Platforma pentru căutarea în fișiere audio

Scopul principal al acestei platforme este acela de a permite identificarea modalității de pronunțare a cuvintelor în diferitele contexte în care pot să apară. Astfel, se permite realizarea unei căutări pornind de la un cuvânt (sau lema asociată acestuia). În plus, pentru a veni în ajutorul utilizatorilor care sunt interesați de anumite proprietăți ale cuvântului, se poate realiza și o filtrare după adnotările asociate acestuia.

Informațiile afișate în urma execuției unei căutări conțin: contextul de apariție al cuvântului (acesta putând fi întreaga frază sau limitat la 5 cuvinte), lema, adnotarea morfo-sintactică. De asemenea, sunt oferite opțiuni de ascultare a sunetului: la nivelul cuvântului căutat sau al întregii fraze.

Interfața de căutare este realizată astfel încât diferiții parametri de căutare specificați sunt salvați în URL-ul generat în browser. Astfel, la identificarea unei căutari de interes, URL-ul generat poate fi salvat de utilizator, fie utilizând funcția "add to favorites", disponibilă în toate browserele, fie trimițând URL-ul generat prin email sau notându-l într-un fișier pentru accesare ulterioară.

Platforma este formată din trei componente: interfața cu utilizatorul, componenta de indexare a informației, componenta de stocare. Cele trei componente sunt separate sub formă de obiecte programatice care expun funcționalitățile disponibile la fiecare nivel.

Componenta de interfață cu utilizatorul se prezintă sub forma HTML pentru a putea fi accesată utilizând un browser. În vederea păstrării compatibilității cu cât mai multe browsere existente în acest moment, s-a încercat simplificarea acesteia și utilizarea de elemente standard. Cu toate acestea, pentru asigurarea redării sunetelor, este necesar un browser compatibil HTML 5. Toate browserele moderne suportă standardul HTML 5, inclusiv Google Chrome, Mozilla Firefox, Safari, Internet Explorer, atât în versiunile desktop, cât și în versiunile mobile.

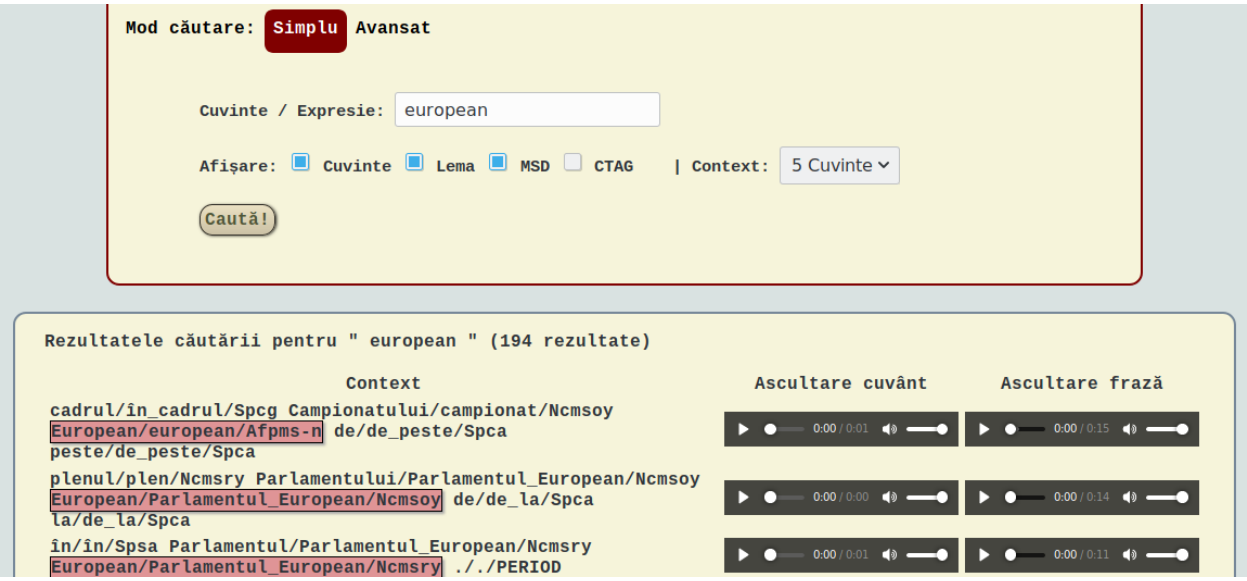
Componenta de indexare a platformei utilizează fișierele aliniate produse prin procese automate. În urma procesării acestora, se asigură îmbogățirea informației utilizând serviciul TTL, prin generarea automată a lemei și adnotării morfo-sintactice pentru fiecare cuvânt. Având în vedere că serviciul TTL returnează adnotările utilizând tag-uri MSD, s-a realizat și o conversie către un set simplificat de tag-uri (CTAG), mai ușor de utilizat și urmărit de către utilizatori.

Componenta de stocare asigură stocarea pe disc a informațiilor, indexate cu ajutorul componentei de indexare, în vederea accesării din interfață. Sunt utilizate tehnologii care permit creșterea volumului de date indexate fără efecte sesizabile asupra vitezei de interogare.

Exemple de căutări

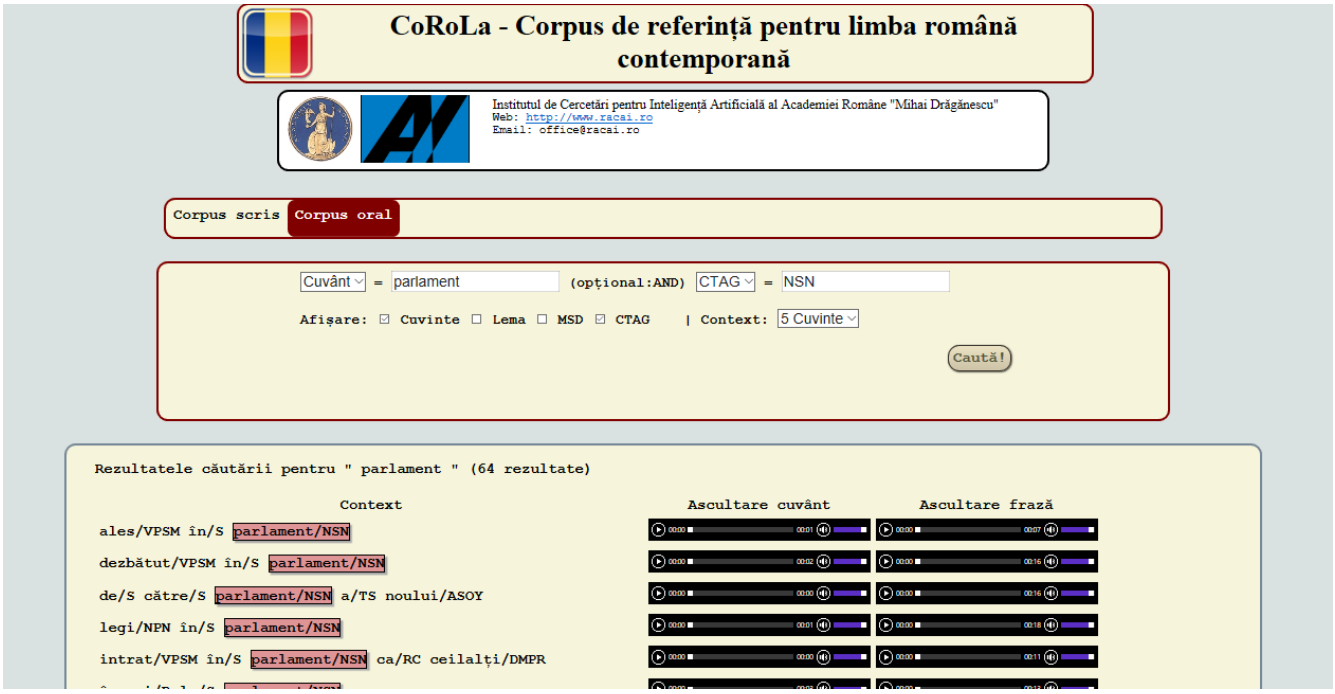
I. Modul de căutare: SIMPLU - permite doar căutarea după forma ocurentă

1. Căutarea unui singur cuvânt



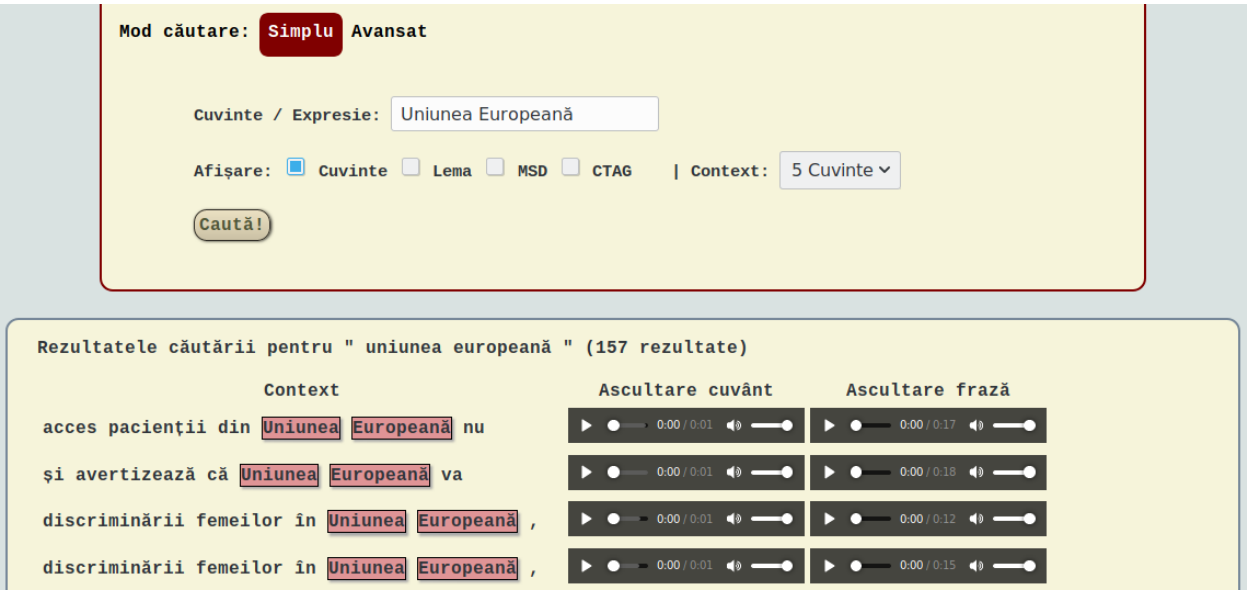
Exemplul 1. Căutare după cuvântul “european”, cu afișare complexă: vezi secțiunea „Afișare” unde s-au bifat: Cuvinte (permite afișarea formei ocurente în text), Lema (permite afișarea lemei), MSD (permite afișarea descrierii morfosintactice).

Exemplul 2. Căutare după cuvântul “european”, cu afișare simplificată: doar Cuvinte este bifat.



Exemplul 3. Căutare complexă introducând cuvântul dorit și adnotarea sub forma CTAG

2. Căutări ale unui șir de cuvinte



Exemplul 4. Căutarea a două cuvinte consecutive și afișarea doar a cuvintelor (fără informație suplimentară)

Figura 5. Căutarea a trei cuvinte consecutive, separate ca tokeni (vezi spațiul introdus între pronume și verbul auxiliar).

II. Modul de căutare: AVANSAT - permite căutarea după forma ocurentă, după lema și/sau descrierea morfosintactică (MSD)

1. Căutarea unui singur cuvânt

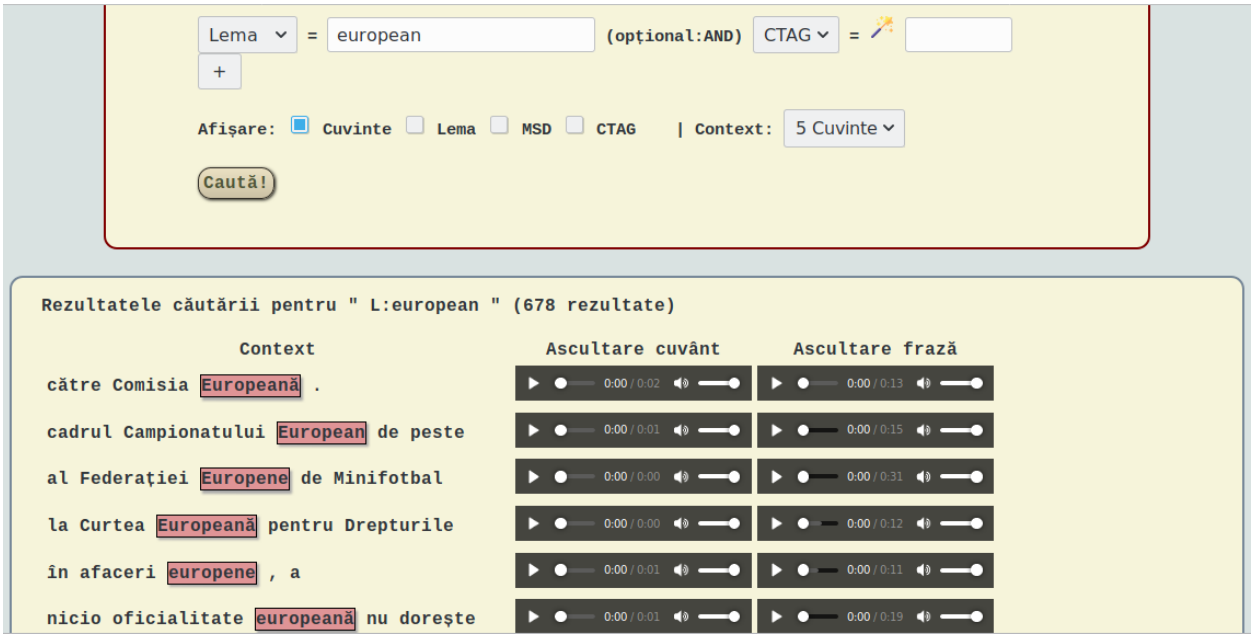


Figura 6. Căutarea ocurențelor formelor din paradigma unui cuvânt (aici: „european”) precizat

prin lema sa.

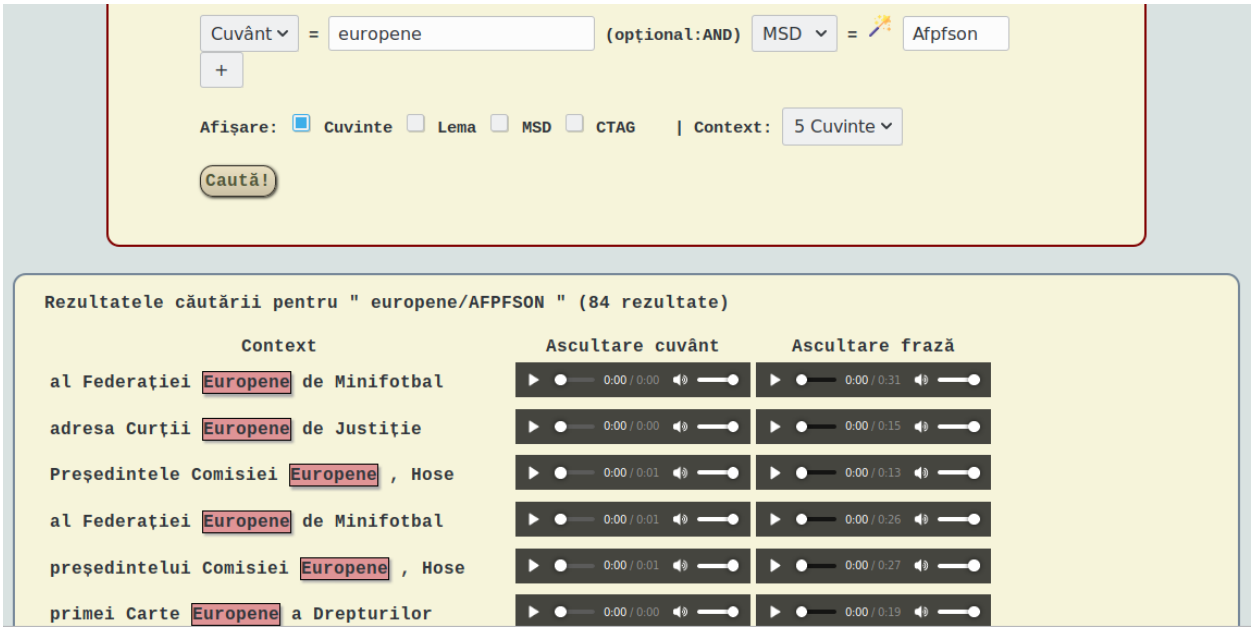


Figura 7. Căutarea ocurențelor adjectivului „europene”, la singular Dat-Genit. Restricțiile asupra valorilor morfologice ale cuvântului introdus sunt formulate în fereastra MSD, folosind valorile posibile din fișierul „Specificații privind etichetele MSD” de pe corola.racai.ro, secțiunea *Interogare*.

Aici, bagheta magică reacționează la clic deschizând o fereastră în care se pot selecta valorile fiecărui atribut. La final, nu uitați să apăsați butonul „Setare TAG”!

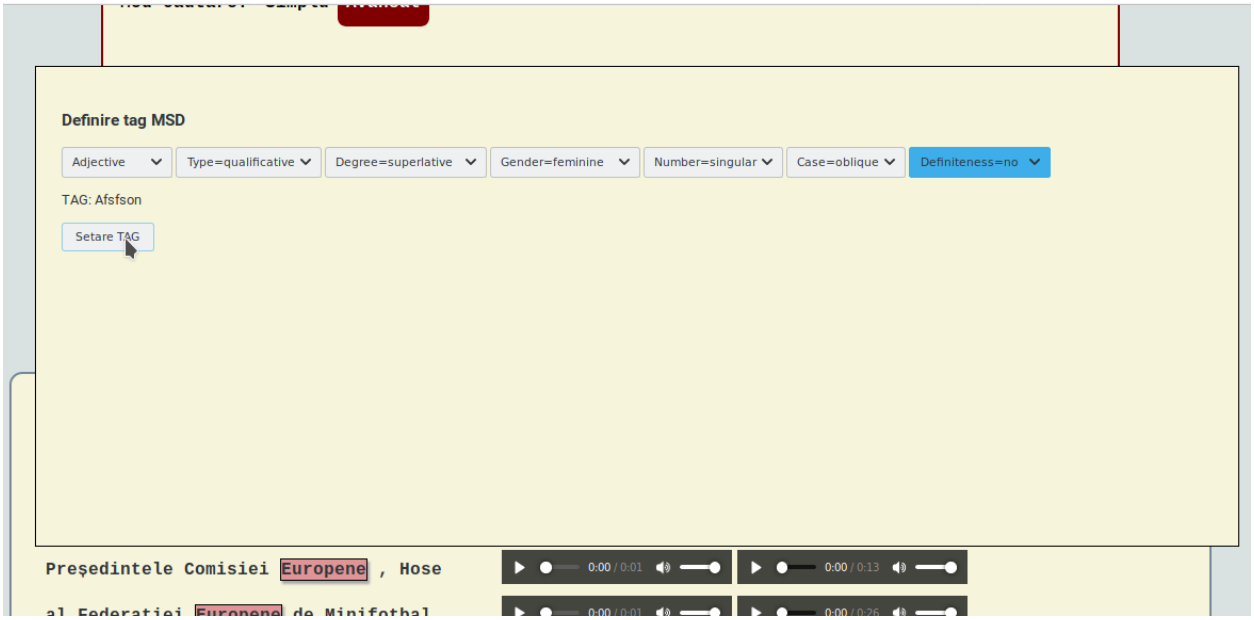


Figura 8. Fereastra pentru formularea descrierii morfosintactice (MSD).

2. Căutări ale unui șir de cuvinte

Pasul I: Prin apăsarea succesivă a butonului cu semnul „+”, se generează atâtea câmpuri câte sunt necesare.

Pasul al II-lea: se completează câmpurile pentru fiecare cuvânt/lemă

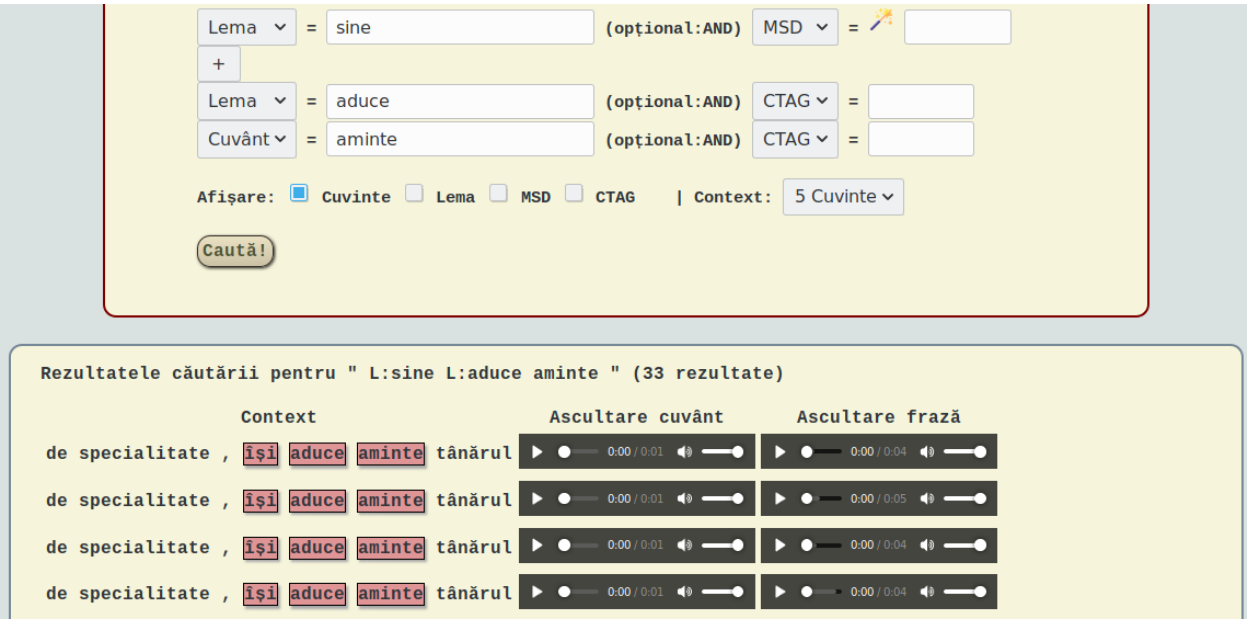


Figura 9. Căutarea șirului: lema „sine” + lema „aduce” + forma ocurentă „aminte”.

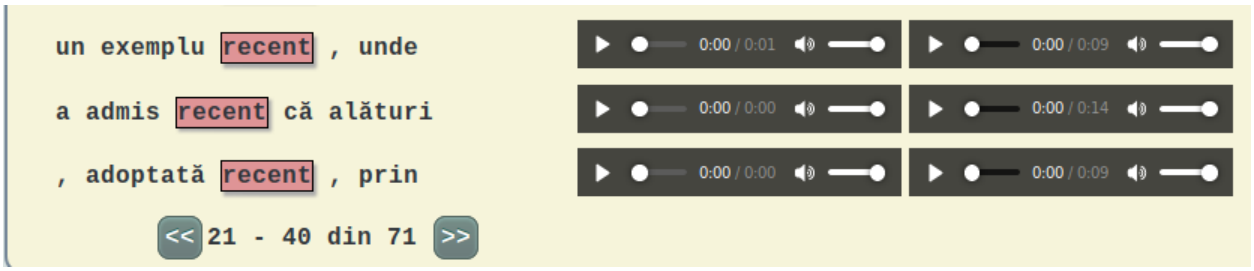


Figura 10. Atunci când exemplele regăsite nu încap pe o singură pagină, se poate naviga prin ele cu ajutorul butoanelor din subsolul paginii.

Referințe

Dan Tufiș, Verginica Barbu Mititelu, Elena Irimia, Vasile Păiș, Radu Ion, Nils Diewald, Maria Mitrofan, Mihaela Onofrei, Little strokes fell great oaks. Creating CoRoLa, the reference corpus of contemporary Romanian, RRL, 3/2019, p. 227-240. - *conține informații despre conținutul corpusului CoRoLa, despre proveniența textelor orale, cantitatea lor și nivelurile de adnotare*