PARTICULARITĂȚI SPECIFICE FOLCLORULUI DIN REPUBLICA MOLDOVA COMPARATIV CU CEL DIN ROMÂNIA. STUDIU PE UN CORPUS ADNOTAT SUB FORMĂ DE ARBORI SINTACTICI

CĂTĂLINA MĂRĂNDUC[*], VICTORIA BOBICEV[**]

Specific Peculiarities of the folklore from the Republic of Moldova compared to Romania's folklore. Study on an annotated corpus in the form of syntactic trees

Abstract: *We have created a subcorpus of the treebank in dependency grammar of the Faculty of Computer Science, Al. I. Cuza University, containing folklore collected from the Republic of Moldova, annotated there by the students of the Technical University of Moldova. Together with an Old Romanian sub-corpus, they were then included in the Universal Dependencies website, under the name Ro_UD_nonstandard treebank, along with 140 other corpora for 60 languages. On this corpus we applied rhyme annotations and statistics. For now, we do not have an automatic rhyme annotation program that counts the syllables of the verses and the Romanian folklore sub-corpus is too small, but we described our corpus and still have some working hypotheses.*

Keywords: *treebank, nonstandard language, oral cultural heritage, folklore, syntactic structure, types of rymes.*

Introducere

În ultima vreme corpusul de tip *Treebank* de dependență al Facultății de Informatică a crescut (are acum 20 094 fraze și 389 281 de cuvinte și semne de punctuație) și a primit un cod de identificare ca resursă internațională, anume: UAIC-RoDia = ISLRN 156-635-615-024-0. Denumirea s-a schimbat puțin ca urmare a faptului că nu au mai fost adnotate traduceri din limba engleză în limbaj standard (cu intenția importării adnotărilor sau a realizării unor corpusuri aliniate), ci texte în orice fel de limbaj nonstandard, mai ales de limbă veche, numindu-se RoDia (Romanian Diachronic). Noi am reușit să explicăm aceste idei conducătorilor

[*] Facultatea de Informatică, Universitatea „Alexandru Ioan Cuza" Iași; Institutul de Lingvistică al Academiei Române Iorgu Iordan – Alexandru Rosetti, București.

[**] Universitatea Tehnică a Moldovei, Chișinău.

proiectului UD (*Universal Dependencies*), care au creat o ramură pentru treebankul nostru, numită UD-Romanian-Nonstandard, alături de treebankul românesc ce face parte din CoRoLa și deci conține limbă română standard. Treebankul din Pennsylvania (*Penn Treebank*) și cel din Republica Cehă, *Prague Dependency Treebank* (PDT), având milioane de cuvinte și fiind aliniate pe o mare parte a lor, fac parte dintre fondatorii proiectului.

Participarea la UD este deosebit de importantă, atât pentru noi, cât și pentru Republica Moldova (Mărănduc, Bobicev 2017). Acolo se află 140 de treebankuri pentru 60 de limbi (și numărul lor e în creștere), toate adnotate în aceleași convenții, cu intenția de a se antrena pe ele un parser universal, adică un analizor sintactic automat care poate fi antrenat pe orice limbă și, după antrenare, analizează sintactic automat texte în acea limbă. Se pot face orice fel de comparații și alinieri între *treebank*-urile adnotate la fel. Există mai multe *treebank*-uri care conțin *Noul Testament* sau *Biblia*. Toate datele de pe acest site sunt *open source*, deci oricine poate avea acces la eșantioane de limbă „moldovenească" și românească pentru a le compara.

Dependency este numele gramaticii computaționale care stă la baza convenției de adnotare. Nu o vom descrie aici, vom menționa doar că relațiile sunt înscrise pe arcuri și toate nodurile sunt cuvinte sau semne de punctuație; relațiile de egalitate sau subordonarea de la mai multe *head*-uri nu sunt permise; coordonarea este dificil de adnotat în limitele acestui model. În schimb, el este foarte flexibil și realizează o mare economie de informație, neconsiderând relațiile și clasificările teoretice ca fiind noduri (nonterminale) în arbore.

Majoritatea limbilor sunt reprezentate în UD prin mai multe *treebank*-uri și se fac periodic, automat, statistici cu asemănări și deosebiri între *treebank*-uri pentru aceeași limbă. În cazul nostru, putem realiza, gratuit și automat, o comparație calitativă și cantitativă între limba română standard și limba română nonstandard.

1. Termenul de limbă română nonstandard

Vom argumenta decizia noastră de a introduce în UD un corpus de limbă română nonstandard. Limba română contemporană standard este obiectul unui mare proiect de curând lansat, CoRoLa (Corpus computațional de referință pentru limba română contemporană, http://corola.racai.ro/), care conține și un *treebank* numit RRT (*Reference Romanian Treebank*) la care am contribuit și noi cu 4 000 de arbori din corpusul care a constituit rezultatul cercetării doctorale a lui Augusto Perez (2014). Dar limbajul natural se manifestă preponderent prin comunicări nonstandard. Este obligatoriu ca o persoană să se exprime conform cu normele academice din DOOM[2], care constituie standardul actual al limbii, atunci când scrie cărți, articole științifice, lucrări de licență, masterat sau doctorat, acte, scrisori oficiale, comunică formal cu superiorii, când susține conferințe, lecții sau cursuri universitare ori în discursul public.

Comunicarea familiară (informală) nu este supusă regulilor și se poate desfășura în moduri oricât de inventive, destinate să obțină efecte asupra

interlocutorului sau se poate desfășura în conformitate cu niște reguli comunicative specifice unui grup restrâns, rezultate din practică îndelungă și din dorința de a abrevia comunicarea. Comunicarea pe *chat*, de care ne-am ocupat, reprezintă aspectul scris al comunicării familiare și este încă și mai creativă. Prin urmare, orice tip de mesaj social-media (*comments*, *twitter*, *chat*, *messenger* etc.) este comunicare nonstandard și – nu mai este nevoie să demonstrăm un lucru îndeobște cunoscut – crește vertiginos ca volum.

Limba română veche, păstrată în tipărituri sau în manuscrise din secolele XVI–XIX este, firește, cât se poate de îngrijită, dar nu o putem considera limbaj standard pentru că nu știm la ce standarde se raportează. Regulile de corectitudine nu erau fixate, erau diferite pentru regiuni, curente, perioade, locul unde scriitorul își face studiile ori originalul după care traduce. Noi ne propunem ca, prin ocerizarea acestor vechi tipărituri, să avem acces direct la textul originiar, iar nu prin intermediul editorilor contemporani oricât de pricepuți, așa încât să putem extrage normele după care textele vechi sunt scrise și să le introducem în memoria computerului. Un grup de cercetători de la IMAT (Institutul de Matematică și Informatică din Chișinău) a realizat un astfel de program de recunoaștere optică a caracterelor din imaginea unei tipărituri, scrise în secolul al XVI-lea sau al XVII-lea cu litere vechi chirilice românești.

Prezentat la o conferință axată pe astfel de programe, DATeCH (*Digital Access of Textual Heritage*)n în iunie 2017, la Gōttingen[2] (Cojocaru, S. *et al*. 2017), programul s-a dovedit compatibil cu cele folosite la ora actuală pentru citirea tipăriturilor vechi gotice sau a inscripțiilor copte.

Ne-am propus să procesăm *Noul Testament* de la Alba Iulia (1648), primul tipărit în România, care nu a făcut obiectul unor proiecte de digitizare așa cum este cazul cu *Biblia de la București* (1688) și este o traducere din altă sursă. Intenția a fost să facem posibilă alinierea lui cu *Noul Testament* grec, latin, armean, vechi slavon, aflate deja pe site-ul UD; începând cu 15 aprilie 2018, am îndeplinit pe jumătate acest obiectiv pentru că am introdus în directorul nostru UD-Romanian-Nonstandard toate cele patru *Evanghelii* și prefețele lor.

Cât despre comunicarea regională, și aceasta se desfășoară potrivit unor reguli, doar că ele diferă de la o regiune la alta și uneori chiar de la un sat la altul. Este acesta un motiv să nu facă obiectul digitizării, al prelucrării limbajului natural? Ce fel de instrumente de procesare vom antrena pe un corpus cum este CoRoLA? Unele care se vor poticni la fiecare fenomen creativ sau la fiecare incorectitudine involuntară. Limbajul natural evoluează.

Apoi, folclorul este o parte a moștenirii culturale, mai ales pentru popoare cu scriere târzie, și e un fenomen care dispare rapid, pe măsură ce toată lumea are acces la cultura scrisă. Cu puțin noroc, tipăriturile vor mai rezista două sute de ani, dar, în câțiva ani, ultimii oameni în vârstă, care mai știu pe de rost texte transmise cândva oral, nu vor mai fi.

[2] http://ddays.digitisation.eu/datech-2017/.

În România, folclorul a fost cules și tipărit încă de pe vremea lui Gustav Weigand, Alecu Russo și a lui Vasile Alecsandri, cu metode tot mai corecte, la început cizelat de poeți, apoi înregistrat pe benzi magnetice transcrise și editate ulterior. Dar în Republica Moldova nu a existat interes pentru publicarea folclorului. Unii profesori de la școli rurale, pensionari, dețin caiete în care au notat texte. Dar s-a publicat extrem de puțin înainte de anii ’90. Aceste creații exprimau sentimente și idei contrare liniei oficiale, dușmănie față de politica dominantă. Chiar dacă erau simple obiceiuri de Anul Nou, cine avea să popularizeze tradițiile unui popor pe care conducătorii sovietelor intenționau să-l deznaționalizeze prin strămutarea dintr-o regiune în alta?

Probabil că a existat și la noi folclor anticomunist care a fost cenzurat. Nu am extins foarte mult cercetarea. În studiul lui Grigore Bostan (1998), din care am extras toate citatele și anexele cu texte populare din Moldova și Bucovina, rezultând 693 de fraze introduse acum în UD, se menționează sursele de unde au fost preluate texte: Cireş, Berdan 1982 şi Covalciuc, Bostan 1993. Alte variante înregistrate în actuala regiune Cernăuţi în zilele noastre au fost găsite în Arhiva de folclor a Catedrei de filologie română şi clasică a Universităţii din Cernăuţi. Studiul lui Grigore Bostan urmărește să demonstreze că există motive, teme și procedee comune folclorului de pe teritoriul României și celui creat de românii din afara lui. Noi urmărim să extindem analiza, fără a nega concluziile acestui cercetător, observând ce anume le deosebește, care sunt particularitățile distinctive ale fiecărei regiuni.

O altă culegere procesată este *Folclor din părțile Codrilor* (Botezatu *et al.* 1973), care cuprinde texte folclorice de diferite genuri, culese în perioada 1962–1972 din satele din regiunea centrală a Republicii Moldova, în expediții folclorice organizate de Academia de Științe a Moldovei, secția folclorică. Cartea este tipărită cu litere chirilice, iar cercetătorii de la IMAT au ocerizat-o și au transpus-o în alfabet latin (Bobicev *et al.* 2016). Prelucrarea textelor extrase din această culegere este în curs; 208 fraze au fost adnotate și introduse până acum în directorul nostru de la UD, iar aproximativ 1000 sunt în lucru.

Cartea este precedată de un cuvânt introductiv în care autorii se străduiesc să demonstreze că fantasticul și superstițiile sunt pe cale de dispariție din mentalul colectiv reflectat în folclor. Materialul prezentat confirmă în parte aceste concluzii comandate ideologic în anul publicării. Folclorul din Republica Moldova fiind mai aproape de contemporaneitate, baladele capătă trăsături de fapt divers de senzație și își pierd motivele fantastice, dacă nu e vorba de motive tradiționale foarte vechi. Însă specia basmului, poezia obiceiurilor și cea a descântecului nu își pierd caracterul mitic și ritualic, care va dispărea doar odată cu ele.

Ambele surse din R. Moldova conțin numele informatorilor, vârsta și localitatea unde trăiesc, uneori și date despre profesie și educație; majoritatea informatorilor au peste 70 de ani; câțiva mai tineri știu doar puține cântece de dragoste.

Până acum, nu am adnotat decât 230 de fraze de folclor din România, extrase din Brăiloiu *et al.* 1978, Bălășel 1967, Datcu 1968, culegeri de care am dispus în 2015 în format electronic, iar Augusto Perez le-a adnotat pentru a fi folosite la dicționarul de *pattern*-uri verbale. Avem în lucru și texte din Amzulescu 1964, culese între 1883 și 1939.

În acest articol ne propunem să facem cunoscut corpusul nostru de folclor, adnotat în convențiile Universal Dependencies și în cele ale *treebank*-ului UAIC-RoDia. De asemenea, vom arăta ce tip de date sunt adnotate pe acest corpus și pot fi folosite la diverse tipuri de cercetare. În fine, vom prezenta sistemul nostru de adnotare a rimelor și datele extrase automat din corpusul adnotat. Vom face o schiță de interpretare a datelor, pe care mărirea corpusului ar putea să o confirme sau, dimpotrivă, să o infirme. Vom trasa direcţii pentru crearea în viitor a unui adnotator automat al rimelor. În ultima parte a articolului, vom prezenta și alte observații asupra corpusului (referitoare la topică, la regionalisme, la frecvența unor circumstanțiale), care au același statut de concluzii provizorii.

2. Prezentarea corpusului

După cum am putut observa, culegerile de folclor din Moldova conțin texte culese după 1962, folclor mai nou decât cel din culegerile românești. În perioada sovietică nu se promovau tradițiile culturale ale națiunilor dominate. În afară de aceasta, unele creații populare exprimau idei politice de opoziție; acestea sunt selectate de preferință în cartea lui Grigore Bostan (1998), publicată în România într-o perioadă în care exista entuziasm pentru libertatea cuvântului. Astfel de texte nu există în celelalte surse citate, probabil au fost cenzurate sau autocenzurate de culegători sau editori.

Noutatea textelor nu are numai consecințe ideologice; fiind un fenomen cu autor colectiv, cu cât este mai veche, o creație folclorică este mai cizelată, toate rimele imperfecte sunt înlocuite, versurile nu mai au măsură variabilă, pentru că fiecare autor când învață și transmite poezia, o mai îmbunătățește; folclorul vechi are mai mulți autori decât cel nou. E adevărat că, în vechime, poeți ca Vasile Alecsandri au modelat ei înșiși poeziile populare publicate, ceea ce metodele științifice actuale nu mai permit. Prin urmare, vechimea este motivul pentru care cele 230 de fraze din folclorul românesc sunt mai corecte din punctul de vedere al rimei decât cele din Republica Moldova.

Frazele au fost adnotate în prealabil în convențiile formatului de bază UAIC, stabilite în perioada 2007, când a început lucrul la *treebank*-ul FII-UAIC, și 2014 (Perez 2014). Acest format este pivotul de la care un program numit TREEOPS obține formatul UD supervizat automat și formatul Semantic al *treebank*-ului (Colhon *et al*. 2017). Trecerea în formatul UD se efectuează cu pierdere de informație semantică deja adnotată, de aceea am conceput formatul Semantic, în care informația semantică este păstrată și îmbogățită (Mărănduc *et all.* 2018). Corpusul de folclor nu a fost încă transpus în formatul Semantic, dar transpunerea este oricând posibilă.

Adnotarea UD este realizată în formatul CONLLU, așa cum sunt toate celelalte 140 de *treebank*-uri pentru cele 60 de limbi care se află pe acest site. Ea conține: pe coloana 1, *id*-ul cuvântului, pe a doua, forma cuvântului în text, pe coloana a treia, *lemma*, pe coloanele 4–6, analiza morfologică, redată prin sigle mai generale care arată doar partea de vorbire (U-POSTAG) apoi prin sigle detaliate care conțin analiza completă (X-POSTAG) și apoi explicarea informației morfologice codificate în X-POSAG; pe coloana a șaptea se află *id*-ul cuvântului *head* (pe care îl determină) și pe coloana a opta, numele relației de dependență pe care o stabilește cu acesta, așa cum apare în setul de relații UD.

Coloana a noua este rezervată pentru analize sintactice îmbogățite, iar coloana a zecea conține informații despre (non)existența spațiului între *token*-i, precum și orice alt tip de informație. În corpusul *Evangheliei* de la 1648 se află sigla evanghelistului, capitolul și versetul, iar în corpusul de folclor am adnotat rima. Exemplu:

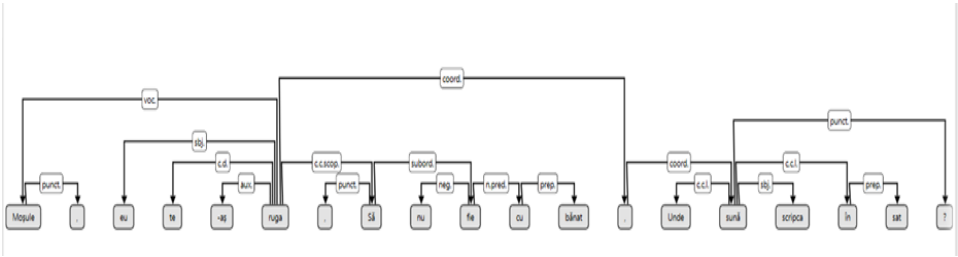

Figura 1. Frază din folclorul din R.Moldova (*Moșule, eu te-aș ruga, Să nu fie cu bănat, Unde sună scripca în sat?*) adnotată în format UAIC.

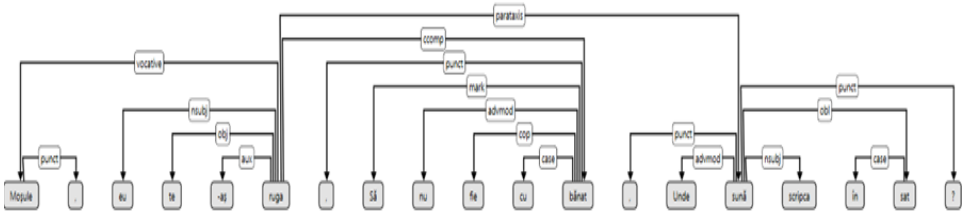

Figura 2. Fraza din figura 1 adnotată în format UD.

În figurile 1 și 2 se poate remarca diferența de concepție între cele două tipuri de adnotare. Adnotarea UAIC stabilește relațiile sintactice cu ajutorul cuvintelor conectori care sunt plasate între cuvintele cu sens deplin, subordonate față de cuvântul regent și *head* pentru cuvântul subordonat. În adnotarea UD, conjuncțiile și prepozițiile, cum ar fi *să*, *cu*, *în* sunt subordonate cuvintelor cu sens deplin și nu pot să-și subordoneze nimic. La fel este tratat și verbul copulativ (*fie*), subordonat de numele predicativ (*bănat*).

În cele ce urmează, se poate observa specificul formatului CONLLU; acesta nu are categoriile marcate urmate de semnul egal (=) și de valoarea lor între ghilimele, ca în XML, ci informația privind valorile morfologice se află pe una dintre cele 10 coloane care exprimă categoria de date. Pe coloana a 10-a apare notația SpaceAfter=No dacă punctuația sau cuvântul cu cratimă este lipit de cuvântul anterior. Această notație este foarte strictă și este destinată calculatoarelor, pentru ca ele să poată reface din analiză textul așa cum apare în limbajul natural.

Dar, cum pe coloana a 10-a se permite adăugarea oricărei informații suplimentare, noi am adăugat aici rimele, pe care le-am adnotat potrivit unei convenții stabilie de noi. În acest exemplu avem doar o rimă, de tip împerecheat (Paired), între cuvintele cu *id* 12 și 18. Vom detalia în comentariile ulterioare constatările cu privire la rimă în cele 3 documente procesate, unul, cu folclor din România, și două, cu folclor din Republica Moldova. Acolo unde există mai multe informații pe o coloană, ele sunt ordonate alfabetic și despărțite prin bară verticală. Este cazul coloanei a 6-a care conține explicarea analizei morfologice. Pe coloana a 10-a am înscris *id*-ul cuvântului cu care rimează și tipul de rimă, dar cum informația SpaceAfter se găsește alfabetic între Rhyme și Type, ea este plasată ca atare și programul nostru de analiză a rimei trebuie să o elimine.

Faptul că rima este codificată face ca aceasta să poată fi extrasă de un program de computer și prelucrată statistic. Alte tipuri de informație, codificate fiind, pot de asemenea să fie extrase, fie din formatul acesta, fie din cel XML al adnotării UAIC. De pildă, cât de mare este ponderea vocativelor, a complementelor de timp, sau a complementelor de loc (care nu sunt adnotate în formatul UD, unde toate circumstanțialele devin *obl, advcl* sau *advmod,* conform convenției lor de analiză morfologică).

```
# sent_id = train-6102
# text = Moșule, eu te-aș ruga, Să nu fie cu bănat, Unde sună scripca în sat?
# Rhyme=Yes
...
8   Să      să      PART    Qs      PartType=Sub                    12      mark    _       _
9   nu      nu      ADV     Qz      Polarity=Neg                    12      advmod  _       _
10  fie     fi      VERB    Vmsp3   Mood=Sub|Person=3|Tense=Pres|VerbForm=Fin
    12      cop     _       _
11  cu      cu      ADP     Spsa    AdpType=Prep|Case=Acc           12      case    _
    _
12  bănat   bănat   NOUN    Ncmsrn
    Case=Acc,Nom|Definite=Ind|Gender=Masc|Number=Sing   6       ccomp   _
    Rhyme=ID18|SpaceAfter=No|Type=Paired
13  ,       ,       PUNCT   COMMA   _       15      punct   _       _
14  Unde    unde    ADV     Rw      PronType=Int,Rel        15      advmod  _       _
15  sună    suna    VERB    Vmip3s
    Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   6       parataxis   _
    _
16  scripca scripcă NOUN    Ncfsry
    Case=Acc,Nom|Definite=Def|Gender=Fem|Number=Sing        15      nsubj   _       _
```

17 în în ADP Spsa AdpType=Prep|Case=Acc 18 case _ _
18 sat sat NOUN Ncmsrn Case=Acc,Nom|Definite=Ind|Gender=Masc|Number=Sing 15 obl _ **Rhyme=ID12**|SpaceAfter=No|**Type=Paired**
19 ? ? PUNCT QUEST _ 15 punct _ _

Iată și un fragment din adnotarea UAIC în XML. În acest format, nu găsim explicația siglelor morfologice. Există 14 tipuri de complemente circumstanțiale, ceea ce reprezintă o mare cantitate de informație semantică. Mai găsim în clar numele caracteristicilor pe care se bazează sutele de reguli cu care TREEOPS transformă convenția UAIC în alte formate. TREEOPS transformă textul din format UAIC–XML în format UD–XML și, după aceea, un alt program numit „xml2conllu" transformă acest UD–XML în CONLLU, adăugând explicațiile de pe coloana a 6-a, numerele de capitol și de verset la *Evanghelii*, SpaceAfter=No, și extrăgând textul frazei din schemă.

<word id="14" form="Unde" lemma="unde" postag="Rw" head="15" chunk="" deprel="c.c.l."/>
<word id="15" form="sună" lemma="suna" postag="Vmip3s" head="13" chunk="" deprel="coord."/>
<word id="16" form="scripca" lemma="scripcă" postag="Ncfsry" head="15" chunk="" deprel="sbj."/>
<word id="17" form="în" lemma="în" postag="Spsa" head="15" chunk="" deprel="c.c.l."/>
<word id="18" form="sat" lemma="sat" postag="Ncmsrn" head="17" chunk="" deprel="prep."/>
<word id="19" form="?" lemma="?" postag="QUEST" head="15" chunk="" deprel="punct."/>
</sentence>

3. Sistemul UD_Romanian-Nonstandard de adnotare a rimei

Potrivit unei definiții clasice, rima este similitudinea sunetelor de la finalul unor versuri, începând cu vocala accentuată. Noi am lărgit puțin definiția, incluzând în conceptul de rimă și cuvinte care nu se află la finalul versurilor, ceea ce am numit rimă internă (Type=Intern). De asemenea, am adnotat și repetițiile, Type=Rep, rime la care coincid toate sunetele unor cuvinte, fie la final, fie în interiorul versului, ba chiar adeseori la început de versuri. Repetițiile contribuie mult la muzicalitate și facilitează memorarea poeziei. Repetițiile, ca și monorima (Type=Mono), pot avea nu doi, ci trei sau mai mulți membri. Am întâlnit mai mulți membri uneori și la Type=Imperf, adică tipul imperfect. Acest fenomen este denumit în statisticile extrase Type Multiple, adică rime având mai mulți membri.

În corpusul nostru, inițiala de vers este marcată, de obicei, prin literă mare și prin semne de punctuație. Prin urmare, rima se situează înainte de cuvintele cu literă mare din interiorul frazei. Nu am studiat bogăția rimei sau locul unde se află accentul în rimă, așa-numitele rime masculine și feminine.

Deci am notat în corpusul UD, pe a zecea coloană, la începutul frazei, unde sunt comments #Rhyme=Yes, sau, după caz, #Rhyme=No, pentru cazul când era vorba de povești, legende toponimice sau descântece în proză. Dacă am notat Yes, este necesar ca în fraza care urmează să apară cel puțin două cuvinte urmate de

notarea unei rime, așa cum avem în exemplul de mai sus. Adnotarea rimei se compune din 2 părți, Rhyme=ID... urmat de una sau mai multe cifre, care reprezintă id-urile cu care rimează cuvântul la finele căruia se află notația, și Type=. Trebuie specificat că nu pot apărea mai multe cifre decât în cazul Type=Mono sau Type=Rep. (Rar Type=Imperf). Pentru a extrage cuvinte între care avem rimă, computerul va selecta cele care au ID-ul notat la Rhyme=ID..., împreună cu cuvântul în dreptul căruia se află notația și le va lista.

După cum se arată în Constantinescu 1973, specifică folclorului este monorima. Mai rar întâlnită în poezia populară veche este rima împerecheată (Type=Paired). Cum aceasta se realizează adeseori eufonic, fără repetarea identică a sunetelor, alternând vocale diferite *meu-tău* sau consoane dure cu perechile surde, sau două consoane lichide diferite, am adnotat aceste situații cu Type=Imperf. Alte situații nu erau de așteptat. Și totuși, în corpusul din Republica Moldova întâlnim situații izolate, în două locuri rima îmbrățișată (Type=Embr). Exemplu:

„Casă mândră am gătat
Cu dulceața grâului,
Cu lacrima vinului,
Cu-n colac de grâu curat”.

În patru situații tot în corpusul din Republica Moldova, am găsit rima încrucișată (Type=Cross). Exemplu:

„Sub o țărnă de pământ
Este-un bujor înflorit
Toată lumea l-a urât
Dumneata l-ai îndrăgit”.

Ambele sunt specifice poeziei culte și probabil se explică prin faptul că în epoca contemporană poeții populari au acces la poezia cultă, ceea ce duce încet-încet la dispariția fenomenului folcloric autentic.

O altă situație pe care am întâlnit-o este a rimei cuvintelor cu cratimă. Cratima are rolul de a uni două silabe într-una singură, ceea ce face posibil ca accentul să fie de o parte a cratimei și sunetele care coincid de alta. Noi însă avem în ambele convenții de adnotare aceste cuvinte scrise cu *id*-uri separate. Prin urmare, am notat rima la cuvântul mai lung, în acest caz la verb. Dacă ar fi notată la ambele părți de vorbire, ar crește artificial numărul de rime, iar dacă ar fi notată la vocala *-o*, ar rezulta că este o similitudine de un singur sunet. Exemplu:

„Că mama care-am av*ut-o*
ai vând*ut-o*
și-ai bău*t-o*”.

Nu am putut adnota nici rimele dintre fraze, deoarece în dependency frazele sunt structuri independente și chiar dacă am nota *id*-ul frazei anterioare, acesta se poate schimba prin restrucutrarea corpusului; de pildă cele 1 200 fraze pe care le-am depus la UD în noiembrie au acum cu totul alte *id*-uri în corpusul actual de 6 300 de fraze. Exemplu:

„Lele Floare, da ești s*urdă*,
Îi duc brânză, nu-i duc *urdă*”.

Avem aici două fraze care în mod vizibil au rimă, dar nu o putem adnota în această convenție și în acest format.

Iată un fragment din lista rimelor extrasă automat din corpusul Republicii Moldova:

sentence 5405 Rhyme=Yes
Rhyme multiple: ID:6,9,13,21 Cuvinte: țară subsuoară grămăjoară școală; Description: Type=Mono; Distance= 3 4 8; Total distance: 15
Rhyme: ID:25,30 words: învățăm intrăm; description: Type=Paired distance=5;
total words: 31
sentence 5406 Rhyme=Yes
Rhyme: ID:3,10 words: învăța uita; description: Type=Paired distance=7;
Rhyme: ID:35,42 words: curcuță luncuță; description: Type=Paired distance=7;
total words: 43
sentence 5407 Rhyme=Yes
Rhyme: ID:6,14 words: tindă grindă; description: Type=Paired distance=8;
Rhyme: ID:21,29 words: casă masă; description: Type=Paired distance=8;
Rhyme multiple: ID:35,40,46,51 Cuvinte: nuci dulci atunci slugi; Description: Type=Mono; Distance= 5 6 5; Total distance: 16
total words: 52
sentence 5416 Rhyme=Yes
Rhyme multiple: ID:14,27,52,65 Cuvinte: aur aur aur aur; Description: Type=Rep; Distance= 13 25 13; Total distance: 51 total words: 66

Pentru a observa dacă rimele acestor versuri au un caracter regulat, comparăm distanța între rime, care trebuie să fie aproximativ egală (4-5 cuvinte), apoi comparăm distanța totală între rime cu lungimea totală a frazei. Dacă distanța totală este mult mai mică decât lungimea frazei, rezultă că fraza are o porțiune fără rimă. În exemplul de mai sus, este cazul primului vers. În fraza numărul 5 416, numărul de cuvinte este foarte mare, iar rima, de fapt o repetiție a cuvântului *aur*, este prezentă în prea puține locuri, la distanțe prea mari. Pe astfel de argumente ne bazăm atunci când afirmăm că rima este mai puțin regulată în corpusul Republicii Moldova, care este mai nou. Uneori fragmente de povestire în proză pot alterna cu fragmente versificate.

În ambele corpusuri comparate, statistica arată că predomină rimele *Paired*, împerecheate, și nu rimele multiple (de mai mult de două cuvinte, care pot fi de tipul Mono, Rep sau Imperf). Este încă o dovadă a disoluției modelului tradițional, care presupunea predominarea monorimei.

În corpusul R. Moldova avem în frazele din ambele culegeri:

902 sentences of which 754 with rhyme; Rhyme total: 3002 with 5123 words; 2144 Paired and 858 Multiple.

În corpusul de folclor din România, care este mai mic, avem în total:

230 sentences of which 187 with rhyme; Rhyme total: 577 with 1027 words; 450 Paired and 127 Multiple.

4. Adnotarea automată a rimei – Proiect

Modul în care am adnotat rima (manual) și mai ales în care am calculat distanța între rime este deocamdată unul aproximativ. Ar fi trebui să calculăm numărul de picioare metrice din versurile populare, care trebuie să fie de trei sau patru picioare metrice bisilabice cu accentul pe prima silabă. Considerând că numărul mediu de silabe dintr-un cuvânt în poezia populară este doi, am aproximat că patru cuvinte ar reprezenta patru picioare metrice. Am dorit să demarăm acest tip de cercetare cu mijloacele de care am dispus. Iată analiza corectă a metricii populare:

Tabelul 1. Analiză metrică a unui vers trohaic.

Ieși	*a-*	*fa-*	*ră*	*soa-*	*ră*	*ma-*	*re*
/	_	/	_	/	_	/	_
Că-ți	*a-*	*duc*	*piep-*	*tă-*	*nă-*	*toa-*	*re*
/	_	/	_	|	_	/	_

Am notat mai sus cu „/” o silabă accentuată și cu „_” o silabă neaccentuată. În cuvintele de mai mult de trei silabe se ia în considerație un accent secundar, pe care l-am notat aici cu „|”. Piciorul metric specific folclorului este troheul, care are schema următoare: „/ _”. Prin urmare, într-adevăr, versurile acestea au câte patru trohei, așa cum cere tiparul clasic folcloric, deși doar primul are patru cuvinte.

După cum vedem, adnotarea „SpaceAfter=No” ar putea fi la un moment dat utilă, pentru că linia de unire fără spații reduce de fapt numărul de două silabe, *că îți*, în una singură, *că-ți*, și astfel avem opt silabe, la fel ca în versul superior.

Pentru adnotarea automată a rimelor ar trebui deci să avem un program care desparte cuvintele în silabe și returnează numărul de silabe, nu de cuvinte, între două rime. Regulile de despărțire în silabe sunt simple. În primul rând, se încarcă în program o listă a literelor care transcriu sunete vocalice și consonantice. Apoi se ia în considerație numărul de consoane dintre două vocale, dintre care cel puțin una trece la silaba următoare și cel mult una rămâne la silaba anterioară. Deci putem avea situațiile: V-CV, VC-CV, VC-CCV, VC-CCCV. Aceste situații au un număr redus de excepții care pot fi descrise sau exemplificate. Mai este nevoie și de o listă a cuvintelor cu hiat, care poate fi încărcată dintr-un lexicon.

Apoi, în funcție de litera majusculă de la inițiala versului și de distanța egală de silabe, se caută rimele și se găsesc similitudini parțiale la finalul cuvintelor. Tipurile de rime pot fi și ele integrate în program sub formă de reguli.

Mai dificilă este problema rimei Type=Imperf, care poate fi foarte diversă. Credem că problema poate fi rezolvată prin încărcarea în program a unui număr cât mai mare de exemple și prin căutarea acestora în locul indicat, înainte de litera majusculă și după numărul obișnuit de silabe al versurilor (care în alte texte poate fi 4 sau 6). Totuși, s-ar putea ca programul automat să nu depisteze unele cazuri de rimă imperfectă.

Regula privind numărul de silabe și inițiala de vers trebuie suspendată pentru Type=Rep, care, așa cum a fost definit mai sus, nu se află neapărat la finalul versurilor și Type=Intern, care se definește tocmai prin rima unor cuvinte dintre care cel mult unul se află la finalul versului.

Ceea ce trebuie să menționăm este că un astfel de program nu poate fi conceput înainte de a fi exersat în practică adnotarea rimei și de a fi stabilit regulile acesteia. Extragerea statisticilor pe care am efectuat-o a avut și ea un rol important în stabilirea regulilor pentru depistarea rimelor. Un alt rol important a fost acela de a ajuta la depistarea erorilor produse prin adnotarea manuală sau a inconsistențelor de definire (la un moment dat denumean cu Type=Imperf doar una dintre rimele unui grup de rime, iar acum le denumim astfel pe toate din grup, deoarece nu putem stabili care este cea imperfectă în raport cu cealaltă sau celelalte).

Tematica

În ambele corpusuri avem exemple din aproape toate speciile folclorice: balade (narațiuni versificate), cântece lirice de dragoste sau de înstrăinare, descântece, orații de nuntă, colinde de Anul Nou, ghicitori, cântece de înmormântare. Poezie cu teme politice găsim numai în culegerea lui Grigore Bostan. Acestea conțin detalii impresionante despre copiii care mor în vagoanele unde sunt transportați.

Cântecele lirice conțin adeseori imagini originale de un mare rafinament:

„Flori albastre, flori adânci
Când le vezi, începi să plângi”.
„Pe pământ vor crește flori
Pentru-ale mele oițe surori”.

În colecția de texte subsumate, în colecția din Republica Moldova, *baladelor,* se regăsesc teme tradiționale românești, precum cele din *Miorița* și *Meșterul Manole,* parte versificate, parte narate, dar și fapte diverse impresionante: fata care s-a înecat, fata care a murit la șezătoare sau cea moartă cu o zi înainte de nunta ei, flăcăul cel mai frumos care moare otrăvit de șapte fete rivale, bărbatul care se întoarce din oastea împăratului la nunta nevestei care crede că el a murit (din care este extras exemplul dezvoltat de mai sus). Finalul dramatic este specific într-adevăr baladei. Caracterul impresionant este realizat prin introducerea unor motive de bocet sau de cântec de înmormântare, prin utilizarea conversației imaginare cu cel decedat și a elementelor ceremonialului de înmormântare.

Nu lipsesc temele actuale, fata a cărei mână a fost tăiată de mașina de treierat, aceasta fiind contaminată cu motivul tradițional al blestemului părintesc. În culegerea lui Gr. Bostan găsim un jurnal versificat al unei fete care este deportată în Kazahstan, de unde fuge după moartea părinților și se întoarce acasă, unde se mărită în alt sat, pentru a-și schimba numele și a-și pierde urma.

Nu poate fi însă vorba nicidecum de realismul basmelor și de dispariția superstițiilor, cum se afirmă în prefața la Botezatu *et al*. 1973. După cum știm, în basmele din folclor se creează o lume fictivă fabuloasă, configurată prin elemente

opuse lumii reale. Modul de configurare în folclorul românesc este unul temporal: „A fost odată, pe când se băteau urșii în coade” (în lumea reală ursul nu are coadă), „pe când se pupau lupii cu mieii” etc.

În folclorul rusesc lumea fabuloasă este delimitată spațial. Prin urmare, în mod firesc la folclorul din R. Moldova vom întâlni clișee cu determinări locale de demarcare a fabulosului. „Se duse până ce a trecut de țara Frigului”, și „a ajuns la un loc”. Motivele fabuloase sunt preluate și în orațiile de nuntă, unde viitoarea mireasă este comparată cu o zână, o stea, o floare, o căprioară, uneori, în mod incoerent, cu toate amestecat:

„La această casă Mândră și frumoasă Steaua ni se lasă, Urma fiarei intră-n casă”.

Mirelui i se atribuie origine împărătească, el pleacă la vânătoare însoțit de mulți oșteni și străbate teritorii întinse, trecând pe la Țaringrad. Ca și în basme, în orațiile de nuntă și în unele colinde valoarea de adevăr este suspendată, nimeni nu crede în adevărul acestor narațiuni care au caracter simbolic, estetic. Șirul de elemente reale negate (puricele potcovit cu șapte ocale de fier) reprezintă mărci de intrare în lumea fabuloasă, iar afirmația povestitorului, cum că nu crede, plasată la finalul poveștii nu exprimă realismul, ci are caracter convențional, este tot o marcă, de astă dată a faptului că se iese din lumea fabuloasă și se reintră în cea reală: „Am încălecat pe o căpșună și am spus o mare minciună”.

Motive fabuloase de negare a realității sunt folosite și în lirica politică:

„Când a face plopul nuci,
Și răchita mere dulci,
Tocma atunci și nici atunci
Vom fi noi nemților slugi”.

În ceea ce privește lirica descântecelor, aceasta se îmbină cu credința religioasă; de obicei descântecele se termină cu:

„Descântecul de la mine,
Leacul de la Maica Domnului”.

Alteori aceasta este invocată ca un simbol al sănătății și al perfecțiunii ce i se dorește celui descântat, întocmai ca în tradiția folclorică de dincoace de Prut:

„Să rămâie Cutare curat, luminat,
Ca maica Domnului care l-a dat”.

Există în descântece și elemente mitologice, pasărea de fier cu puii de fier, omul negru, lupul, ursul cu forțe malefice, o bătrână cu un ochi de foc și unul de apă. Fata căreia i se face descântecul de iubire încalecă pe doi cuci. Nu avem aici o dispariție a superstițiilor, ci o afirmare a lor, în sensul că autorul descântecului modelează o realitate spre care speră că pacientul va evolua, deci speră că realitatea dată, inițială, va putea fi modificată în conformitate cu modelul.

Ceea ce apare și în colinde și în descântece ca marcă a slăbirii funcției rituale este umorul. Acesta este mai frecvent în folclorul din Republica Moldova, poate fiindcă e mai nou, sau poate fiindcă nu avem un corpus de folclor din România destul de dezvoltat.

Un colind ca cel de mai jos desființează clar potențialul de urare benefică pentru casa colindată. El reprezintă o ironie la adresa celor care pleacă la colindat fără a cunoaște texte tradiționale lungi și se grăbesc să primească răsplata:

> „Plugușorul cu tărâță Cu doisprezece pui de mâță, Unul scurt și unul lung, Dați-mi rubla că mă duc! Hăi!"

Ironia mai apare și în orațiile de nuntă, cu referire la raporturile între mireasă și soacră:

> „Ți-am adus o noră tare, Să-ți ajute la ciubăr, Și la scărmănat de păr".

Prezentarea cu umor a relațiilor de familie și a defectelor omenești apare și în folclorul românesc, mai ales când este vorba despre strigături. Totuși remarcăm bogăția formelor de umor în folclorul din Republica Moldova.

Alte particularități

Corpusul din Moldova și Bucovina conține numeroase auxiliare cu forme abreviate, ambiguie: „a" sau „o" pot însemna „va", „ar", „a", viitor, condițional sau perfect compus; trebuie decodate în funcție de alte verbe din context. POS-taggerul le adnotează eronat ca fiind particula infinitivului sau articol nedefinit. Este vorba de o transcriere mai fidelă a pronunției regionale sau orale nonstandard. Trebuie totuși ca POS-taggerul să fie antrenat pe astfel de texte pentru a putea procesa limbaj nonstandard.

Există puține elemente lexicale în folclorul din Republica Moldova care nu se întâlnesc în România. Spre exemplu: doi *logani* (arbori), *locai* (îngrijitori de cai), fântână *stoborâtă*, *ohoti*, *dojesc*, *jugnesc*, *mișulezi*, *zacon*, *smină*, *maladeț*, *copică*, *rublă daraft*, *beșlici*. Unele cuvinte sunt pronunții sau cuvinte regionale similare cu cele din regiunea Moldovei de la Vest de Prut: *chicurat*, *ista*, *bortă*, *curechi*, *chiroane*.

Sunt frecvent folosite denumiri de monezi care au caracter istoric și geografic.

Așa cum se arată în Cojocaru 2016, în conversația nonstandard a moldovenilor care fac parte dintr-o populație bilingvă apar cuvinte rusești; dar se renunță la acestea de îndată ce ia parte la conversație cineva care nu știe limba rusă, ceea ce demonstrează că aceste cuvinte nu sunt intrate în limbă, ci sunt interferențe ce țin de convenția situației de discurs. Uneori, studenții moldoveni din Iași folosesc astfel de cuvite de neînțeles pentru colegii lor ca să se distreze sau ca să comunice între ei codificat, ca într-un fel de argou.

Ceea ce prezintă adeseori deosebiri este topica, ordinea cuvintelor în comunicare. De aceea credem că este important că avem un corpus analizat sintactic pentru a face comparații, iar acestea vor fi mai interesante când corpusul va crește. De exemplu, libertatea topicii, bine ilustrată în corpusul din Republica

Moldova, este specifică limbii române vechi și populare. Aceste elemente de sintaxă ar putea fi mult mai bine studiate având la dispoziție corpusuri mult mai extinse, comparabile și prin dimensiune, și prin conținut. Dăm un singur astfel de exemplu, în care complementul direct neaccentuat *ne-* este un argument avansat înaintea auxiliarului *au* al verbului regent al regentului său real (*bate*):

„Când le-am spus că n-au dreptate
Ei ne-au început a bate”.

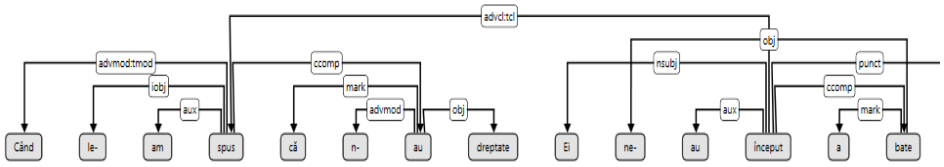

Figura 3. Frază din folclorul R. Moldova cu argument (obiect direct) mult avansat.

Atunci când fraza începe cu un circumstanțial, în corpusul de folclor românesc acesta este unul de timp, pe când în corpusul moldovenesc este de preferință local.

Ceea ce pare asemenea la prima vedere în cele trei corpusuri analizate este prezența obligatorie a invocării elementului vegetal, dar nu am efectuat un studiu aprofundat pentru a observa elemente distinctive. Sigur este că elementul vegetal selectat are nu doar funcția de a rima cu elementul central al poeziei, ci și valențe simbolice; de pildă, pelinul, care este amar, este invocat în texte care descriu starea de suferință a emitentului, exprimând o hiperbolizare a suferinței. Alte plante pot avea diverse alte simboluri care trebuie descifrate sau sunt exprimate direct:

„Cunună de brebenoc
Ca să fie de noroc”.

Concluzii

Pentru a continua acest studiu, este necesar să mărim dimensiunea celor două corpusuri, mai ales al celui din România, și să ne orientăm și în România asupra folclorului cules în deceniile 6 şi 7 ale secolului trecut, pentru a vedea în ce măsură evoluția tipurilor tradiționale spre poezia cultă și disoluția tradiției și a ritualului în umor este comparabilă. Nu știm în ce măsură vom putea să o facem, deoarece interesul pentru folclor a scăzut în ultima vreme în România, invers proporțional cu cel din Republica Moldova, și folclorul nou este mai puțin antologat decât cel vechi.

Tiparele sintactice cu topică liberă trebuie de asemenea studiate cu mai multe exemple.

Pentru a continua analiza rimei la altă scară, vom încerca să facem un program automat de recunoaștere și adnotare a rimei. Nu există, deocamdată, studii privind analiza formală a textului poetic; avem date doar despre un singur demers, dedicat studiului rimei în poezia cultă germană. În ce privește poezia populară, nu am găsit corpusuri în format electronic și credem că este urgent ca acestea să fie create, fiindcă, după dispariția acestui fenomen, nu vom mai avea posibilitatea de a-l studia și conserva.

Este important ca studiul folclorului și studiul rimei să se bazeze pe un corpus aflat în UD, care este o platformă de mare vizibilitate internațională, având în vedere că, astfel, putem atrage atenția altor cercetători spre această tematică.

SURSE

Amzulescu 1964 = Al. I. Amzulescu, *Balade populare româneşti,* Bucureşti, Editura pentru Literatură, 1964.

Bălășel 1967 = Teodor Bălășel, *Folclor din Oltenia şi Muntenia. Texte alese din colecţii inedite,* Vol. II, București, Editura pentru Literatură, 1967.

Bostan 1998 = Grigore C. Bostan, *Poezia populară românească în spaţiul carpato-nistrean,* Iaşi, Editura Cantes, 1998.

Botezatu *et al.* 1973 = G. Botezatu, V.A. Cirimpei și I.D. Ciobanu, *Folclor din părțile Codrilor*, Chișinău, Editura Știința, 1973.

Brăiloiu *et al.* 1978 = C. Brăiloiu, Emilia Comişel şi Tatiana Găluşcă-Cârşmariu, *Folclor din Dobrogea,* Bucureşti, Editura Minerva, 1978.

Cireş, Berdan 1982 = Lucia Cireş, Lucia Berdan, *Descântece din Moldova. Texte inedite,* în „Caietele arhivei de folclor", nr. 1, Iași, 1982.

Covalciuc, Bostan 1993 = D. Covalciuc, Gr. Bostan, *Folclor din Ţara Fagilor*, Chişinău, Editura Hyperion, 1993.

Datcu 1968 = Iordan Datcu, *Folclor din Oltenia şi Muntenia. Texte alese din colecţii inedite,* vol. III, București, Editura pentru Literatură, 1968.

BIBLIOGRAFIE

Bobicev *et al.* 2016 = Victoria Bobicev, Tudor Bumbu, Victoria Lazu, Victoria Maxim, and Daniela Istrati, *Folk poetry for computers: Moldovan Codri's ballads parsing* in *Proceedings of the 12th International Conference Linguistic Resources and Tools for Processing the Romanian Language*, ConsILR, 2016, p. 39–50.

Bobicev *et al.* 2017 = Victoria Bobicev, Cătălina Mărănduc, and Cenel-Augusto Perez, *Tools for Building a Corpus to Study the Historical and Geographical Variation of the Romanian Language*, in *Proceeding of the First Workshop on Language technology for Digital Humanities in Central and (South-) Eastern Europe* (LTDH4CSEE 2017). *Varna, Bulgaria,September 8, 2017*, Shoumen, Incoma Ltd., 2017, p.10–20.

Cojocaru *et al.* 2017 = Svetlana Cojocaru, Al. Colesnicov, Ludmila Malahov, *Digitization of Old Romanian Texts Printed in the Cyrillic Script*, in *Proceedings of Second International Conference on Digital Access of Textual Cultural Heritage (DATeCH 2017). Gōttingen 1–2 June 2017*, New York, ACM, 2017, p. 143–148.

Cojocaru 2016 = Valentina Cojocaru, *Marcatori discursivi în limba română vorbită în Republica Moldova: aspecte pragmatice și sociolingvistice*. Teză de doctorat, Facultatea de Litere, Universitatea din București, 2016.

Colhon *et al.* 2017 = Mihaela Colhon, Cătălina Mărănduc, Cătălin Mititelu, *A Multiform Balanced Dependency Treebank for Romanian*, in *Proceedings of Knowledge Resources for the Socio-Economic Sciences and Humanities, (KnowRSH). Varna, Bulgaria September 7, 2017*, Shouma, Incoma Ltd, 2017, p. 9–18.

Constantinescu 1973 = Nicolae Constantinescu, *Rima în poezia populară românească,* București, Editura Minerva, 1973.

DOOM^2 = Ioana Vintilă-Rădulescu, (ed.) *Dicționarul ortografic, ortoepic și morfologic al limbii române.* Ediția a doua, București, Editura Univers Enciclopedic, 2005.

Mărănduc, Bobicev 2017 = Cătălina Mărănduc, Victoria Bobicev, *Non Standard Treebank Romania – Republic of Moldova in the Universal Dependencies*, in *Proceedings of Conference on Mathematical Foundations of Informatics* (MFOI-2017) November 9–11, 2017, Chișinău, Moldova, p. 111–116.

Mărănduc *et al.* 2018 = Cătălina Mărănduc, Cătălin Mititelu, Victoria Bobicev, *Syntactic Semantic Correspondence in Dependency Grammar*, in *Proceedings of 16th International Workshop on Treebanks and Linguistic Theories*, Prague, Jan. 23–24, 2018, p.167–180.

Perez 2014 = Cenel-Augusto Perez, *Resurse lingvistice pentru procesarea limbajului natural*. Teză de doctorat, Iași, Facultatea de Filologie, Universitatea „Alexandru Ioan Cuza”, 2014.