CZU: 811.111:34

CORPUSURI ADNOTATE PENTRU PRELUCRAREA TEXTULUI JURIDIC

Nicoleta BAGHICI,
Universitatea de Stat „Alecu Russo” din Bălți

Abstract*: Legal terms can be mapped from annotated corpuses. A "semantic annotation" presents a much more precise description of the knowledge contained in the texts and of its semantics in the legal field. A semantic annotation has to be well defined, easy to understand by the experts in the field, and not ambiguous. To meet these requirements, a semantic annotation must be based on a formal model of the domain. Texts with terms annotated with the corresponding legal concepts describing this field will improve the process of extracting information from texts and documents with legal content for Romanian as well. Texts will help achieve far more qualitative results in Automated Translations by disambiguating legal terms.*
Annotation of legal texts can not be achieved without specific tools. In this case study, we will proceed to analyzing the Constitution concept with the PaLinKa software. In recent years, the need to produce reusable cells has led to an increasing use of XML encoding in annotation. As a result, annotation can not be applied using simple text editors. In addition, discourse annotation is usually complicated and requires specialized tools. In this subchapter, we will present the most important features of an annotation speech tool.
PaLinKa, the instrument presented in this paper, meets all these requirements, being appropriate for annotation of the legal texts.

Cuvinte cheie: terminologie juridică, adnotare, texte juridice, adnotări semantice, PaLinKa, unitate lexicală (LU), rol semantic (SR)..

FrameNet a stat la baza extragerii „automate a cadrelor sintactico-semantice. Aceste cadre vizualizează legătura dintre sens şi structura sintactică prin care acesta este redat. Ca bază pentru marcarea semantică sunt utilizate cadrele semantice – structuri conceptuale ce reprezintă evenimente, obiecte și proprietăți. Fiecare cadru include un set de elemente semantice (engl. *frame elements*). Fiecărui cadru îi este asociat un număr de cuvinte–unităţi ale lexiconului (engl. *lexicon units*)care evocă sensul cadrului dat (Fillmore : 2003).

Termenii juridici pot fi cartați din corpusuri adnotate. O „adnotare semantică” prezintă o descriere mult mai precisă a cunoștințelor conținute în texte şi a semanticii acesteia în domeniul juridic. O adnotare semantică trebuie bine definită, să fie uşor de înţeles de experţii din domeniu şi să nu fie ambiguă. Pentru a respecta aceste cerinţe, o adnotare semantică trebuie

să bazeze pe un model formal al domeniului. Textele cu termenii adnotaţi cu conceptele corespunzătoare din domeniul juridic, care descrie acest domeniu, va îmbunătăţi procesul de extragere a informaţiei din texte şi documnte cu conţinut juridic și pentru limba română. Textele vor contribui la obţinerea unor rezultate mult mai calitative în Traduceri Automate prin dezambiguizarea termenilor juridici (Fillmore:2002).

Pledăm pentru ideea prin care corpusurile adnotate să fie constant reactualizate. În vederea corectitudinii datelor este una din multele premise care pot elucida probleme persistente în folosirea în limbajul comun: exemple din presă.

This is an area where, by *acts* of commission and omission on the part of ... and laity over many decades, we have, I confess, *acted* shamefully (Кубрякова 1997: 435) ”Cautionary tales of Baba Yaga *acted* to deter children from wandering ... It *acts* as a place for refugees to flee from war, a source of income for ...” (Hugo : 2013)

Verbul	Definiție	Semantic Roles	Exemple adnotate
ACT(V) - a acți / a secomporta / a interpreta / a juca	To carry out an action / To appear or seem to / Something done that has legal significance/	[ROLE] / [PERFORMER]	, declaring what Officer shall then act as President , such Officer shall act accordingly ,
ABRIDGE(V) – prescurta/a reduce a limita/ a restring drepturile/	to reduce length of (a written work) by condensin or rewriting / to curtail; diminish	[AGENT]/ [ATTRIBUTE]/ [CAUSE]/ [DIFFERENCE]/ [ITEM]	which shall abridge the privileges or immunities of citizens of the United S itizens of the United States , or in any way abridged ,

Folosirea în limbajul comun: exemple din presă: (Turner : 2015).

Atenționăm că polisemantismul termenilor poate fi o altă instigare pentru construirea corpusului juridic. Termenii juridici în limba română în domeniul penal, procesual penal, pot evoca diferite scenarii, și diferite cuvinte pot evoca același scenariu. Cuvântul *accuse* are câteva cadre semantice, evocînd cîteva unități lexicale (Judgment_communication), (Notification_of_charges), (Judgment) –ca un cuvânt mai general decât unitatea lexicală *charge.* În timp ce *accuse* (a acuza) evocă cadrul CRIMINAL_INVESTIGATION , unitatea lexicală *accuse* (to charge) evocă cadrul CHARGING. La fel și cu *abrodge* sau *act.*

Domeniile juridic și penal includ multe exemple de cuvinte polisemantice. Cuvântul (engl. to testify), care semnifică *a depune mărturie* ,evocă cadrul CRIMININAL_INVESTIGATION, cadrul PROBATORY_HEARING, și cadrul COURT_EXAMINATION. Cuvântul (a mărturisi) evocă aceleași cadre. Deși locuțiunea *a depune mărturie* (engl. to testify) și verbul *a mărturisi* (engl. to give evidence) pot fi considerate ca sinonime în anumite contexte, acestea prezintă o variație de sens. Cuvântul *a depune mărturie* în general este legat de martorul unei infracțiuni, atunci când el / ea decide să depună mărturie în mod voluntar, în timp ce cuvântul (engl. *to give evidence*) este folosit în contexte când persoana este impusă de o autoritate să depună mărturie. Diferența de sens dintre ele ar putea fi modelată într-un lexicon juridic bazat pe crearea a două cadre diferite: martor și acuzat (engl. WITNESS AND ACCUSED). Cadrul MARTOR va fi evocat de verbul *to tesify* (*a depune mărturie)* și elementul ACUZAT va fi evocat de verbul (*a mărturisi*). Adnotatorul trebuie să analizeze contextul în care este introdusă unitatea lexicală, în scopul de a alege cadrul însușit(Fillmore:1968).

Bazându-ne pe analiza resurselor terminologice de nominalizare *ale elementelor modelului mental a realității în aspectul juridic,* putem face următoarea concluzie: în diferite cadre modelate în discursul juridic, se disting caracteristici invariante și componente în interiorul cărora pentru fiecare cadru, există diferențe și trăsături personalizate. Într-adevăr, după cum a mai fost menționat de către cercetători, tipul cadrului ca reprezentare mentală este determinat de specificul reprezentării codificate.

Adnotarea discursului juridic nu poate fi realizată fără ajutorul unor instrumente specifice.

În acest studiu de caz, vom purcede la analizarea conceptului *Constitution* cu ajutorul softului PaLinKa. În ultimii ani, necesitatea de a produce corpusuri reutilizabile a condus la o utilizare tot mai mare de codificare XML în adnotare. Drept urmare, adnotarea nu poate fi aplicată folosind editoare de text simple. În plus, adnotarea discursului este, de obicei complicată și necesită instrumente specializate. În acest subcapitol, vom prezenta cele mai importante caracteristici ale unui instrument de discurs de adnotare (Fillmore:1976).

PaLinKa, instrumentul prezentat în această lucrare, întrunește toate aceste cerințe, fiind adecvat pentru adnotarea discursului juridic.

Documentul *Constitution* a fost împărțit în 9 fișiere XML ce se deschid respectiv cu un alt program Notepad ++ și care este destul de simplu de instalat pe PC.(U.S. Constitution).

```
[MARKER]
NAME:CRIME
BGCOLOR:10,17,2
FGCOLOR:248,248,250
ATTR:ID=# ;defines an unique id for each tag
PREFIX:ACT
PUT_IN_TREE:1
HOT_KEY:F6

[MARKER]
NAME:MESSAGE
BGCOLOR:243,109,132
FGCOLOR:248,248,250
ATTR:ID=# ;defines an unique id for each tag
PREFIX:ACT
PUT_IN_TREE:1
HOT_KEY:F6

[MARKER]
NAME:ACT
BGCOLOR:163,243,109
FGCOLOR:8,17,2
ATTR:ID=# ;defines an unique id for each tag
PREFIX:ACT
PUT_IN_TREE:1
HOT_KEY:F6
```

Fig. 1. O parte din fișierul de preferințe folosit pentru adnotare

Acestea sunt cîteva preferințe ce au fost folosite pentru adnotarea cuvintelor de referire. NAME: - denumirea tagului poate fi ales de

adnotator, în funcție de cadrele și rolurile semantice RS pe care dorește să le evidențieze, ACT, CRIME, MESSAGE. Prin urmare, pe ecranul principal al programului nu se afișează tagurile XML, astfel încât textul poate fi ușor de citit. În scopul de a identifica tag-urile prezente în text, am specificat o bază de culori pentru a afișa textul adnotat și pentru a avea limitele marcate în mod explicit, în cazul dat, au fost afișate și etichetate patru fișiere de taguri care nu se repeta (în total 140 taguri), nici după denumire, nici după nuanța lor cromatică.

Adnotarea coreferențială este în mod notoriu o însărcinare consumatoare de timp și muncă. Am marcat legăturile coreferentiale dintre entități, din text în mod manual. De obicei, fiecare entitate primește un ID unic și un link între cele două entități ce a fost marcat folosind aceste ID-uri. Aceste ID-uri sunt gestionate automat de programul PaLinKa. Unele link-uri se referă la mai mult de o singură entitate. Acest fapt poate fi, de asemenea, codificat .

Lanțurile coreferentiale pot fi identificate rapid prin utilizarea arborelui entităților în partea dreaptă a ecranului (vezi figura 1) sau prin evidențierea lor. Fiecare cadru conține mai multe unități lexicale LU, unele dintre aceste, de asemenea, includ alte LU. Având în vedere această bogăție de tag-uri, unul din avantajele de a folosi acest program este, în mod notoriu, că el ascunde tagurile XML, folosind culori pentru fiecare etichetă. În afară de aceasta, este posibil să adaptăm programul pentru a marca începutul și sfârșitul fiecărei etichetări, folosind un caracter ales de noi. Această caracteristică s-a dovedit, de asemenea, a fi utilă pentru adnotarea întreprinsă. Este posibil să se observe în Figura 1, unde limitele fiecărui tag sunt marcate prin paranteze patrate. Fiecare tag (rol sintactic RS) are o culoare aparte, deoarece am analizat exclusiv verbele din Constituția SUA, aceastea având culoarea galbenă. De exemplu LU CHOOSE (v), are următoarele RS COGNIZER / POSSIBILITIES / CHOSEN, iar cadrul este Choosing (Anexa 3). COGNIZER ia decizii pentru CHOSEN (fie acesta un element sau un curs de acțiune), dintr-un set de POSSIBILITIES. COGNIZER poate avea un INTENDED_PURPOSE pentru CHOSEN (Atkins:1994).

Fig. 2 Exemplu LU și RS marcate în text

CHOSEN identifică entitatea sau cursul de acțiune, care este ales dintre POSSIBILITIES.

COGNIZER îl alege pe CHOSEN din toate POSSIBILITIES.

COGNIZER face o alegere dintre un set de POSSIBILITIES. POSSIBILITIES sunt de obicei exprimate prin intermediul unei fraze oblice, care indică alternativa sau alternativele, sau de către o propoziție subordonată (alegerea de a face sau alegerea să nu o facă), de obicei, condusă de „sau" sau „dacă ".

Schematic va fi reprezentat astfel:

Tabelul 1. Relația semantică pentru verbul CHOOSE

№	Subiect	verb	CHOSEN	COGNIZER	POSSIBILITIES
1	The Members	chosen	every second year	by the people of the several States	

2		choose	their other Officers...a President pro tempore	The Senate	
3		choose	heir Speaker and other Officers	The House of Representatives	
4		choose	a President	The House of Representatives	whenever the right of choice shall have devolved upon them
5		choose	a Vice President	Senate	whenever the right of choice shall have devolved upon them

Verbul *choose* (a alege), după tabelul de mai sus, necesită obligator CHOSEN si COGNIZER (ES nucleare), optional însă POSSIBILITIES (ES non- nucleare). Verbele cu sensuri aferente precum *pick, select* sau *opt*, urmează să aibă aceleași însemnătate. Fiecare dintre aceste verbe indică sau evocă diferite aspecte ale cadrului.Verbul *choose* (a alege), se focalizează pe CHOSEN și COGNIZER, avînd ca fundal POSSIBILITIES sau alte explicații. Ideea este că cunoașterea sensului oricărui dintre aceste verbe necesită să se știe ce are loc într-un anturaj juridic, în cazul dat, cunoașterea conținutul Costituției SUA și cunoașterea sensului structurat de cadru oferă fundal și motivație pentru categoriile reprezentate de cuvinte. Cuvintele, care sunt material lingvistic, evocă cadrul (în mintea unui vorbitor / ascultător); interpretul (o exprimare sau un text în care apar cuvintele) invocă cadrul. O descriere completă a acestor verbe trebuie să includă, de asemenea, informații cu privire la proprietățile lor gramaticale și diverse modele sintactice în care acestea apar. Ce elemente sau aspecte ale cadrului pot fi realizate ca subiect al verbului, ca obiect, în cazul în care există, și care va fi forma de suprafață dintre celelalte? Care dintre aceste elemente sunt opționale și care sunt obligatorii? De exemplu în propoziția, *The Members chosen every second year by the people of several States, The Members* este subiectul , *chosen* verbul, *every second year* – TIME, *by the*

people of several States – COGNIZER. Dintre acestea, CHOSEN and COGNIZER sunt obligatorii, şi reprezentate în forma de suprafaţă de grupuri nominale.

Definirea cuvintelor în ceea ce privește cadrele și prototipurile oferă o abordare utilă a problemei limită pentru categoriile lingvistice. Pentru a ilustra această abordare, este mai degrabă necesar de definit ca un cadru de fundal, decât în termeni referitori la toate circumstanțele neobișnuite în care ar putea fi utilizat cuvântul. Faptul că verbul *choose* ar putea să apară în contexte care nu se potrivesc cu prototipul ne sugerează că vorbitorii sunt dispuși să extindă cadrul cuvântului sau să creeze un nou cadru.

Alt concept aplicat este cel de perspectivă. În exemplul *The Senate choose their other Officers ... a President pro tempore,* evocând cadrul *Choosing,* În timp ce se menționează toate elementele cadrului, raport asemănător anturajului juridic, prioritar ar fi punctul de vedere al celui care alege – COGNIZER. Similar, propoziția *their other Officers... a President pro tempore was chosen by the Senate* este raportul anturajului juridic din perspectiva celui care alege (COGNIZER).

Referinţele Bibliografice

BTS Atkins, *Analysing the verbs of seeing: a frame semantics approach to corpus lexicography. In Susanne Gahl, Johnson, Christopher, and Dolbey, Andrew, Proceedings of the Twentieth Annual Meeting of the Berkeley Linguistics Society*. Berkeley Linguistics Society, 1994.

https://framenet.icsi.berkeley.edu/fndrupal

Charles Fillmore,*The case for case. In E Bach and Harms, R, Universals in Linguistic Theory*, Universals in Linguistic Theory. Holt, Rinehart & Winston, New York, 1968, p.14-25.

Charles J Fillmore, *Frame semantics and the nature of language*. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 1976, p.20-32.

Charles J Fillmore, Petruck, Miriam RL, Ruppenhofer, Josef, and Wright, Abby. *FrameNet in Action: The Case of Attaching. IJL*, 2003, p.16:297-332.

Charles J Fillmore. In *Proceedings of 19th International Conference on Computational Linguistics*, Taipei. COLING, 2002.

Fourth Amendment – U.S. Constitution, http://constitution.findlaw.com/amendment4/amendment.html#sthash.Vzntl5je.dpuf (visited 20.12.2017).
Sama Hugo, ”*The Vanquishing of Witch Baba Yaga***”,**USA, 2013, http://screen.artshub.com.au/news-article/reviews/film/sama-hugo-/the-vanquishing-of-witch-baba-yaga-245187 (visited 21.04.2014).
Turner, Nancy,„*Republicans Violate 26th Amendment And 'Abridge' Youth Voting Rights In NC Turner, Nancy “American University Law Review*”, The Young and the Restless: How the Twenty-Sixth Amendment Could Play a Role in the Current Debate over Voting Laws, July 1, 2015.(visited 10.02.2016).
Кубрякова Е.С. *Части речи в когнитивном аспекте*. - М., 1997, p. 435.