovsky J., Knippen R., Litmann J., Sauri R. (2005a) – „**Temporal and event information in natural language text**”, 2005

Pustejovsky J., Litmann J., Sauri R., Verhagen M. (2005b) – „**Annotating Time and Events in Language**”, 2005

Sauri R., Verhagen M., Pustejovsky J. – „**SlinkET. A Partial Modal Parser for Events**” în *Proceedings of LREC 2006*, Genoa, Italy, pp.1332-1337, 2006

Schmid H. – „**Probabilistic Part-of-Speech Tagging Using Decision Trees**”, *International Conference on New Methods in Language Processing*, 1994

Verhagen M. – „**Times Between The Lines - Embedding a Temporal Closure Component in a Mixed-Initiative Temporal Annotation Framework**”, 2004

## Anexa 1

Lista completă a cuvintelor și expresiilor care vor fi adnotate cu tagul SIGNAL împreună cu frecvența lor de apariție în corpusul TimeBank:

Lema	Frecv	Lema	Frecv	Lema	Frecv	Lema	Frecv
after	56	effective	1	meanwhile	4	soon after	1
ahead of	1	ended	13	on	33	still	4
already	13	ending	1	once	5	subsequent	1
as	14	followed	2	over	14	subsequently	3
as early as	1	followed by	2	not	10	then	5
as of	1	following	4	n't	15	thereafter	1
as soon as	2	follows	3	past	1	through	15
at	11	for	52	pending	1	throughout	2
at least until	1	four times	1	previous	1	to	3
at the same time	4	from	19	previously	11	until	25
before	23	if	37	prior	1	when	35
between	1	immediately	1	prior to	2	while	6
by	20	in	161	repeatedly	1	within	8
can	10	in anticipation of	1	shortly before	1	would	7
before, during and after	1	into	3	since	17	yet	5
during	13	late	3	since then	1	's	8
earlier	6	later	7	so far	1		