Premise pentru o tehnologie de recunoaştere automată a vorbirii în limba română aplicată domeniului juridic

Melania Duma[1], Corina Giurgea[2], Mihaela Ordean[2,] Paula Zălhan[2]

[1]Universitatea Babes-Bolyai din Cluj-Napoca – UBB
Str. M. Kogălniceanu nr. 1, Cluj-Napoca, 060042
E-mail: duma_melania@yahoo.com

[2]iQuest Technologies Cluj-Napoca – iQuest
Str.Motilor nr.4-6, Cluj-Napoca, 40001
E-mail: corina.giurgea@iquestgroup.com; mihaela.ordean@iquestgroup.com; paula.zalhan@iquestgroup.com;

Rezumat. Gândit pentru a veni în întâmpinarea unui obiectiv din strategia IT a Ministerului Român de Justiție, proiectul JustASR, in curs de derulare, urmăreşte concepția şi proiectarea elementelor tehnice necesare implementării unui sistem de interacțiune om-calculator bazat pe detectarea automată a limbajului pentru limba română având un scop restricționat la un domeniu particular. Proiectul îşi propune să dezvolte nucleul unei tehnologii de recunoaştere automată a vorbirii cu aplicabilitate directă în sălile de judecată ale României. Articolul de față se doreste a fi o prezentare succintă a premiselor şi a unor elemente lingvistice de natură fonetică care țin de concepția proiectului. După o introducere în problematica tehnologiei de recunoaştere automată a vorbirii, articolul conține o trecere în revistă a eforturilor naționale din domeniul recunoaşterii vorbirii. Sunt prezentate în continuare provocările specifice limbii române care constituie punctul de pornire al efortului ştiințific şi tehnic al tehnologiei care face obiectul proiectului JustASR. Cea mai consistentă secțiune a articolului, conținând partea aplicată a principiilor teoretice, este cea dedicată aspectelor fonetice care țin de realizarea dicționarului care va fi utilizat în proiect. Sunt prezentate criteriile generale utilizate pentru selecția cuvintelor care vor fi incluse în dicționar, precum şi categoriile de cuvinte din care s-a făcut extragerea.

Cuvinte cheie: recunoaşterea vorbirii, corpus, dicționar, procese fonetice.

1.Introducere

Vorbirea, cea mai naturală formă de comunicare între oameni, a reprezentat încă de acum 80 de ani o provocare în proiectarea şi construirea unei maşini

care să mimeze comportamentul uman sub aspectul capacității de a vorbi natural şi de a răspunde corect limbii vorbite conform Rabiner & Huang (2005).

Un interes major pentru studiul interacțiunii om-calculator prin intermediul limbajului a manifestat şi manifestă nu doar comunitatea ştiințifică, ci şi mediul de afaceri. În condițiile fuziunii dintre sistemele de calcul şi cele de comunicații, tehnologiile de procesare a limbajului natural capătă o importanță deosebită.

Mai mult, astăzi, aceste tehnologii sunt chemate să soluționeze problemele societății contemporane. Astfel, o serie de evenimente de tipul intervențiilor armate sau al calamităților naturale din ultimii ani au reliefat greutățile misiunilor umanitare sau a celor de salvare datorită necunoaşterii de către participanții la aceste misiuni a limbilor vorbite de populația afectată.

Pe de altă parte, se estimează că, până la sfârşitul acestui secol, peste o jumătate din limbile existente vor dispărea (Besacier et al, 2014), existând un risc de diminuare a diversității lingvistico-culturale a omenirii.

Tehnologiile de procesare a limbajului uman pot şi trebuie să aducă un aport major în soluționarea problemelor creionate mai sus prin: îmbunătățirea utilitarelor de procesare a textului, realizarea de dicționare electronice, continua dezvoltare de sisteme avansate de procesare de tip TTS (eng. *Text to Speech*) sau SR (eng. *Speech Recognition*). Spre exemplificare, sistemele de traducere automată pot constitui o alternativă la translatorii umani, insuficienți la nivelul impus de evenimente de tipul celor amintite mai sus în timp ce sistemele de recunoaştere automată a vorbirii ASR (eng. *Automatic Speech Recognition*) pot contribui la salvarea unor limbi amenințate cu dispariția (cele mai multe vorbite şi nu scrise) prin transcrierea acestora.

În mare, tehnologia vorbirii în cadrul interacțiunii om-calculator presupune una sau ambele din următoarele: (i) o tehnologie care sintetizează automat vorbirea având ca intrare un text şi (ii) o tehnologie care recunoaşte automat limbajul. Recunoaşterea automată a vorbirii are ca scop determinarea, cu o cât mai mare exactitate a şirului de cuvinte pronunțat de un vorbitor şi este utilă în aplicații de tip comandă-şi-control, dictare automată.

Cele două tehnologii prezentate mai sus stau la baza dezvoltării de sisteme avansate de procesare de tip TTS respectiv ASR, integrate într-o măsură tot mai mare în produse comerciale (în special în produse software)

destinate unor domenii precum telecomunicațiile, centrele de informare (call-center-e) automatizate, învățarea electronică (e-learning), sau medicină şi justiție.

Tehnologia ASR este matură pentru limbile răspândite în lume: engleza, chineza, spaniola sau franceza. Marea majoritate a limbilor din țări în curs de dezvoltare au fost, până nu de mult, sub un con de umbră. Mai mult, studiile din cadrul unor proiecte ample în care au fost implicate rețele de excelență şi/sau asociații profesionale (Besacier et al, 2014) au reliefat faptul că o serie de limbi cu un puternic potențial economic, multe dintre ele aflate în topul celor mai vorbite 20 de limbi ale lumii, intră în categoria aşa numitelor limbi cu deficit de resurse (Cucu et al, 2014).

Termenul de **limbă cu deficit de resurse** (eng. *under-resourced language*) a fost introdus de Krauwer în anul 2003 şi de Berment în 2004 (Besacier et al, 2014) pentru a desemna o limbă caracterizată prin unele, dacă nu prin toate aspectele care urmează: absența unui sistem de scriere unic sau cu ortografie aplicată consistent, prezența limitată pe web, lipsa de expertiză lingvistică, lipsa de resurse pentru procesarea limbajului natural (corpusuri monolingve, dicționare electronice bilingve, dicționare pentru pronunție, corpusuri adnotate fonetic/sintactic).

Cercetările efectuate în acest domeniu trebuie să răspundă la două mari provocări. Prima dintre acestea vizează rezolvarea unor probleme specifice limbii cu resurse limitate, probleme care nu au apărut în cazul limbilor cu resurse lingvistice dezvoltate. De exemplu, sistemele fonetice asociate unor limbi cu resurse limitate şi dificultatea de a crea dicționare pentru pronunție constituie una dintre problemele care trebuie rezolvată în vederea dezvoltării de sisteme ASR sau TTS. A doua mare provocare o constituie optimizarea costurilor ataşate dezvoltării unor sisteme de procesare a limbajului, costuri, de multe ori, mari.

În ultimii 7 ani, echipele de cercetare multidisciplinare s-au focalizat pe crearea de resurse lingvistice scrise şi acustice, pe portarea tehnologiilor, pe dezvoltarea de algoritmi şi metode pentru adaptarea modelelor acustice şi lingvistice pentru limbi cu deficit de resurse. Începând cu 2008 se organizează bianual un workshop internațional, *Spoken Language Technologies for Under-resourced languages*, iar revista *Speech Communication* a avut în 2014 un număr special (56/2014) dedicat procesării limbilor cu resurse limitate (Besacier et al, 2014 b).

Cu toate că este limbă oficială a Uniunii Europene, limba română este considerată ca făcând parte din categoria limbilor cu resurse limitate. Încadrarea în această categorie s-a făcut pe baza rezultatelor studiilor efectuate în cadrul rețelei de excelență METANET. În cadrul acestei rețele reunind 60 de centre de cercetare din peste 34 de țări, s-au întocmit, pe baza unor criterii unitare, o serie de studii care analizează resursele şi tehnologiile lingvistice disponibile pentru 30 de limbi europene şi prezintă situația sprijinului acordat tehnologiilor limbajului. Rezultatelor studiilor privind limba română le-a fost dedicată o monografie ce se constituie într-un volum distinct al seriei (Trandabăț et al, 2012).

După ştiința noastră, până la această oră nu există pe piața romanească o soluție comercială de recunoaştere automată a vorbirii dezvoltată în România. Trebuie menționat aici ca există pe piață implementări care utilizează recunoaşterea vorbirii fară ca acestea să menționeze însă aplicațiile comerciale sursă utilizate.

Când vine vorba de limba română, majoritatea companiilor care dezvoltă produse comerciale care integrează o soluție de recunoaştere automată a vorbirii cer o bază de date cu înregistrări pe baza cărora implementările sunt personalizate. Un exemplu de companie care a lansat pe piață o aplicție comercială de recunoastere a vorbirii in limba română este Nuance Communications, Inc. Conform unui comunicat de presă postat pe situl Nuance (http://www.nuance.com), compania a lansat in anul 2012 aplicația comerciala Dragon Dictation pentru iPhone, iPad si iPod. Aplicația utilizeaza recunoaşterea vorbirii în vederea dictării rapide a mesajelor de tip SMS şi poştă electronică, a informației care se doreşte a fi actualizată pe rețele sociale sau obținută prin căutare pe web.

În anul 2010, în vederea îndeplinirii condițiilor legale şi a recomandărilor Comisiei Europene privind transparența actului public de justiție, Ministerul de Justiție din Romania a organizat o licitație publică urmărind achiziționarea unei soluții de recunoaştere a vorbirii pentru sălile de judecată, oferind peste 2.5 Mil EUR. Pe piața din România, nici un furnizor de software nu a fost capabil să livreze un astfel de produs. În urma acestui „eşec" în strategia IT a ministerului pentru perioada 2013-2017, publicată în aprilie 2013, Ministerul de Justiție reiterează interesul pentru un produs care să realizeze transcrierea automată a discuțiilor din sălile de judecată.

În acest context, în dorința de a răspunde provocării şi totodată oportunității comerciale descrise mai sus, a fost gândit proiectul JustASR. Proiectul, aflat in curs de derulare in cadrul Programului National de

Cercetare – Parteneriate in Domenii Prioritare, are ca obiective concepția şi proiectarea elementelor tehnice necesare implementării unui sistem de interacțiune om-calculator bazat pe detectarea automată a limbajului pentru limba română, având un scop restricționat la un domeniu particular: realizarea unei transcrieri automate , de calitate, pentru audierile publice din sălile de judecată din România.

Articolul de față nu îşi propune să intre în aspectele tehnice legate de dezvoltarea nucleului tehnologiei de recunoaştere a vorbirii în limba română aplicată domeniului juridic, care face obiectul proiectului JustASR. Aceste aspecte vor constitui subiectul unei alte publicații propuse de Sofroni & Stan (2015).

Prezentul articol investighează o serie de aspecte practice, de natură fonetică, dar necesare în realizarea unui dicționar, respectiv a unui corpus de atrenament pentru un sistem de recunoaştere automată a vorbirii pentru limba română.

2. Sinoptic privind cercetările efectuate la nivel național în vederea dezvoltării unui sistem ASR pentru limba română

Deşi, la această oră, nu este încă disponibil pe piața românească un produs dezvoltat in România care să integreze tehnologia de recunoaştere a vorbirii pentru limba română, demersuri în acest sens au fost întreprinse de mai multe grupuri de cercetare din România, cu rezultate demne de luat în calcul.

În teza lui de doctorat, Boldea (2003) a colectat, pentru prima dată, o bază de date fonetice, acoperind direct sau prin combinații, fonemele limbii literare române. El a dezvoltat un algoritm pentru identificarea limitelor unităților de semnal şi a efectuat adnotarea majorității semnalelor înregistrate. Se cuvine a preciza aici că fonemul reprezintă unitatea minimală a nivelului fonetic ce individualizează un sunet articulat şi care e capabilă să diferențieze sensul.

La Universitatea Tehnică din Cluj-Napoca, grupul de cercetare coordonat de către Prof. G. Toderean a dezvoltat un sistem bazat pe Modele Markov Ascunse Discrete (eng. *Hidden Markov Models*) pentru recunoaşterea cuvintelor izolate descris în Toderean & Buza (2008). În teza sa de doctorat, elaborată sub îndrumarea Prof. G. Toderean, Margit Antal (2006) a

investigat recunoşterea vorbirii şi a vorbitorului pe baza fonemelor şi a propus modele pentru recunoaşterea şi clasificarea fonemelor.

Grupul Prof. H. N. Teodorescu de la Universitatea Tehnică Iaşi, a dezvoltat „Sunetele Limbii Române" (SroL) conţinând 1500 de înregistrări distincte cu diferite formate de precizie şi codificări descrise în (Teodorescu et al, 2010). În privinţa SroL, lucrarea Teodorescu (2010) furnizează exemple despre modul în care se pot extrage elemente relevante din semnale destinate recunoaşterii limbajului natural.

O largă contribuţie în procesarea limbajului natural este oferită la Universitatea Politehnică Bucureşti sub îndrumarea Prof. Corneliu Burileanu, printr-o muncă cumulată de-a lungul anilor. Lucrarea (Popescu et al, 2008) prezintă o metodă computaţională simplă, pentru urmărirea vorbitorului în dialogurile multi-laterale, bazată pe traversarea unui arbore decizional construit în funcţie de o regresie liniară neghidată de tip Maximum Likelihood. Rezultatele arată o performanţă de 80% în procesul de urmărire a vorbitorului.

Menţionăm aici ca pentru evaluarea performanţei unui sistem ASR trebuie calculată rata de eroare (word-error-rate-WER), obţinută prin compararea textului vorbit cu cel recunoscut de către software. Această rată de eroare este calculată ca număr total de substituiri, inserări şi ştergeri de cuvinte necesare pentru a face cele două texte identice, raportat la numărul total de cuvinte. Acurateţea procesului ASR este definită ca opusul ratei de eroare, adică Acurateţea = 1-WER.

Grupul condus de Prof. Burileanu a început să avanseze către problematica recunoaşterii limbajului în vorbirea spontană în limba română în cadrul proiectului PNII IDEI nr. 114/2007. Numeroase articole au fost publicate în acest sens (Petrea et al, 2009), (Petrea et al, 2010). Utilizând Hidden Markov Modelling Toolkit (HTK) (HTK toolkit http://htk.eng.cam.ac.uk/), experimentele preliminare au arătat o performanţă de 70% acurateţe (Burileanu et al, 2010). Prin redefinirea metodei şi modelarea incertitudinii, prin detectarea de Modele Gaussiene Mixte (eng. *Gaussian Mixture Models*) şi introducerea log-probabilităţilor, ei au reuşit să înregistreze o performanţă de 18.5% WER, utilizând un corpus de 10000 de cuvinte distincte, pronunţate de numai 5 vorbitori.

Într-un articol recent (Cucu et al, 2014) se raportează construirea, de către echipa din cadrul laboratorului de cercetare SpeeD din Universitatea Politehnica din Bucureşti (UPB), a celui mai mare vocabular al limbii romane (peste 60000 cuvinte) destinat dezvoltării unui sistem ASR.

Articolul citat este centrat pe prezentarea metodologiei propuse de autori pentru portarea/adaptarea unui corpus asociat unui domeniu specific dintr-o limbă cu resurse dezvoltate (în cazul de față, un corpus asociat jurnalismului de ştiri în limba franceză) într-o limba cu resurse limitate (în cazul de față limba română), utilizând metode de traducere automată bazate pe statistica oferită de Maşini Statistice de Traducere sau SMT (eng. *Statistical Machine Translation*). În acelaşi articol, modelele pentru restaurarea diacriticelor limbii române, respectiv cel pentru conversia grafem-fonem, propuse, dezvoltate şi testate de autori, sunt evaluate şi comparate cu alte modele elaborate pentru limba română.

Studiul lui Schiopu (2010) propune o comparație între sistemele ASR dezvoltate la Cluj şi la Iaşi (prezentate mai sus), dar nu ia în considerare dezvoltările Prof. C. Burileanu.

Grupul condus de către Acad. Dan Tufiş de la Academia Română a desfăşurat o muncă de cercetare în direcția dezvoltării resurselor computaționale lingvistice pentru limba română. Merită menționate în acest sens: Proiectul European CLARIN (proiectul stabilieşte o infrastructură de cercetare formată din resurse şi tehnologii în scopul diminuării fragmentării curente şi oferirea unei infrastructuri stabile, persistente, accesibile şi extensibile), proiectul eDTLR realizat în cooperare cu grupul Prof. Dan Cristea de la Universitatea Al. Ioan Cuza din Iaşi – acest proiect a condus procesul de digitizare a Dicționarului Tezaur al limbii române eDTLR (2013) sau proiectul SEE – ERA.NET (construirea resurselor lingvistice şi a unor modele de traducere automatizată cu ajutorul calculatorului în limbile balcanice şi de origine sud-slavonă).

Toate aceste resurse merită luate în considerare, pentru că ele au reuşit să clarifice aspecte controversate şi variate ale modelării limbii române, reprezentând o bază reală pentru construirea unei tehnologii comerciale pentru recunoaşterea automată a limbajului natural. Ca o remarcă, se observă că aceste grupuri de cercetare au muncit izolate unele față de celelalte, fiecare utilizând propria bază de date cu semnale de voce înregistrate.

Până foarte recent, nici unul dintre grupurile de cercetare amintite mai sus, nu a reuşit să înregistreze şi să proceseze semnale vocale provenind de la un număr mare de vorbitori care utilizeaza cuvinte extras dintr-un vocabular de scară medie. Proiectul RASC - Romanian Anonymus Speech Corpus, derulat în cadrul Institutului de Cercetări pentru Inteligență

Artificială al Academiei Române, îşi propune ca obiectiv principal tocmai crearea unei colecţii mari de resurse care să poată fi utilizate în aplicaţii care implică prelucrarea limbajului vorbit. Colectarea corpusului de voce se face într-un mod inedit, prin „crowd-sourcing" conform (Dumitrescu et al, 2014).

Portalul RASC (http://rasc.racai.ro/) a fost deschis tuturor celor care doresc să se înregistreze citind propoziţii dintr-un set prestabilit, asigurându-se astfel atât variabilitatea condiţiilor de înregistrare reflectate în calitatea semnalului obţinut (medii cu sau fără zgomot de fundal, înregistrări cu sau fără microfon) cât şi variabilitatea în vorbire (înregistrări provenind de la persoane de sex diferit, de vârste diferite, provenind din diversele regiuni ale ţării). Se cunoaşte faptul că fiecare persoană are un tract vocal diferit şi că un vorbitor al unei limbi având caracteristici generale clare poate rosti una şi aceeaşi propoziţie cu mici variaţii. Toate aceste variaţii se vor transpune şi pot fi observate în reprezentările digitale ale semnalelor de vorbire. La ora redactării acestui articol, conform informaţiei disponibile pe site-ul proiectului, corpusul conţinea 3056 de propozitii înregistrate, avea o mărime 1060 MB şi era pus la dispoziţia comunităţii de cercetare (ca open-source) în vederea dezvoltării unor sisteme de recunoaştere a vorbirii continue în limba română.

3. Aspecte lingvistice ale unui proiect de recunoaştere a vorbirii în limba română

Particularităţile gramaticale distinctive ale limbii române în cadrul limbilor romanice şi deficitul de resurse în ceea ce priveşte utilizarea românei în contexte informatice determină o dublă provocare pentru orice demers ce presupune digitalizare şi aplicare de algoritmi.

Dezvoltarea unei tehnologii de recunoaşterea automată a vorbirii pentru limba română necesită, în primul rând, o înţelegere a particularităţilor fonetice şi morfologice ale limbii, iar, în al doilea rând o modelizare a regulilor sintactice derivate din acestea în funcţii.

3.1. Specificul limbii române

La nivel sintactic limba română se individualizează în cadrul limbilor romanice prin: nelexicalizarea subiectului prenominal, dublarea de către un

pronume personal a unui grup nominal lexical, dublarea cliticelor, concordanța negativă și negația dublă.

La nivel morfologic, limba română se evidențiază printr-un sistem flexionar bogat, în care substantivele, adjectivele şi pronumele au cinci cazuri şi două numere. Verbele au trei persoane atât la singular, cât şi la plural pentru cele cinci moduri personale, la care se adaugă modurile nepersonale: infinitivul, participiul şi gerunziul, în total un verb putând număra treizeci şi trei de forme. Adjectivele şi substantivele pot avea forme articulate şi nearticulate, un substantiv putând avea în total cinci forme, iar un adjectiv şase.

La nivel fonetic, flexiunile cuvintelor limbii române pot să aibă alternanțe în interiorul rădăcinii, la care se adaugă sufixe morfologice şi desinențe (Trandabăț et al, 2012).

O caracteristică a limbii române cu impact fonetic pentru ASR este folosirea diacriticelor: *ă*, *â*, *î*, *ş* şi *ţ*, care apar în aproape 30-40% dintre cuvintele dintr-un text (Cucu et al, 2014).

În cazul în care diacriticele lipsesc, mesajul transmis s-ar putea să fie neclar, aceasta determinând atât lizibilitate scăzută, cât şi ambiguitate. Deşi pe Internet sunt disponibile pentru descărcare diferite corpusuri româneşti de ştiri fără diacritice, acestea nu pot fi utilizate în dezvoltarea unui sistem ASR pentru limba română (Cucu et al, 2014). O adnotare ASR trebuie să conțină diacritice şi să elimine ambiguitatea cuvintelor care pot avea forme distincte cu şi fără diacritice.

Pentru transcrierea fonetică automată a unui vocabular, există în momentul de față două abordări posibile. Primul tip de abordări se bazează pe reguli de conversie grafem-fonem. Pentru aceste abordări contează numărul de reguli necesare în dezvoltarea sistemului ASR. Dacă pentru limba engleză, numărul acestora este foarte mare (în jur de 1500 de reguli), pentru limba română (care este o limba fonetică) numărul regulilor poate fi considerabil mai mic. Al doilea tip de abordări sunt cele bazate pe maşini de învățare. Sistemele care folosesc maşini de învățare se bazează, în principal, pe cunoştințe dobândite dintr-un set de învățare, care sunt mai apoi aplicate datelor din aplicația ASR. Aceste cunoştințe sunt de obicei exprimate prin arbori de decizie sau prin rețele neuronale.

3.2. Principii lingvistice în recunoaşterea automată a vorbirii

Recunoaşterea automată a vorbirii presupune interpretarea unui flux de intrare auditiv şi transpunerea sa în text. Realizarea unei interpretări cât mai fidele este condiționată de câteva etape premergătoare:

- Realizarea unui corpus de antrenament şi a unui corpus de testare;
- Realizarea unei parsări, adică a unei analize discrete a elementelor ce individualizează un anume nivel de limbă.

Analiza lingvistică implicată în recunoaşterea automată a vorbirii se concretizează în parsări de tip fonetic, morfologic şi sintactic. Articolul de față prezintă o parte din criteriile fonetice de care trebuie să se țină cont în evidențierea interacțiunii dintre nivelele limbii pentru un sistem de recunoaştere automată a vorbirii pentru limba română.

Dicționarul nu este doar o înşirare de cuvinte, ci un set de cuvinte ordonat dupa criterii fonetice, morfologice sau sintactice, aşa cum evidențiază şi Cristea (2005).

Criteriul fonetic, conform cu DOOM (2005) are la bază interacțiunea dintre unitățile sale minimale, fonemele. Un dicționar realizat pe astfel de principii va include cuvintele şi transcrierea lor fonetică. Selecția cuvintelor va avea în vedere acoperirea inventarului de foneme ale limbii române, mai precis a instanțelor acestor foneme ce poartă numele de alofone. Alofonele sunt rezultatul interacțiunii dintre sunet şi contextul său diagnostic.

Criteriul morfologic, în conformitate cu GALR (2008), are la bază interacțiunea dintre unitățile sale minimale, morfemele. Morfemele sunt de două tipuri: lexicale (partea stabilă a vocabularului) şi gramaticale (partea flectivă a vocabularului). Dicționarul va include partea lexicală a cuvântului şi variațiile sale în funcție de context: poziționare sau regulă sintactică.

Criteriul sintactic, corespunzător cu GALR (2008), are la bază interacțiunea dintre grupuri sintactice, iar acest fapt nu generează un dicționar în sine, ci reglementează gradul de filtrarea şi de acces al intrărilor din dicționarul ce conține deja elemente de fonetică şi de morfologie.

În practică dicționarul va conține cuvinte şi un cumul de criterii, fonetice şi morfologice. Regulile sintactice se adaugă în calitate de constrângeri ce acționează la nivel macro-sintagmatic.

3.3. Realizarea dicționarului şi a corpusului

Pentru un proiect de recunoaştere a vorbirii, dicționarul include intrările

lexicale atent selectate după o serie de criterii. Cum proiectul de față se adresează serviciilor publice din zona juridică, selectarea contextelor reprezentative pentru acest domeniu va ține cont atât de jargonul de specialitate, cât şi de cuvintele uzuale.

Criteriul frecvenței

Cuvintele trebuie selectate pe baza unui corpus de texte juridice (comunicate, sentințe, transcrieri din cadrul proceselor), pe baza corpusurilor realizate ad-hoc şi a celor existente prezentate în Dascălu (2002), Dascălu & Pop (2003).

Criteriul tematic

Corpusul va acoperi texte ce țin de domeniile: apărare, condamnare, ordin, rugăminte, solicitare informațiişi expunere argumente. Realizarea unui dicționar pentru domeniul juridic trebuie să conțină două categorii de cuvinte:

a. cuvinte ce țin de limbajul comun, deci de o zonă invariabilă a vocabularului:

- *prepoziții*: cu Acc (cu, pe, peste, în, din, la etc.), cu G-D (împotriva, contra, asupra, conform, potrivit, contrar etc);
- *conjuncții:* coordonatoare (dar, iar, însă, ci, şi, sau, ori) şi subordonatoare (pentru că, din moment ce, ca să, dacă etc.);
- *pronume şi adverbe relative*: care, cine, ce, unde, când, cum, cât;
- *pronume*: personale (eu, tu, el), demonstrative (acesta, acela, acelaşi), nehotărâte (oricare, fiecare, altul), negative (nimeni, nimic);
- *nume*: zile, luni, ora, instituții/locații, nume persoane, grade de rudenie;
- *expresii de mod* (aşa, repede, încet, corect, incorect, conform procedurii, contrar procedurii);
- *expresii de timp* (acum, atunci, cândva, acum <x> ore/zile/luni/ani, ieri, ora/ziua/săptămâna/luna trecută, anul trecut, când ….., de <x> ore/zile/săptămâni/luni/ani, din <data/anul> până în <data/anul>);
- *expresii de loc* (aici, acolo, în+<loc>, la+<loc>).

b. cuvinte specializate:

- *verbe/expresii de stare*: a fi, a exista, a sta, a se mira, a se învecina, a se înrudi;

- *verbe/expresii de opinie*: a crede, a considera, a pune sub semnul întrebării, a contesta, a admite, a fi de acord, a nu fi de acord;
- *verbe/expresii de acțiune*: a lua, a da, a crea, a realiza, a evidenția, a convoca, a elibera, a pune (sub acuzare), a trimite, a primi, a fura, a comite o infracțiune, a defăima, a omorî, a vătăma, a răni, a trece, a supraveghea, a învăța, a repara, a vinde, a cumpăra, a se căsători, a divorța, a strânge, a elimina, a introduce, a facilita, a media, creşte, a descreşte, a îmbătrâni, a îngriji, a acorda atenție, a ezita, a minți, a spune adevărul, a aduce dovezi, a utiliza, a contesta;
- *verbe/expresii performative*: mă scuzați, mă/ne angajăm să, te/vă acuz, te/vă autorizez, te/vă anunț, te/vă avertizez, te/vă declar, te/vă felicit, te/vă insult, te/vă întreb, te/vă previn, te rog/vă rog, te/vă sfătuiesc, îți/vă interzic, îți/vă jur, îți/vă mulțumesc, îți/vă ordon, îți/vă promit, îți/vă recomand, îți/vă sugerez;
- *verbe/expresii de senzație*: i se urăşte, i se face frică, i se face teamă, a se teme, i se face rău, îi place, îi pare bine, îi pasă, îl doare, îl uimeşte, îl îngrozeşte, a iubi, a urî;
- *verbe/expresii de aspect/modale* : a începe, a continua substantiv/de+Participiu, a termina substantiv/de+Participiu, a putea să, trebuie să, e nevoie de/să, a avea nevoie de/să, e posibil să, e obligatoriu să, e necesar să, e interzis să, e permis să, e recomandat să;
- *substantive* de tip actanți: inculpat, reclamant, pârât, petent, avocatul apărării, avocatul acuzării, magistrat, judecător, grefier, procuror, polițist, organele de poliție, martor;
- *substantive* de tip acuzații: coruptie, înşelăciune, fals, uz de fals, crimă, defăimare, evaziune;
- *substantive* de tip sancțiuni: pedeapsa cu închisoare între <x> zile/luni/ani şi <x> zile/luni/ani, plata retroactivă amendă, prestarea de muncă în folosul comunității, privare de libertate, arest - arestare – în stare de arest - arest la domiciliu;
- *substantive* de tip nume locuri: închisoare, penitenciar, tribunal;
- *substantive* de tip elemente ale procesului juridic: acuzare, dovedirea nevinovăției, chemarea în instanță a martorilor/pârâților, amânarea procesului/a sentinței.

Corpusul reprezintă un set de enunțuri înregistrate care au fost în prealabil:

- Selectate sau construite pentru a acoperi inventarul relevant fonetic şi morfologic, din dicționar;
- Parsate în conformitate cu regulile fonetico-morfologice convenite de specialişti.

Cu cât enunțurile sunt rostite de mai multe persoane în diverse contexte de claritate, comportamentul elementelor gramaticale şi interpretarea acestora se poate evidenția mai bine. Acuratețea şi relevanța corpusului sunt dublate de gradul de individualizare a regulilor gramaticale pentru limba respectivă.

3.4. Nivelul fonetic. Generalități

Pentru o analiză corectă şi coerentă a nivelului fonetic, trebuie să se înțeleagă de la bun început o serie de aspecte fundamentale ale acestui nivel:

1. Sunetele ce alcătuiesc un cuvânt sunt individualizate prin intermediul fonemelor, adică prin intermediul reprezentării lingvistice asociate unui sunet.
2. Fonemul este o reprezentare ideală a sunetului. Ceea ce interesează este ocurența sunetului într-un context diagnostic, adică alofonul. Alofonul unui sunet este rezultatul influenței exercitate asupra acestuia de contextul precedent şi/sau succedent.
3. Decelarea inventarului de alofone specifice unei limbi se realizează prin identificarea proceselor fonetice specifice acelei limbi.
4. Gradul variabil de sonoritate a fluxului vocal determină existența vocalelor şi a semivocalelor.
5. Adiacența vocală-vocală se realizează doar în silabe diferite şi poartă numele de hiat.
6. Adiacența vocală-semivocală (în diverse poziții) se realizează în aceeaşi silabă şi poartă numele de diftong (o vocală+o semivocală) sau de triftong (o vocală+2 semivocale).

Din această prezentare schematică, se poate uşor observa necesitatea înțelegerii şi a interpretării valorii sunetului: (i) în contextul adiacent sau diagnostic din care face parte; (ii) în contextul apartenenței sale la un tip de proces fonetic care îi modifică atributele; (iii) în contextul silabei din care face parte.

În continuarea acestui articol, vom face o scurtă prezentare a patru aspecte esențiale pentru o bună realizare a analizei fonetice pentru limba română şi pentru sistemele ce o utilizează în prelucrare: problema diftongilor şi a triftongilor, problema hiatului, probleme de omofonie şi omografie, problema stabilirii unui inventar minimal de procese fonetice.

3.5. Diftongii şi triftongii vs. hiatul în limba română

Elementele *a*, *ă*, *î*, *ă* sunt mereu vocale în limba română. Rezultă de aici că sunetele *e*, *i*, *o*, *u* vor putea fi fie vocale, fie semivocale în funcție de contextul din care fac parte.

Diftongul este asocierea unei vocale şi a unei semivocale în aceeaşi silabă. În funcție de poziționarea semivocalei în raport cu vocala, rezultă:

- 14 diftongi descendenți:
  - Semivocala *i*: ai (e.g. hai), ăi (e.g. a destăinui), îi/âi (e.g. pâine), oi (e.g. doi), ui (e.g. a uita), ei (e.g. idei), ii (e.g. fii);
  - Semivocala *u*: au (e.g. lucrau), ău (e.g. rău), îu/âu (e.g. grâu), ou (e.g. lingou), iu (e.g. viu), eu (e.g. meu), uu (e.g. perpetuu).
- 10 diftongi ascendenți:
  - Semivocala *i*: ia (e.g. iarbă), ie (e.g. iertare), io (e.g. niciodată), iu (e.g. iubire);
  - Semivocala *e*: ea (e.g. începea), eo (e.g. vreo);
  - Semivocala *u*: ua (e.g. ploua), uă (e.g. nouă), uî/uâ (e.g. plouând);
  - Semivocala *o*: oa (e.g. moară).

Triftongul este asocierea unei vocale şi a două semivocale în aceeaşi silabă. În funcție de poziționarea semivocalelor în raport cu vocala, rezultă:

- 11 triftongi descendenți: eau (e.g. aveau), eai (e.g. primeai), iai (e.g. suiai), iau (e.g. miau), oai (e.g. leoaică), iei (e.g. miei), ieu (e.g. maieu), iou (e.g. maiou), uăi (e.g. rouăi), uai (e.g. înşeuai), uau (e.g. înşeuau);
- 3 diftongi ascendenți: eoa (e.g. leoarcă), ioa (e.g. lăcrămioară), uea (e.g. înşeuează).

În momentul în care ocurențele de mai sus, adică grupările de sunete de mai sus apar în silabe diferite acestea nu mai sunt interpretate ca diftongi sau triftongi, ci ca hiat. Hiatul este deci prezența a două vocale în silabe diferite. Hiatul implică şi fenomenul fonetic de stop glotal cu apendice palatal prezentat mai jos în cadrul proceselor fonetice.

(1) Aţi mai furat doi cai cu un an înainte.

Cuvântul *cai* conţine diftongul descendent [aï], în timp ce *î-na-in-te* conţine hiatul [a-^{j}i] unde sunetul *i* este concretizat ca alofon cu stop glotal şi apendice palatal. Modul în care gruparea *ai* se pronunţă diferă, dar în transcrierea vorbirii, gruparea trebuie redată identic ca grafie.

Manifestarea grupurilor de vocale şi semivocale ca hiat vs. diftong/triftong implică trăsături articulatorii diferite, cu implicaţii pentru scriere, motiv pentru care, pentru fiecare grupare, vor trebui identificate cazurile de omofonie şi omografie. De multe ori, aceste situaţii sunt motivate sintactic sau semantic.

3.6. Probleme de omofonie şi omografie

Fenomenul în care o singură grafie sau scriere are drept corespondent două sau mai multe pronunţii diferite poartă numele de omografie. Invers, în situaţiile în care o anumită grupare se pronunţă identic, dar scrierea e diferită, vorbim de omofonie.

(2) Mai ştii când m-ai lovit?

Cuvântul *mai* şi expresia *m-ai* se pronunţă identic. Cu toate acestea, grafia este diferită din motive ce ţin de morfologie şi sintaxă: *mai* este adverb, pe când *m-ai* este o combinaţie de pronume personal în Acc şi auxiliar verbal. Se poate observa astfel motivul pentru care dicţionarul trebuie să conţină elemente din toate palierele limbii.

Oferim în cele ce urmează câteva exemple în care omofonia, omografie şi schimbarea de sens conlucrează, fapt ce are drept consecinţă o complexitate sporită a regulilor lingvistice la nivel fonetic.

(3) Grup identic - Pronunţie identică – Scriere diferită:

(3.1.) sau vs. s-au
Sau inculpaţii au adus noi dovezi, *sau* avocaţii lor *s-au* schimbat.

(3.2.) mai vs. m-ai
În luna *mai* *m-ai* contactat să mai povestim despre afaceri.

(3.3.) cai vs. c-ai
C-ai văzut, că n-ai văzut niciunul din cei doi *cai* nu mă priveşte.

(3.4.) nai vs. n-ai
N-ai cumva un *nai*?

(3.5.) iar vs. i-ar
I-ar oferi o nouă şansă *iar*.

(3.6.) Iaşi vs. i-aş
I-aş fi dat documentele la *Iaşi* dacă aş fi avut suficient timp.
(3.7) cei vs. ce-i
Ce-i poți oferi când *cei* mai bogați nu au soluții?
(3.8) Mia-i vs. mia-i vs. mi-ai
Tot asta *mi-ai* spus: „*Mia-i* de încredere, iar mia-i la mine."

(4) Grup identic - Pronunție diferită – Scriere diferită:
ie vs. i-e
Cu această *ie* *i-e* rece în celulă.

(5) Grup identic – Pronunție diferită – Scriere identică:
(5.1.) Diftongul ascendent *iü* vs. Diftongul descendent *ïu* vs. Hiatul iu
iulie vs. *vişiniu* vs. el ştiu
iunie vs. *eu ştiu* vs. el ştiu
(5.2.) Diftongul descendent *aï* vs. Hiatul ai
haină (substantiv) vs. haină (adjectiv)
(5.3.) Diftongul descendent *âï* vs. Hitaul âi
eu mă *bâlbâi* (acum) vs. el se bâlbâi (ieri)
(5.4.) Diftongul descendent *aï* (context anterior: consoană) vs. diftongul descendent aï (context anterior: vocală)
cai vs. ştiai
(5.5.) Pronunția hiatului ca *diftong* în funcție de subiectivitatea vorbitorului (rapiditate în vorbire)
a-gre-si-u-ne vs. a-gre-*siu*-ne
şti-ut vs. *ştiut*
glo-ri-e vs. glo-*rie*
a-bre-vi-e-re vs. a-bre-*vie*-re

3.7. Procese fonetice

Procesele fonetice se referă la transformările pe care un fonem le suportă în contextul său diagnostic. Rezultatul proceselor fonetice vor fi alofonele, adică instanțele clasei fonem. Procesele fonetice ale limbii române au fost studiate şi documentate de Rosetti (1963), Stati (1972) şi Neamțu (2002).

Palatalizarea şi labializarea

Palatalizarea şi labializarea acționează doar în cadrul unei silabe, iar fenomenele presupun influența vocalelor şi/sau a semivocalelor *e* şi *i* (cazul palatalizării), respectiv a vocalelor şi/sau a semivocalelor *o* şi *u* (cazul

labializării) asupra consoanei care le precedă şi asupra vocalei care le succedă.

(6) via-ță – alofonele palatalizate sunt [v] şi [a] sub influența lui *i.*

(7) no-uă – alofonele labializate sunt [n] sub influența lui *o* şi [ă] sub influența lui *u.*

Palatalo-labializarea reprezintă influența semivocalelor *eo*, *io* şi a diftongilor ascendenți *ïo*, *ëo*, *ïu* asupra consoanei care îi precedă şi/sau asupra vocalei care le succedă.

(8) pleoa-pă – alofonele palatalo-labializate sunt [l] şi [a] sub influența lui *eo*

Nazalizarea

Nazalizarea este un proces ce influențează vocalele care sunt urmate de consoanele nazale *m* sau *n*, atunci când *m* sau *n* sunt urmate de o consoană. Consoana nazală şi consoana care îi urmează trebuie să facă parte din aceeaşi silabă doar în cazul lui *m*.

(9) an-te-ri-or – alofonul nazalizat e [a] sub influența lui *n+cons.*

Fricativizarea

Fricativizarea reprezintă efectul consoanelor *f* sau *v* asupra consoanele *m* sau *n* care le precedă.

(10) in-for-ma-ți-e – alofonul fricativizat e [n] sub influența lui *f.*

Fricativizarea este produsă însă şi când consoanele *s*, *z*, *ş* sau *j* afectează consoana *n* care le precedă.

(11) în-jo-si – alofonul fricativizat e [n] sub influența lui *j.*

Velarizarea

Velarizarea este procedeul fonetic care afectează consoana *n*, atunci când *n* este urmată de sunetele *k*, *g* sau *h*.

(12) en-gle-ză – alofonul velarizat e [n] sub influența lui *g.*

Bilabializarea

Bilabializarea este procesul fonetic în urma căruia consoana *n* se pronunță *m*. Acest fenomen apare când *n* este urmat de *p*, *b* sau *m*.

(13) în magazie – alofonul bilabializat este [n] sub influența lui *m.*

Desonorizarea

Desonorizarea constă în asurzirea sau amuțirea sunetelor *m*, *n*, *l* când sunt precedate de o consoană în poziție finală de cuvânt.

(14) calm – alofonul desonorizat e [m] sub influența poziției sale finale şi a lui *l.*

Aspirarea

Aspirarea este procesul fonetic prin care sunetele *p*, *b*, *t*, *d*, *k*, *g* sunt însoțite de un uşor sunet *h*.

(15) cap – alofonul aspirat e [p] sub influența poziției sale finale.

Neutralizarea

Neutralizarea este procesul fonetic în urma căruia sunetele *s*, *z* se confundă când sunt urmate de *m*, *n*, *l*.

(16) snob – alofonul neutralizat e [s] sub influența lui *n.*

Anularea stopului glotal

Anularea stopului glotal apare atunci când două vocale se află în hiat. În fața celei de-a doua vocale se aude un sunet mai slab (asemănător unei semivocale) asemănător lui *i* (notat ^{j}) sau lui *u* (notat ^{w}).

(17) am-bi-ți-^{j}e (stop glotal cu apendice palatal)
sca-^{w}un (stop glotal cu apendice labial

Scurtarea vocalică postconsonantică

Scurtarea vocalică postconsonantică este procesul fonetic în urma căruia *i*, aflat la final de cuvânt, are o pronunțare scurtă, lipsită de intensitate sonoră

(18) Tu vezi pomi.

Ceea ce procesele fonetice scot în evidență este faptul că vocala *a*, de exemplu, va avea asociat diferite spectre acustice şi diferite reprezentări ale trăsăturilor sale fonetice în funcție de procesul fonetic care acționează asupra sa contextual. Aceste ocurențe trebuie avute în vedere în realizarea corelațiilor dintre reprezentarea grafică a lui *a*, în acest caz, şi multitudinea modurilor de pronunție ale sunetului corespondent în vorbirea articulată.

4. Concluzii

Perspectivele din punctul de vedere al aplicabilității unui sistem automat de recunoaştere a vorbirii sunt nelimitate. Obținerea unor rezultate caracterizate de acuratețe şi eficiență în transcrierea fluxului oral pentru limba română depinde de modul discret de organizare a nivelurilor ierahice ale limbii. Analiza fonetică este primul pas într-un astfel de demers, dar limitările sale, în absența introducerii unor reguli de morfologie şi de sintaxă, sunt evidente. Fenomenele de omografie şi omofonie, spectrul proceselor fonetice şi prezența grupărilor diftong/triftong/hiat determină un model lingvistic complex care necesită cercetări colaborative de lungă durată.

Mulțumiri

Studiul de față a fost posibil ca urmare a finanțării de catre UEFISCDI a proiectului JustASR în cadrul programului Parteneriate, prin contractul 14/01.11.2013, cod înregistrare proiect PN-II-PT-PCCA-2013-4-1644.

Referințe

Antal, M., *Contributions to Speech and Speaker Recognition.* Cluj-Napoca, Technical University of Cluj-Napoca, 2006.

Besacier, L., Barnard, E., Karpov, A., Schultz, T. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 85-100.

Besacier, L., Barnard, E., Karpov, A., Schultz, T. 2014 b. Introduction to the special issue on processing under-resourced languages. *Speech Communication* , 83-84.

Boldea, M. *Contributions to Automatic Continuous Speech Recognition for the Romanian Language*. Timişoara: Politehnica University, 2003.

Burileanu, C., Popescu, V., Buzo, A., Petrea, C., Ghelmez-Hanes, D. 2010. Spontaneous Speech Recognition for Romanian in Spoken Dialogue Systems. *Proceedings of the Romanian Academy series A*, 83-91.

Cinque, G., Giusti, G. *Advances in Romanian Linguistics*. John Benjamins Publishing, 1995.

Cristea, D. 2005. Resurse lingvistice si tehnologii ale limbajului natural. Cazul limbii române, *Prelegeri Academice*, Academia Româna - Filiala Iaşi, vol. III, nr. 3, Iaşi.

Cucu, H., Buzo, A., Besacier, L., & Burileanu, C. (2014). SMT-based ASR domain adaptation methods for under-resourced languages: Application to Romanian. *Speech Communication*, 195-212.

Dascălu Jinga, L., Pop, L. *Dialogul în limba română vorbită*. Oscarprint, 2003.

Dascălu Jinga, L. *Corpus de română vorbită (CORV). Eşantioane*. Oscarprint, 2002.

*** *DOOM (Dicţionarul ortografic şi ortoepic al limbii române)*. Editura Universul Enciclopedic, 2005.

eDTLR – The Thesaurus Dictionary of the Romanian Language, 2013, http://metashare.info.uaic.ro/repository/browse/edtlr-the-thesaurus-dictionary-of-the-romanian-language/c05a74c063d611e28e8252540060617d5d02cd135a8b4759904cb7a956d41aa7/

Dumitrescu, S.D., Boros, T., Ion, R., 2014, Crowd - sourced, automatic speech - corpora collection - building the Romanian Anonymous Speech Corpus, presentation at LREC Workshop, http://www.ilc.cnr.it/ccurl2014/LREC2014-Workshops_CCURL2014-Oral-7_Presentation.pdf

*** *GALR (Gramatica Academiei Limbii Române)*. Editura Academiei Române, 2008.

*** *HTK Speech Recognition Toolkit*, http://htk.eng.cam.ac.uk/, 2009.

Neamțu, G.G. *LRC. Fonetica limbii române*. curs susținut la Facultatea de Litere, Universitatea Babeş-Bolyai, Cluj-Napoca, 2002.

****NuanceCommunications,Inc.* Nuance Dragon Dictation and Dragon Search App Now Available in the Czech Republic and Romania. Press release from May 11, 2012 http://www.nuance.com/company/news-room/press-releases/czechromaniaapps.doc

Petrea, C. S., Buzo, A., Cucu, H., Pasca, M., & Burileanu, C., 2010. Speech Recognition Experimental Results for Romanian Language. *Proceedings of the 6th European Conference on Intelligent Systems and Technologies.*

Petrea, S., Ghelmez-Hanes, D., Buzo, A., Burileanu, C. 2009. Statistical Results in the Context of Romanian Spontaneous Speech Recognition . *Proceedings of the 13rd International Conference Speech and Computers SPECOM 2009.* St. Petersburg.

Popescu, V., Burileanu, C., Caelen, J. 2008. Unsupervised Speaker Tracking in a Speech Recognition Module for Multi-Party Human-Computer Dialogue . *16th European Signal Processing Conference EUSIPCO 2008.*

Rabiner, L., Huang, Y. (2005). Automatic Speech Recognition – A Brief History of the Technology, în K. Brown, *Encyclopedia of Language and Linguistics Second Edition*, Elsevier, 523-541.

Rosetti, Al. *Introducere în fonetică*. Editura Ştiinţifică, 1963.

*** *RASC (Romanian Anonymous Speech Corpus)*, http://rasc.racai.ro/, 2015.

Schiopu, D. 2010. A Comparative Study of Three Speech Recognition Systems for Romanian Language, *The 5th International Conference on Virtual Learning.*

Sofroni, V., Stan, A., (2015), O Descriere generală a Arhitecturii Sistemelor Automate de Recunoaştere Vocală. *Revista Romana de Interactiune Om-Calculator 8(1),* 45-64.

Stati, S. (1972). Unităţile limbii, în A. Graur, S. Stati, L. Wald, *Tratat de lingvistică generală*. Editura Academiei. 221-233

Teodorescu, H. 2010. AI Tools for Speech Analysis Applied to the Romanian Language. *Proceedings of the 4th Conference on European Computing.* WSEAS Press.

Teodorescu, H., Pistol, L., Feraru, M., Zbancioc, M., & Trandabat, D. (2010). *Sounds of the Romanian Language Corpus (SRoL).* Retrieved from http://www.etc.tuiasi.ro/sibm/romanian_spoken_language/index.html

Toderean, G., Buza, O. (2008). *Research report on project Interactive System based on Voice for Blind People.* Cluj Napoca: Technical University of Cluj Napoca.

Trandabăţ, D., Irimia, E., Barbu Mititelu, V., Cristea, D., Tufis, D. *The Romanain Language in the Digital Age-Limba Romana in Era Digitala.* Springer, 2012.