

MODELE FORMALE DE REPREZENTARE A INFORMAȚIILOR LEXICALE ȘI TERMINOLOGICE ÎN PROIECTUL *SIASTRO*

1. Introducere

În momentul de față există o varietate largă de aplicații foarte complexe de tratare, cu mijloace informatice, a limbajului natural – NLP [Natural Language Processing]. Așa cum se arată în G. Francopoulo *et alii* (2006a), optimizarea producerii, întreținerii și extinderii resurselor lexicale este unul dintre aspectele cruciale ale aplicațiilor NLP. Un al doilea aspect implică optimizarea procesului de integrare a acestor resurse în aplicațiile lingvistice. De-a lungul timpului s-au produs, cu eforturi mari, numeroase resurse lexicale pentru diverse limbi. Ar fi benefic (Segura Bedmar *et alii*, 2006) ca toate aceste resurse să poată fi integrate pentru a forma resurse globale extinse.

Pe de altă parte, dezvoltarea tehnico-științifică a dus la apariția terminologiilor specifice diverselor domenii de specialitate, iar procesul de globalizare a generat, pe lângă schimbul de produse și tehnologii, și schimbul de documentații. În acest context, aplicațiile de gestiune terminologică au o importanță crucială.

Până în ultimul deceniu, aplicațiile NLP și cele terminologice au urmat un curs paralel, fiecare reprezentându-și structurile de date de așa manieră încât să răspundă cerințelor formale și condițiilor tehnice existente în momentul respectiv. Informațiile din lexicoanele aplicațiilor NLP vizează în primul rând atributele gramaticale (lexicale, morfologice, sintactice și semantice) și, foarte puțin, atribute specifice termenilor. Aplicațiile de gestiune terminologică (marea lor majoritate abordând orientarea conceptuală în terminologie) uzează foarte puțin de informația gramaticală, prezentând pe larg, în schimb, atribute ale termenilor: domeniul, sistemul conceptual, conceptul desemnat, definiție, contexte, informații de utilizare, statut, autoritate normativă etc.

Evoluția tehnică și științifică a dus la apariția unor sisteme informatice din ce în ce mai complexe. S-a simțit nevoia ca, pe de o parte, aplicațiile NLP să poată exploata termenii (care necesită o tratare specială, fiind, în general, expresii multicuvânt care funcționează ca unități de sine stătătoare), iar pe de altă parte, aplicațiile de gestiune terminologică să poată utiliza facilitățile oferite de aplicațiile

NLP (unele dintre cele mai utile fiind recunoașterea terminologică avansată și extragerea termenilor). Mai mult decât atât, în prezent se caută metode de a facilita interschimbul de date între diverse produse informatice. S-a ajuns astfel să se caute elaborarea unor modele de reprezentare a informațiilor lexicale care să permită exploatarea acestora de către cele mai variate aplicații, atât de natură terminologică, cât și din clasa NLP. Există mai multe modele de organizare a informațiilor care încearcă să răspundă acestor exigențe, unele dintre ele (de exemplu LMF¹) în curs de standardizare ISO. Fiecare are avantaje și dezavantaje, iar cercetările în domeniu sunt foarte dinamice.

Proiectul **SIASTRO**², care are ca temă realizarea unui sistem informatic pentru analiza sintagmatică a textelor în limba română, își propune, ca o primă aplicație practică, implementarea unui extractor de termeni și înregistrarea acestora într-o colecție de date terminologice. Prin urmare, el se află la punctul de întâlnire a NLP cu sistemele de gestiune terminologică. S-a căutat să se abordeze un mod de structurare și reprezentare a datelor lingvistice și terminologice în conformitate cu modelele formale existente. Acest lucru permite extinderea și, eventual, integrarea aplicațiilor noastre în cele dezvoltate în acest domeniu.

Am început, așadar, demersul nostru prin analiza modelelor existente, căutând să alegem un mod de reprezentare care să fie cât mai apropiat unui model standard, să răspundă cerințelor analizei sintagmatice pentru limba română și, totodată, să adapteze și să extindă atributele *Dicționarului Morfologic Român*³ realizat de colectivul RoLingva⁴.

1.1. Standarde și modele pentru resursele lexicale

Există numeroase încercări de a standardiza procesele și resursele lingvistice. Amintim aici câteva dintre cele mai importante proiecte care au avut drept scop crearea de modele pentru reprezentarea resurselor lingvistice:

- GENELEX (1990-1994) – a fost un proiect EUREKA, unul dintre scopurile sale fiind acela de a proiecta un model global de reprezentare a oricărui tip de informație lexicală, într-un mod neutru, independent de aplicație și neatașat unei teorii lingvistice specifice.
- EAGLES⁵ (Expert Advisory Group on Language Engineering Standards) – o inițiativă a Comisiei Europene în cadrul programului DG XIII *Linguistic*

¹ Lexical Mark-up framework: <http://tagmatica.fr./doc/ISO24613cdRev9.pdf>

² Proiectul *Sistem informatic pentru analiza sintagmatică a textelor în limba română. Fundamentare teoretică și implementare – SIASTRO* se desfășoară în cadrul programului 86 CEE-X-II 03 / 31.07.2006. Realizatorii sunt: Universitatea „Babeș-Bolyai” din Cluj-Napoca (Facultatea de Litere), coordonator, Software ITC SA (prin colectivul *RoLingva*), Institutul de Lingvistică și Istorie Literară „Sextil Pușcariu” din Cluj-Napoca și Universitatea Tehnică din Cluj-Napoca (Catedra de Calculatoare). Vezi *supra*, Tămâianu-Morita 2008.

³ http://www.rolingva.ro/aplicatii_dictionar.php

⁴ <http://www.rolingva.ro/>

⁵ <http://www.ilc.cnr.it/EAGLES/home.html>

Research and Engineering, a avut ca scop accelerarea producerii de standarde pentru:

- resurse lingvistice la scară foarte largă;
 - mijloacele de tratare a unor astfel de resurse prin formalisme ale lingvisticii computaționale și aplicații informatice diverse;
 - mijloace de atestare și evaluare a unor astfel de resurse, aplicații și produse.
- ISLE⁶ (International Standards for Language Engineering) – este atât numele unui proiect, cât și numele unui întreg set de activități coordonate din domeniul tehnologiei limbajului uman (HLT – Human Language Technology). ISLE a funcționat sub egida EAGLES; scopul său a fost să dezvolte standarde într-un cadru internațional, în contextul inițiativei EU-US International Research Cooperative.
 - MULTEX⁷ (Multilingual Text Tools and Corpora) – înglobează mai multe proiecte care vizează elaborarea de standarde pentru codificarea textelor și a corpusurilor și dezvoltarea aplicațiilor și a resurselor lingvistice; proiectul a stabilit linii directoare pentru NLP și traducerea automată (MT – Machine Translation) bazată pe corpusuri.
 - PAROLE (Preparatory Action for Linguistic Resources Organisation for Language Engineering) – proiect al Uniunii Europene, a avut ca scop armonizarea resurselor lexicale și a corpusurilor pentru limbile Uniunii Europene.
 - SIMPLE⁸ – proiect sponsorizat de Programul Cadru IV, a reprezentat prima încercare de a dezvolta lexicoane semantice pentru un număr mare de limbi, cu un model comun care codifică „tipuri semantice” structurate și cadre semantice; dezvoltat în strânsă legătură cu PAROLE.

1.2. Standarde și modele pentru resursele terminologice

Marea diversitate a reprezentării informațiilor terminologice a dus la apariția unor standarde internaționale care stabilesc principii, metode de lucru și modele de reprezentare pentru a se putea realiza diseminarea și interschimbul datelor terminologice. În cadrul ISO (*International Organisation for Standardisation*) a fost creat comitetul tehnic 37 (TC37 – *Terminology and other Language Resources*) care a elaborat numeroase standarde, dintre care amintim:

- ISO 704:2000 *Terminology work – Principles and methods*;
- ISO 12620:1999 *Computer applications in terminology – Data categories*;
- ISO 1087-1:2000 *Terminology work – Vocabulary – Part 1: Theory and application*;

⁶ http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm

⁷ <http://aune.lpl.univ-aix.fr/projects/multext/>

⁸ <http://www.ub.es/gilcub/SIMPLE/simple.html>

- ISO 1087-2:2000 *Terminology work – Vocabulary – Part 2: Computer applications*;
- ISO 12616:2002 *Translation-oriented terminography*;
- ISO 12200:1999 *Computer applications in terminology – Machine-readable terminology interchange format (MARTIF) – Negotiated interchange*;
- ISO/FDIS 24613 *Language resource management–Lexical markup framework(LMF)*.
www.tc37sc4.org/new_doc/ISO_TC37_N130_rev9_LMF_15March2006.pdf

1.3 Modele pentru integrarea resurselor lexicale și a celor terminologice

Există, de asemenea, o mulțime de inițiative a diverselor organizații și consorții, care își propun să atingă același scop: reprezentarea uniformă a informațiilor lexicale în vederea realizării interschimbului între diverse aplicații și a reutilizării resurselor, precum și integrarea informațiilor terminologice și a celor lexicale în aceeași gamă de aplicații.

TEI⁹ (Text Encoding Initiative), lansat inițial în 1987, este un standard internațional și interdisciplinar de reprezentare a informațiilor textuale. Din 2000 s-a înființat un consorțiu care-i sprijină activitatea. Standardul TEI prevede modele de reprezentare atât a informațiilor lexicale, cât și a celor terminologice.

Consortiul **SALT**¹⁰ (Standards-based Access service to multilingual Lexicons and Terminologies), format din grupuri academice, guvernamentale, comerciale și asociații din Europa și SUA a fost creat tocmai pentru aceasta și a elaborat formatul **XLT**.

În același scop se desfășoară activitatea grupului **OSCAR**¹¹ (Open Standards for Container/Content Allowing Re-use) din cadrul asociației **LISA** (Localisation Industry Standards Association).

OLIF¹² este un consorțiu sub egida căruia s-a elaborat un standard deschis cu același nume (OLIF – Open Lexicon Interchange Format), pentru interschimbul de informații lexicale și terminologice.

2. Modelele analizate în cadrul proiectului SIASTRO

Dintre modelele de reprezentare a informației lexicale ne-am oprit la studierea **LMF** (Lexical Markup Framework) și **OLIF** (Open Lexicon Interchange Format), care se apropie cel mai mult atât de cerințele formale ale gramaticii limbii române, cât și de cele impuse de realizarea tehnică a obiectivului propus.

⁹ <http://www.tei-c.org/index.xml>

¹⁰ <http://www.ttf.org/salt/index.html>

¹¹ <http://www.lisa.org/sigs/oscar/>

¹² <http://www.olif.net/>

2.1. LMF – Lexical Markup Framework

LMF – Lexical Markup Framework este o propunere de standard ISO 24613:2006. În momentul de față este un document de lucru. Lexical Markup Framework este un metamodel abstract care oferă un cadru comun, standardizat, pentru construcția lexicoanelor computerizate. LMF asigură codificarea informației lingvistice într-un mod care permite reutilizarea ei în diverse aplicații. El oferă o reprezentare comună, care poate fi partajată între aplicații, a obiectelor lexicale, incluzând aspectele morfologice, sintactice și semantice. LMF răspunde unor cerințe de normalizare a lexicoanelor utilizate în aplicațiile NLP (Francopoulo *et alii* 2006, p. 27) astfel:

1. Reprezintă cuvinte în limbi în care sunt posibile mai multe ortografii (nativă sau transliterată).
2. Reprezintă morfologia unor limbi pentru care reprezentarea tuturor formelor flexionale nu poate fi gestionată; pentru astfel de limbi, doar reprezentarea în intensiune este singura de luat în considerare.
3. Asociază ușor formele scrise cu formele rostite pentru orice limbă.
4. Reprezintă cuvinte compuse complexe.
5. Reprezintă expresii multi-cuvânt fixe, semi-fixe sau flexibile.
6. Reprezintă comportări sintactice specifice.
7. Permite corespondența complexă a argumentelor între descrierile sintactice și semantice.
8. Permite organizarea semantică bazată pe SynSet-uri (ca în WordNet) sau pe predicate semantice (ca în FrameNet).
9. Reprezintă resurse lingvistice multilingve de dimensiuni mari pe baza pivoților inter-limbi sau pe baza legăturilor de transfer.

LMF utilizează Unicode ca sistem de codificare a caracterelor. Constantele lingvistice sunt specificate în Data Category Registry (DCR), așa cum sunt definite de ISO TC37. Specificația LMF este conformă principiilor de modelare ale **UML** (Unified Modeling Language)¹³.

LMF este compus din următoarele componente:

- Un model central (**core package**), care cuprinde un metamodel, adică scheletul structural al LMF; acesta descrie ierarhia informațiilor incluse într-o intrare lexicală. Modelul central este completat cu resurse diverse care fac parte din definiția LMF. Aceste resurse cuprind:
 - categorii de date specifice, utilizate de varietatea tipurilor de resurse asociate cu LMF; aceste categorii de date sunt relevante pentru metamodel și sunt asociate cu extensii ale modelului central;

¹³ UML a fost elaborat în 1997 de *Object Management Group* (OMG) cu scopul de a oferi un limbaj de proiectare comun și stabil, care să servească la dezvoltarea aplicațiilor informatice. UML a devenit un limbaj de modelare standard.

- restricțiile care guvernează relațiile categoriilor de date cu metamodelul și cu extensiile sale;
- proceduri standard pentru exprimarea categoriilor de date în XML.
- Extensiile modelului central:
 - extensii pentru lexicoane electronice [en: machine readable lexicons];
 - extensii pentru lexicoanele NLP.

2.1.1. Metamodelul LMF

Modelul LMF este format din clase UML, asocieri între clase, precum și un set de categorii de date conforme cu ISO 12620, reprezentate ca perechi atribut-valoare.

Cei care creează un lexicon trebuie să utilizeze clasele specificate în pachetul central LMF (**LMF core package**). Se pot utiliza clasele suplimentare definite în extensiile standardului (**LMF extensions**). De asemenea, cei care creează un lexicon trebuie să definească selecția categoriilor de date.

2.1.1.1. Pachetul central LMF

Pachetul central LMF este un metamodel care oferă un mijloc flexibil de a construi modele LMF și extensii. Reprezentarea UML a pachetului central este următoarea¹⁴:

Clasa *DataBase* este o clasă unică și reprezintă resursa lingvistică în întregime. *DataBase* este container pentru unul sau mai multe lexicoane.

Clasa *Lexicon* este container pentru toate intrările lexicale ale unei limbi în baza de date. Un *Lexicon* trebuie să conțină cel puțin o intrare lexicală (*Lexical Entry*).

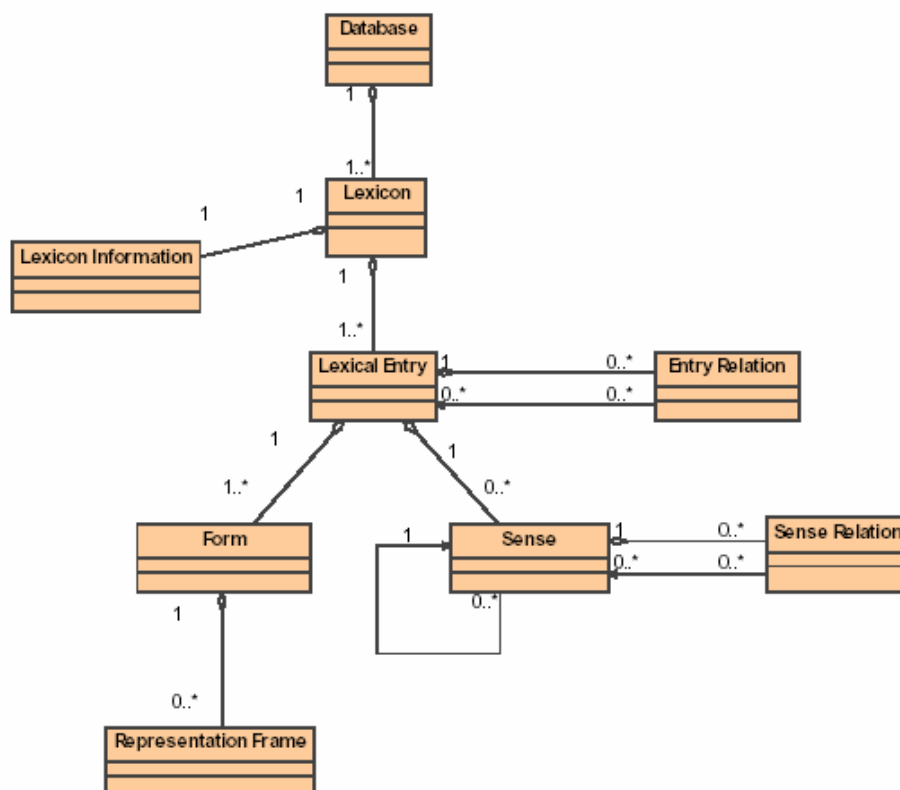
Clasa *Lexical Information* conține informații administrative și alte atribute generale. Această clasă descrie informațiile generale referitoare la un lexicon.

Clasa *Lexical Entry* este un container pentru reprezentarea componentelor de nivel superior ale limbii. În consecință, numărul de reprezentări pentru cuvintele simple, expresiile multi-cuvânt și afixe din lexicon este egal cu numărul intrărilor lexicale din lexiconul respectiv. *Lexical Entry* este un container pentru tratarea claselor *Form* și *Sense*. Prin urmare, *Lexical Entry* tratează relațiile dintre forme și semnificațiile asociate lor. O intrare lexicală – *Lexical Entry* – poate avea una sau mai multe forme și poate avea zero sau mai multe sensuri.

Clasa *Entry Relation* este o clasă de referințe încrucișate care poate lega două sau mai multe intrări lexicale din cadrul aceluiași lexicon sau din lexicoane diferite. Atributele conținute în această clasă descriu tipul relațiilor.

Clasa *Form* reprezintă o variantă lexicală a formei scrise sau vorbite a intrării lexicale. O formă conține un șir Unicode care reprezintă forma cuvântului și categoriile de date care descriu atributele formei cuvântului. O clasă *Form* poate conține mai multe variante ortografice (lema, pronunția, despărțirea în silabe). Clasa *Form* are două subclase: *Lemmatized Form* (care poate conține numai leme) și *Inflected Form* (care poate conține numai forme flexionate).

¹⁴ <http://tagmatica.fr/doc/ISO24613cdRev9.pdf>



Dacă există mai multe forme ortografice de reprezentare a cuvântului (transliterații, romanizări, pronunții), clasa *Form* poate fi asociată cu clasa *Representation Frame*. Aceasta conține o reprezentare ortografică specifică și una sau mai multe categorii de date care descriu atributele acelei reprezentări.

Clasa *Sense* conține atributele care descriu semnificația intrării lexicale. Ea permite reprezentarea ierarhică a sensurilor: o parte a sensului poate fi în relație cu o altă parte a aceluiași sens.

Clasa *Sense Relation* este o clasă de referințe încrucișate care poate lega două sau mai multe sensuri pentru o singură limbă sau între mai multe lexicoane. Ea poate conține atribute care descriu tipul relațiilor semantice.

2.1.1.2. Extensiile LMF

Toate extensiile se conformează pachetului central LMF. O extensie nu poate fi utilizată pentru a reprezenta datele lexicale independente de pachetul central. În funcție de tipul datelor lingvistice, o extensie poate depinde de o altă extensie. Din punctul de vedere al UML, o extensie este un pachet UML.

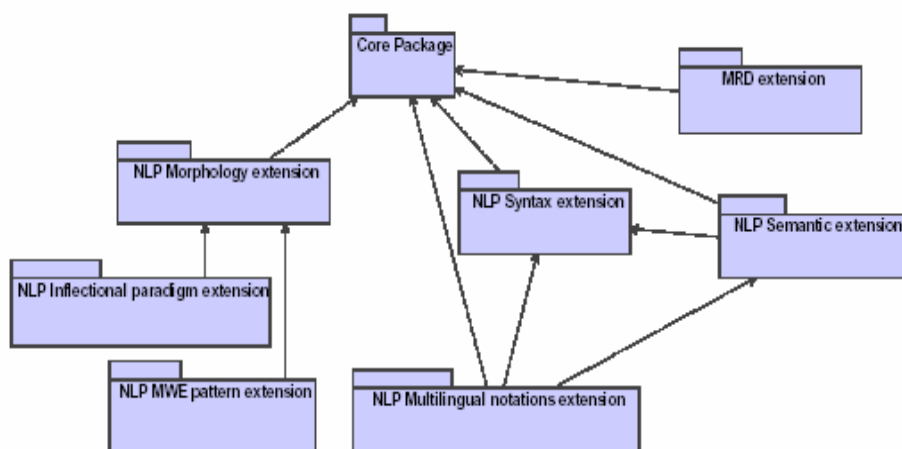
Mecanismul realizării unei extensii include:

- crearea subclaselor pe baza principiilor de modelare UML;
- adăugarea de noi clase;
- restricții asupra cardinalității și tipului asocierilor;
- permiterea punctelor diferite de ancorare pentru asocieri;
- selecția categoriilor de date.

Extensiile NLP ale modelului LMF cuprind:

- extensia NLP pentru morfologie;
- extensia NLP pentru paradigma de flexiune;
- extensia NLP pentru modelul multi-cuvânt (MWE – MultiWord Expression);
- extensia NLP pentru sintaxă.

Dependențele între diferitele extensii sunt reprezentate în diagrama următoare¹⁵:



2.1.2. LMF pentru datele terminologice

Deși LMF a fost conceput pentru reprezentarea informațiilor lexicale, există posibilitatea ca el să poată fi folosit și pentru reprezentarea informațiilor terminologice, realizându-se astfel dezideratul de a avea o reprezentare uniformă a celor două tipuri de informații.

Deși în G. Francopoulo *et alii* (2006a) se folosește expresia *LMF pentru lexicoane specializate*, caracteristicile prezentate sunt cele specifice colecțiilor de date terminologice:

1. un număr mare de expresii multi-cuvânt (MWE – Multiword Expression)
 - termenii sunt, în general, compuși din mai multe cuvinte;
2. un număr mare de variante ortografice, care includ abrevierile și acronimele – standardele terminologice prezintă posibilele forme prescurtate ale termenilor: forma scurtă, forma troncată, sigla, acronimul, abrevierea;

¹⁵ <http://tagmatica.fr/doc/ISO24613cdRev9.pdf>

3. includerea informațiilor specifice domeniului: definițiile terminologice, coduri specifice – standardele terminologice prevăd informații de descriere noțională – definiții, contexte de utilizare etc.;
 4. specificarea domeniului de cunoștințe – informație esențială în terminologii.
- Soluțiile pe care le oferă LMF pentru tratarea caracteristicilor terminologice sunt:
- a) expresiile multi-cuvânt pot fi folosite, în particular, pentru a reprezenta termeni;
 - b) diferitele forme prescurtate ale termenilor pot fi reprezentate ca variante ortografice sau cu ajutorul relației de sinonimie din LMF;
 - c) fiecare element LMF poate fi îmbogățit cu perechi atribut-valoare care să reprezinte informațiile de domeniu/subdomeniu, definiție etc.
- ### 2.2. Modelul OLIF – Open Lexicon Interchange Format

OLIF este un standard deschis pentru codificarea informațiilor lexicale și terminologice. Conceput în anii '90 ca un mijloc de a asigura interschimbul de informații între aplicațiile NLP (în special traducerea automată) și bazele de date terminologice, OLIF a evoluat într-un standard care oferă mijloace diverse pentru facilitarea reprezentării și interschimburilor datelor lingvistice.

Prototipul OLIF a fost dezvoltat în cadrul proiectului OTELO (Open Translation Environment for Localization) al Uniunii Europene (încheiat în 2000). În martie 2000 s-a constituit consorțiul OLIF¹⁶, al cărui scop este de a elabora un standard bazat pe prototipul OLIF.

Versiunea a doua a OLIF a fost elaborată în cadrul inițiativei SALT (Standards-based Access service to multilingual Lexicons and Terminologies); SALT este un consorțiu de instituții academice, guvernamentale, comerciale din Europa și Statele Unite, care are ca scop elaborarea de formate și instrumente care să faciliteze interschimbul de informații între bazele de date terminologice și lexicoanele utilizate de aplicațiile NLP.

Implementat ca XML Schema (OLIF v.2.1) și ca XML DTD (OLIF v.2), prototipul OLIF oferă un set reprezentativ de trăsături lexicale și terminologice utile în aplicații ca:

- interschimbul de informații lexicale/terminologice;
- managementul lexicoanelor și terminologiilor;
- extragerea termenilor [engl.: *term extraction*];
- limbaj controlat;
- regăsirea informației [engl.: *information retrieval*];
- dezvoltarea glosarelor;
- crearea de ontologii.

¹⁶ Consorțiul **OLIF** este format din SAP, Microsoft, LionBridge X, Trados, Systran, IBM, DFK, BASIS Technology, SLS Smart Logic Solution.

2.2.1. Structura generală a OLIF

Datele sunt organizate în trei grupe esențiale de informații:

1. Antetul (*header*) conține informații comune tuturor intrărilor lexicale/terminologice;
2. Corpul (*body*) conține intrările lexicale/terminologice propriu-zise;
3. Resursele partajate (*shared resources*) conțin informații suplimentare referite de intrările lexicale/terminologice (de exemplu, sursele bibliografice).

2.2.2. Structura corpului (*body*) unui fișier OLIF

Corpul fișierului OLIF este o listă de intrări care conțin date grupate conform caracterului lingvistic/lexical/terminologic al informațiilor reprezentate. Deoarece scopul principal al OLIF este acela de a oferi o punte de legătură între lexicoanele utilizate în aplicațiile de procesare a limbajului natural (NLP) și aplicațiile de management al terminologiilor, el a fost proiectat având în vedere atât punctul de vedere lexical, cât și cel terminologic. Structura unei intrări (*entry*) OLIF reflectă, prin urmare, acest caracter hibrid, fără a fi orientată pe leamnă, ca majoritatea lexicoanelor, și fără a fi orientată pe concept, ca majoritatea bazelor de date terminologice. Pentru a reprezenta atât informațiile necesare lexicoanelor, cât și cele din terminologii, în OLIF se optează pentru o structură flexibilă, orientată pe semnificația cuvântului [eng.: *word-sense orientation*]. O intrare este o *colecție de informații monolingve* asupra unei *semnificații specifice* a unui cuvânt sau a unei expresii; între intrări pot fi *referințe încrucișate*, care reflectă diferite *relații* în cadrul aceleiași limbi; echivalențele între intrările dintr-o limbă oarecare și intrările corespunzătoare în altă limbă sunt date sub formă de *relații de transfer*.

O intrare OLIF are următoarea componentă de informații:

- un grup unic de informații, a cărui prezență este *obligatorie*, care conține datele corespunzătoare unei singure limbi, numit **monolingual (mono)**;
- o mulțime *opțională* de grupuri de informații – **cross-reference (cross-Refer)** – care reprezintă relațiile intrării respective cu alte intrări corespunzătoare aceleiași limbi; fiecare grup **cross-reference** reprezintă o singură relație; pot exista mai multe grupuri cross-reference ale aceleiași intrări; în aceste grupuri sunt reprezentate relațiile de sinonimie, omonimie, antonimie etc.;
- o mulțime *opțională* de grupuri de informații – **transfer** – care reprezintă relațiile dintre intrarea respectivă și intrările corespunzătoare din alte limbi; fiecare grup **transfer** reprezintă o singură relație de transfer, unidirecțională; în aceste grupuri sunt specificate legăturile spre echivalențele în alte limbi corespunzătoare intrării respective;
- un grup *opțional* de categorii de date generale – **general (generalDC)** care poate fi inclus în oricare din grupurile **monolingual**, **cross-reference** sau **transfer**.

În cele ce urmează, prezentăm acele aspecte ale OLIF care ne interesează în demersul nostru de inventariere a trăsăturilor lexico-morfologice și sintactice care să fie reprezentate în lexicon pentru realizarea analizei sintagmatiche. Prin urmare, nu vom inventaria trăsăturile specifice reprezentării conceptelor și termenilor. De asemenea, vom extrage din standardul OLIF doar partea care interesează reprezentarea monolingvă, ignorând aspectele care vizează transferul între mai multe limbi.

2.2.3. Categoriile esențiale de date

Categoriile esențiale de date [engl.: *key data categories*] – **keyDC** – specifică semnificația unei intrări date într-o limbă anume și sunt obligatorii în grupul **monolingual** al fiecărei intrări.

Există cinci categorii esențiale de date:

- *forma canonică* [engl. *canonical form*] – **canForm** – șirul de caractere corespunzător intrării, reprezentat în formă canonică, conform specificațiilor OLIF; este, în esență, forma lematizată a intrării;
- *limba* [engl. *language*] – **language** – este specificarea limbii corespunzătoare intrării;
- *partea de vorbire* [engl. *part of speech*] – **ptOfSpeech** – specifică partea de vorbire a intrării;
- *domeniul* [engl. *subject field*] – **subjField** – specifică domeniul de cunoștințe corespunzător intrării;
- *semnificația* [engl. *semantic reading*] – **semReading** – identificatorul clasei semantice, utilizat pentru a distinge semnificații diferite ale intrărilor care au valori identice pentru *forma canonică*, *limbă*, *parte de vorbire* și *domeniu*.

entry	mono	keyDC	<table style="border-collapse: collapse;"> <tr> <td style="padding: 2px 5px;">canForm</td> <td style="padding: 2px 5px;"><i>tabel</i></td> </tr> <tr> <td style="padding: 2px 5px;">language</td> <td style="padding: 2px 5px;"><i>ro</i></td> </tr> <tr> <td style="padding: 2px 5px;">ptOfSpeech</td> <td style="padding: 2px 5px;"><i>noun</i></td> </tr> <tr> <td style="padding: 2px 5px;">subjField</td> <td style="padding: 2px 5px;"><i>general</i></td> </tr> <tr> <td style="padding: 2px 5px;">semReading</td> <td style="padding: 2px 5px;">32</td> </tr> </table>	canForm	<i>tabel</i>	language	<i>ro</i>	ptOfSpeech	<i>noun</i>	subjField	<i>general</i>	semReading	32
canForm	<i>tabel</i>												
language	<i>ro</i>												
ptOfSpeech	<i>noun</i>												
subjField	<i>general</i>												
semReading	32												

Exemplu¹⁷ de categorii esențiale de date, dat sub formă de structuri de trăsături

Versiunea 2.0/2.1 a modelului a fost extinsă cu scopul de a permite atât reprezentarea datelor orientată pe concept (ca în majoritatea colecțiilor de date terminologice), cât și reprezentarea orientată pe leme (ca în colecțiile de date lexicografice). Identificatorii de leme și de concept sunt prezenți la nivelul superior

¹⁷ Exemplele sunt adaptate după Susan McCormick, OLIF Consortium *The Structure and Content of the Body of an OLIF v.2.0/2.1 File*

al intrării (*entry*). Identificatorul de concept permite organizarea intrărilor ca semnificații echivalente asociate aceluiași concept, iar identificatorul de leamnă permite organizarea intrărilor cu semnificații diferite, care au aceeași leamnă:

$$\left[\text{entry} \left[\begin{array}{l} \left\{ \begin{array}{l} \text{conceptID} \\ \text{lemaID} \end{array} \right\} \\ \text{mono} \end{array} \right] \right]$$

Informațiile relative la un singur cuvânt sau termen (care sunt, prin urmare, asociate unei singure limbi), sunt reprezentate în intrarea OLIF în grupul **mono** (**monolingual**); acest grup cuprinde datele esențiale (**keyDC**), obligatorii, precum și alte categorii de date (**monoDC**), grupate, la rândul lor, în funcție de natura informațiilor – lingvistică, lexicală sau terminologică – în date administrative (**monoAdmin**), morfologice (**monoMorph**), sintactice (**monoSyn**) și semantice (**monoSem**). Datele esențiale **keyDC** identifică în mod univoc intrarea.

Pentru identificarea intrărilor și pentru simplificarea reprezentării referințelor încrucișate și a transferurilor, modelul OLIF prevede mai multe posibilități de a specifica identificatorii numerici (care pot fi referiți în cadrul relațiilor), asociați fie grupului **mono** (**monoUserId** sau **monoUniversalId**), fie categoriilor esențiale de date (**keyDCUserId** sau **keyDCUniversalId**), fie intrării (**lemmaUserId**).

Categoriile de date

Categoriile de date din OLIF referitoare la informațiile lexicale sunt ilustrate în figura următoare; prezentarea este sub forma structurilor de trăsături (categoriile de date scrise cu aldine sunt obligatorii), astfel: [denumire *valoare* (explicație)].

Structura intrării are forma:

$$\left[\text{entry} \left[\begin{array}{l} \text{lemmaUserId } nb \\ \text{mono} \left[\begin{array}{l} \text{monoUserId } nb \\ \text{monoUniversalId } nb \\ \text{keyDC } [...] \\ \text{monoDC } [...] \\ \text{generalDC } [...] \end{array} \right] \end{array} \right] \right]$$

(identificator al utilizatorului pentru leamnă)
(identificator al utilizatorului)
(identificator universal)
(date esențiale)
(alte categorii de date)
(informații generale)
(grupează datele monolingve ale unei intrări) ()

unde grupul monolingual – **mono** este:

mono	monoUserId	<i>nb</i>	(identificator al utilizatorului)
	monoUniversalId	<i>nb</i>	(identificator universal)
	keyDC	[...]	(date esențiale)
	monoDC	[...]	(alte categorii de date)
	generalDC	[...]	(informații generale)

Categoriile esențiale de date, *keyDC* au următoarea structură:

keyDC	keyDCUserId	<i>nb</i>	(identif. al utilizatorului)	
	keyDCUniversalId	<i>nb</i>	(identif. universal)	
	canForm	<i>string</i>	(forma canonică)	
	language	<i>string</i>	(limba)	
	ptOfSpeech	noun	(substantiv)	(partea de vorbire)
		verb	(verb)	
		adj	(adjectiv)	
		adv	(adverb)	
		prep	(prepoziție)	
		conj	(conjunție)	
		det	(determinant)	
		part	(particulă verbală)	
		auxverb	(verb auxiliar)	
pron		(pronume)		
punc	(semn de punctuație)			
other	(alta)			
subjField	<i>string</i>	(domeniu)		
semReading	<i>string</i>	(definiție)		

Categoriile de date asociate grupului *mono* (*monoDC*) sunt date suplimentare, care pot fi: informații administrative – *monoAdmin*, informații morfologice – *monoMorph*, informații sintactice și semantice – *monoSyn* și, respectiv, *monoSem*.

Astfel, informațiile administrative relative la datele lexicografice conțin date despre despărțirea în silabe – *syllabification*, despre structura șirului care formează intrarea – *entryFormation* (abreviere, acronim, cuvânt simplu, cuvânt compus, expresie, nespecificat), despre tipul expresiei – *phraseType* (expresie multi-cuvânt, expresie fixă, lexicalizată, colocație, idiom, nespecificată), autorul intrării – *originator* sau statutul intrării – *adminStatus* (intrare nouă, intrare verificată, implicită, exclusiv pentru MT, învechită, nespecificat).

Informații morfologice – *monoMorph* sunt cele care se referă la transcrierea structurii morfologice – *morphStruct*, la clasa de flexiune – *inflection*, cele care indică centrul sintagmei – *head*, genul gramatical – *gender*, cazul – *case*, numărul – *number*, persoana – *person*, timpul – *tense*, modul – *mood*, aspect – *aspect*, gradele de comparație – *degree* sau tip de verb auxiliar – *auxType* (în funcție de limbă).

Informații sintactice – *monoSyn* sunt cele care descriu comportarea generală a șirului de intrare – *synType*, poziționarea nemarcată a șirului de intrare, din punct de vedere sintactic – *synPosition*, descriu tranzitivitatea verbelor – *transType*, structura de constituenți a unui șir multi-cuvânt – *synStruct*, valența șirului de intrare – *synFrame* etc.

Informații semantice – *monoSem* se referă la definiția – *definition*, genul – *natGender* (masculin, feminin, nespecificat) și la o categorie care reprezintă tipul unei intrări care respectă o anumită clasificare semantică – *semType*.

Grupul de date generale – *generalDC* are forma:

generalDC	updater	<i>string</i>	(persoana care a operat modif.)	(categorii de date generale)
	modDate	<i>date</i>	(data modif.)	
	example	<i>string</i>	(exemplu)	
	usage	<i>string</i>	(utilizare)	
	note	<i>string</i>	(comentariu al lexicografului/ terminologului)	

3. Informațiile lexico-morfologice și sintactice utilizate în proiectul SIASTRO

Proiectul **SIASTRO** continuă cercetările grupului *RoLingva* de la Software ITC SA asupra morfologiei. Acesta a realizat *Dicționarul Morfologic Român*¹⁸,

¹⁸ http://www.rolingva.ro/aplicatii_dictionar.php

dicționar în format electronic, care acoperă lexicul *DEX* și care conține, printre alte componente, un analizor morfologic.

În ce privește modelele formale de reprezentare a informației lexicale s-a considerat că cel mai apropiat de structura existentă a dicționarului morfologic și de cerințele impuse de analiza sintagmatică a limbii române este OLIF.

Intrările *Dicționarul Morfologic Român* au asociate atribute lexico-morfologice și fonologice. O parte dintre aceste atribute pot servi scopului propus; altele (de exemplu, cele fonologice) nu sunt necesare analizei sintagmatice; altele trebuie rafinate. Pentru realizarea analizei sintagmatice, fiecare clasă lexicogramaticală trebuie îmbogățită cu atribute sintactico-semantice care nu sunt prezente în *Dicționarul morfologic*.

În vederea stabilirii mărcilor specifice claselor și subclaselor lexico-morfologice, care să facă posibilă realizarea automată a analizei sintagmatice, s-a procedat la examinarea atributelor relevante ale fiecărei clase și a valorilor acestora.

Pentru selectarea/stabilirea atributelor și valorilor ce privesc categoriile morfologice s-a utilizat în principal GALR 2005. S-a apelat, de asemenea, și la materialul științific oferit de D. D. Drașoveanu, G. G. Neamțu și D. Bejan, ca și la cadrul în sens larg și fundamental teoretic oferit de lingvistica integrală a lui E. Coșeriu. Există cazuri în care s-a considerat că opțiunile terminologice și/sau teoretice ale acestora sunt mai adecvate obiectivelor proiectului decât cele din GALR 2005, iar acestea sunt menționate explicit în studiul dedicat categoriilor morfologice în cauză.

Deoarece ca primă aplicație a analizorului sintagmatic este prevăzută realizarea unui extractor de termeni, ne-am orientat spre studierea acelor formate de reprezentare a datelor lexico-morfologice și sintactice care să permită și extinderea spre bazele de date terminologice. În ce privește terminologia, ca teorie, am adoptat concepția noțională (orientată pe concept), care stă la baza modelelor formale analizate.

Premisele elaborării tehnologiei lingvistice pentru limba română au fost create de studiile teoretice, realizate de grupul *RoLingva* asupra morfologiei. A fost realizat primul model formalizat al morfologiei cu reguli și metode care pot acoperi în totalitate sistemul flexionar din limba română. În afară de informațiile referitoare la elementele morfologice și ortografice, modelul păstrează informații incipiente despre elemente fonetice, cum ar fi silabația fonetică, accent, împreună cu regulile de migrare a acestuia în cadrul flexiunii. Modelul morfologic respectă regulile gramaticale clasice și a fost realizat ca un sistem deschis pentru a putea acoperi cu ușurință și alte domenii ale gramaticii limbii române. El permite extinderile necesare pentru a prelua informațiile care vor rezulta din formalizarea sintaxei limbii române.

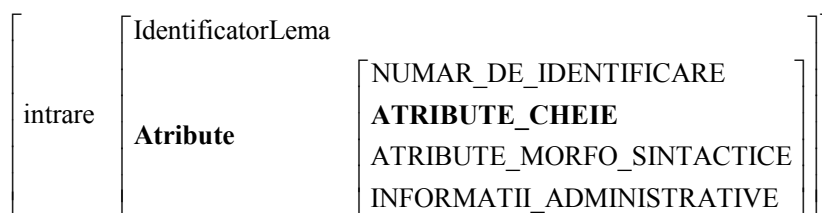
Limba română este o limbă care flexionează puternic și permite realizarea de numeroase cuvinte noi prin sufixare și prefixare, cuvinte care, la rândul lor, pot avea propria flexiune. Numărul formelor flexionate este mare și, practic, se poate pierde controlul asupra acestora, dacă se va face un simplu inventar al tuturor formelor însoțite de categoria gramaticală reprezentativă. Din acest motiv, în

tezaurul de cunoștințe lingvistice s-a preferat obținerea acestor forme prin metode procedurale, pe bază de reguli, plecând de la o *rădăcină* condensată, liste de *sufixe* și *prefixe*, și o *clasă flexionară părinte*, care pot descrie în totalitate o intrare lexicală. Combinațiile care se pot realiza astfel sunt foarte numeroase și pot acoperi un lexic foarte bogat.

Pentru proiectul **SIASTRO** s-a făcut o documentare asupra atributelor existente în tezaurul de cunoștințe lingvistice, ținându-se cont de importanța lor în lexiconul care se va crea. De asemenea, s-a făcut o paralelă între atributele utilizate de sistemele standardizate și atributele care pot fi preluate din tezaurul de cunoștințe lingvistice.

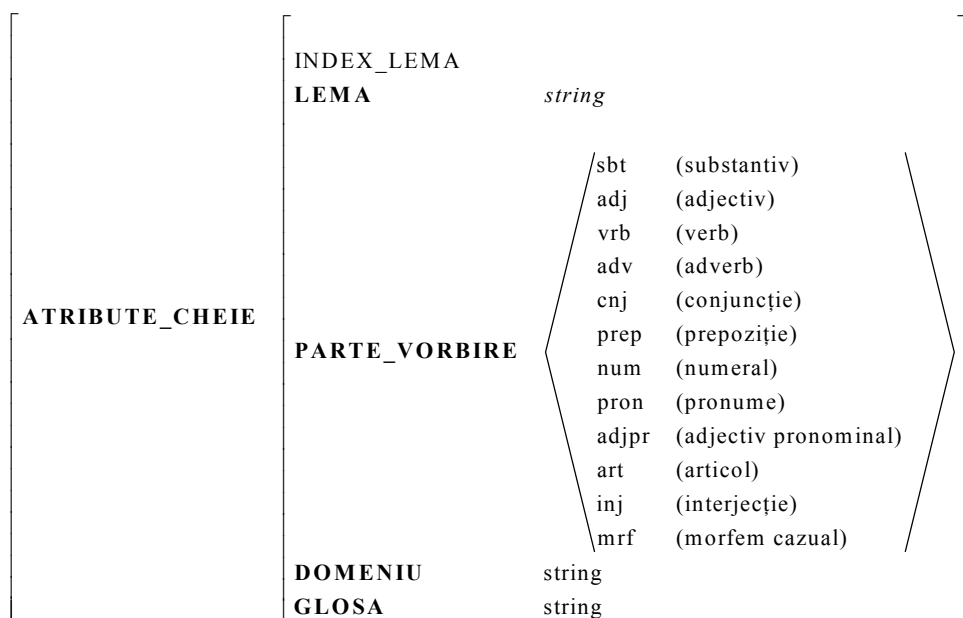
S-a creat, în felul acesta, un model de reprezentare a atributelor morfo-sintactice și s-au dezvoltat procedurile care generează lexiconul îmbogățit cu aceste atribute.

Astfel, pornind de la modelul OLIF, s-a creat o **structură de intrare** de forma:



Structura intrării SIASTRO

în care **atributele cheie** sunt:



Atribute-cheie SIASTRO

Pentru fiecare clasă lexico-gramaticală s-au stabilit atribute și valori specifice. Aceste atribute se împart, la rândul lor, în mai multe categorii:

- atribute lexico-morfologice propriu-zise (de exemplu, TIP, TEMA pentru clasele lexico-gramaticale flexionale, GEN, pentru substantive etc.);
- atribute sintactice: VALENȚA, POZIȚIE (pe care o poate avea adjectivul în raport cu substantivul determinat), DETERM (reprezintă comportamentul determinării adjectivului antepus), PROXIMITATE_REGENT (se aplică la adjectivele pronominale care pot fi postpuse și reprezintă proximitatea față de regent) etc.;
- atribute care facilitează procedura de generare a lexiconului din informațiile existente în *Dicționarul Morfologic*: CONVERSIE (dacă este posibilă conversia de la o clasă lexico-gramaticală la alta, cu specificarea claselor respective), PREFIXE (dacă admite derivarea cu prefixe, iar în caz afirmativ, se specifică prefixele respective), MEMBRU ÎN LOCUȚIUNE și STRUCTURA LOCUȚIUNII etc.

4. Concluzii

În vederea stabilirii structurii lexiconului *SIASTRO*, studiile s-au desfășurat pe două direcții: una de analiză a modelelor formale existente pentru descrierea datelor lexicale și terminologice, iar a doua, de analiză aprofundată a atributelor lexicale, morfologice și sintactice necesare realizării analizei sintagmatice pentru textele scrise în limba română.

Până în prezent s-au realizat procedurile de înscriere a atributelor noi în dicționar și a celor de generare a lexiconului necesar analizorului sintagmatic.

În viitor urmează să se adauge în lexicon structurile multi-cuvânt și informațiile de descriere terminologică. Așa cum este concepută acum structura lexiconului, aceste extensii se vor integra în mod natural în lexiconul existent.

REFERINȚE BIBLIOGRAFICE

- ISLE Meta Data Initiative 2003 = *Metadata Elements for Lexical Descriptions*, Version 1.1c, MPI Nijmegen, http://www.mpi.nl/IMDI/documents/Proposals/IMDI_Lexicon_1.1c.pdf
- ISO/CD 24613:2006 *Language resource management – Lexical markup framework(LMF)*. www.tc37sc4.org/new_doc/ISO_TC37_N130_rev9_LMF_15March2006.pdf, <http://tagmatica.fr/doc/ISO24613cdRev9.pdf>
- Antoni-Lay, Marie-Hélène, Gil Francopoulo, Laurence Zaysser (1994), *A Generic Model For Reusable Lexicons: The Genelex Project*. <http://perso.orange.fr/laurence.zaysser/lc94.html>
- Calzolari, Nicoletta, Alessandro Lenci, Francesca Bertagna, Antonio Zampolli (2002), *Broadening the Scope of the EAGLES/ISLE Lexical Standardization Initiative*, in *Proceedings of the 3rd*

- workshop on Asian language resources and international standardization*, Volume 12. <http://ucrel.lancs.ac.uk/acl/W/W02/W02-1204.pdf>
- Francopoulo, Gil, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, Claudia Soria (2006a) *LMF for Multilingual, Specialized Lexicons*, in Pierre Zweigenbaum, Stefan Schulz and Patrick Ruch (editors), *LREC 2006 Workshop on Acquiring and Representing Multilingual, Specialized Lexicons: the Case of Biomedicine*, Genova, Italy, 2006, ELDA. <http://estime.spim.jussieu.fr/~pz/lrec2006/Francopoulo.pdf>
- Francopoulo, Gil, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, Claudia Soria (2006b) *Lexical Markup Framework (Lmf) For Nlp Multilingual Resources*, in *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, pages 1–8, Sydney, July 2006. © 2006 Association for Computational Linguistics
- Jacquemin, Christian, Evelyne Tzoukermann (1999), *NLP for Term Variant Extraction: Synergy between Morphology, Lexicon, and Syntax*, in T. Strzalkowski (ed), *Natural Language Processing Information Retrieval*, p. 25-74, Kluwer, Boston, MA, 1999 <http://citeseer.ist.psu.edu/cache/papers/cs/22948/http:zSzzSzwww.limsi.frzSzIndividuzSzjacquemizSzFTPzSzjacqtzou-NLIR97.pdf/jacquemin99nlp.pdf>
- Lieske, Christian (2001), *The Open Lexicon Interchange Format (OLIF) Comes of Age*, Machine Translation Summit VIII 2001, www.olif.net/documents/olifMtSummitVIII.pdf
- Mc Cormik, Susan, Christian Lieske, Alexander Culum (2004), *OLIF v.2: A Flexible Language Data Standard* <http://www.olif.net/documents/olifMtSummitVIII.pdf>
- Mc Cormik, Susan (2005), *The Structure and Content of the Body of an OLIF v.2.0/2.1 File*. www.olif.net
- Melby, Alan K. (2000), *SALT: Standards-based Access service to multilingual Lexicons and Terminologies*. www.ttt.org
- Monachini, Monica, Francesca Bertagna, Nicoletta Calzolari, Nancy Underwood, Costanza Navarretta (2003), *Towards a Standard for the Creation of Lexica* www.elra.info/services/standard_lexica.pdf
- OLIF Consortium (2005), *OLIF – The Open XML Language Data Standard* www.olif.net
- Peters, Wim (2002), *Resurse Lexicale*. http://phobos.cs.unibuc.ro/roric/Ro/lex_introduction.html
- Romary, Laurent (2001), *Un modèle abstrait pour la représentation de terminologies multilingues informatisées: TMF – Terminological Mark-up Framework*, in *Cahiers GUTenberg* no. 39-40, mai 2001, <http://www.gutenberg.eu.org/pub/GUTenberg/publicationPDF/39-romary.pdf>
- Romary, Laurent, Marc Van Campenhoudt (2001), *Normalisation des échanges de données en terminologie: le cas des relations dites «conceptuelles»*, in *Conférence TIA-2001*, Nancy, 3 et 4 mai 2001, www.termisti.refer.org/tia4.pdf
- Romary, Laurent, Salmon-Alt S., Francopoulo G. (2004), *Standards going concrete: from LMF to Morphalou*, in *Workshop on Electronic Dictionaries*, Coling 2004, Geneva, Switzerland, <http://acl.ldc.upenn.edu/coling2004/W10/pdf/4.pdf>
- Segura Bedmar et alii 2006 = Segura Bedmar, Isabel, José L. Martínez Fernández, Paloma Martínez, *Including deeper semantic information in the Lexical Markup Framework: a proposal*, in *Proceedings of 5th Slovenian and 1st international Language Technologies Conference, 2006 IS-LTC'06* nl.ijs.si/is-ltc06/proc/15_Segura.pdf
- Tămăianu-Morita, Emma (2008) *Tipologia sintagmelor în modelul D. D. Drașoveanu. Posibile aplicații în proiectul SIASTRO*, supra, p. 137-150
- TEI *Guidelines for Electronic Text Encoding and Interchange. XML-compatible edition*, Edited by C. M. Sperberg-McQueen and Lou Burnard; XML conversion by Syd Bauman, Lou Burnard, Steven DeRose, and Sebastian Rahtz © 2001, 2002, 2004 by the TEI Consortium www.tei.org
- Wright, Sue Ellen (1999) *TO11: SALTing the Alphabet Soup. TC Forum 1998-2001* - <http://www.tc-forum.org> - file last updated 17 Oct 1999

FORMAL MODELS FOR REPRESENTATION OF LEXICAL
AND TERMINOLOGICAL INFORMATION IN *SIASTRO* PROJECT
(*Abstract*)

This paper outlines the research carried out within the *SIASTRO* project on the formal models for a uniform representation of lexical and terminological information. In this framework, we describe the structure of the lexicon that will be used in the phrase analysis of texts written in Romanian.

*Universitatea „Babeș-Bolyai”
Facultatea de Litere
Cluj-Napoca, str. Horea, 31*