Corpus de ghiduri turistice generale despre România, redactate în română şi franceză. Constituire, tratare, instrumente de lucru şi tipuri de analize

Ioana Daniela Balauta

Corpus de ghiduri turistice generale despre România, redactate în română și franceză. Constituire, tratare, instrumente de lucru și tipuri de analize

Ioana – Daniela Bălăuță
PhDstudent, Doctoral School of Socio-Human Sciences, Ștefan cel Mare University, Suceava
Doctorante, Ecole Doctorale SLPCE (Sciences du Langage, Psychologie, Cognition et Education), Université de Poitiers, France

Rezumat: În lucrarea noastră, ne propunem să abordăm modul în care constituim un corpus electronic și cum trebuie să-l prelucrăm în vederea exploatării cu programe de lectură. Obiectivele noastre sunt acelea de a prezenta metodologia utilizată pentru obținerea unui corpus electronic plecând de la texte imprimate, precum și unele posibilități de editare și formatare a datelor pentru a le aduce în forma acceptată de parametrii de funcționare a programelor. Vom expune funcționarea programelor de analiză de date textuale ***IRaMuTeQ***, ***Lucon*** și ***Cordial Pro***, cu scopul de a prezenta tipurile de analize posibile, având în vedere că, instrumentele statistice, utilizate în mod judicios, pot susține demersurile cercetătorului, pentru a observa mai bine proprietățile lingvistice ale corpusului supus analizei.

Considerăm că analiza statistică a corpusului nostru de ghiduri turistice oferă un teren important pentru abordarea diferitelor clase gramaticale care țin de specificitatea generală a genului discursiv al ghidului turistic. De exemplu, realizarea unei analize comparative a ghidurilor prin intermediul distribuției substantivului ne permite să identificăm câmpurile tematice conceptuale cele mai relevante, pe care se bazează construirea din punct de vedere lingvistic a referentului – România.

Credem că utilizarea instrumentelor informatice ne permite să dăm mai multă vizibilitate rezultatelor cantitative, valabile și pertinente, pentru a le insera în abordarea calitativă. Trebuie să remarcăm că, programele informatice, care asistă munca cercetătorului cu analize automate și calcule statistice, au câștigat notorietate în cercetările lingvistice actuale, în general, însă, alegerile metodologice și interpretarea rezultatelor aparțin cercetătorului.

Cuvinte cheie: *corpus electronic, editare, formatare, programe de lectură, analize statistice*

1. Introducere

Abordarea faptelor de limbă pe bază de corpus este proprie analizelor lingvistice realiste, deoarece, credem noi, acest demers se bazează pe date reale, atestate, și nu se aplică pe exemple fabricate pentru a justifica o teorie sau un studiu. În acest sens, ne raliem opiniei lui Bénédicte Pincemin [1999], [2020], care consideră corpusul ca fiind apanajul unei lingvistici descriptive care îl observă pentru a-i reconstitui regularitățile.

Precizăm că lucrul pe un corpus prelucrat pe calculator în scopul cercetării presupune o adaptare a datelor lingvistice disponibile, în privința tratării pe care ne-o propunem, ținând cont de criteriile de constituire utilizate dar și de limitele corpusului. În acest sens, vom prezenta metodologia utilizată pentru obținerea corpusului electronic, adică culegerea datelor pentru constituirea materială, începând cu scanarea și ocerizarea textelor imprimate ale ghidurilor turistice incluse în Tabelul 1. Menționăm că fișierele electronice obținute trebuie prelucrate, adică editate, curățate și formatate cu parametrii ceruți pentru buna funcționare a programelor utilizate pentru lectura și analiza corpusului.

Vom detalia, de asemenea, funcționarea programelor de analiză (*IRaMuTeQ, Lucon, Cordial Pro*) în ceea ce privește tratarea datelor lingvistice pentru fiecare limbă separat, română și franceză, având în vedere că instrumentele statistice, utilizate în mod judicios, pot susține demersurile cercetătorului. Considerăm că analiza statistică a corpusului de texte oferă un teren important pentru abordarea diferitelor categorii gramaticale care țin de specificitatea generală a genului discursiv al ghidului turistic. De exemplu, realizarea unei analize comparative a ghidurilor prin intermediul distribuției substantivului ne permite să identificăm câmpurile tematice conceptuale cele mai relevante, pe care se bazează construirea din punct de vedere lingvistic a referentului – România.

Vom prezenta cele două tipuri de abordări ale corpusului *corpus-based* și *corpus-driven*, utilizate în tradiția anglo-saxonă, pe care le considerăm complementare, precum și abordarea cantitativă și cea calitativă în lingvistica de corpus.

2. Conceptul de corpus și lingvistica de corpus. Elemente de reflecție asupra opțiunilor metodologice de lucru pe corpus

În această secțiune a cercetării noastre, ne propunem să facem o trecere în revistă a noțiunii de corpus în lingvistica franceză cu scopul de a fundamenta opțiunea noastră pentru utilizarea corpusului în cercetare și pentru a preciza poziționarea noastră față de aceste studii.

Prin intermediul cercetărilor întreprinse, referitor la noțiunea de „corpus", trebuie să remarcăm că, în lingvistica franceză, acest concept primește diferite accepții care oscilează după definițiile menționate de către François Rastier [2004], între : «une collection de données langagières, un échantillon de langage»[1] - «ensemble de mots (ou *sac de mots*)»[2], - «un ensemble d'énoncées (ou *sac de phrases*)»[3], - «un ensemble de textes, de segments de texte, de phrases»[4]. François Rastier [2004] alege o perspectivă diferită asupra noțiunii de corpus, «Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications.»[5]

Remarcăm că pentru John Sinclair (considerat un reper important pentru cercetările din spațiul francez), din punct de vedere lingvistic, corpusul este «une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon au langage»[6] [Sinclair, 1996 : 4, citat și tradus de către Benoît Habert, 2000 : 11].

Ne raliem opiniei lui Jacques Guilhaumou [2002 :6] conform căruia, corpusul desemnează un ansamblu determinat de texte pe care se aplică o metodă de cercetare definită. În științele limbii, în accepția lui Jean-Philippe Dalbera [2002]: «un corpus est un ensemble d'éléments sur lequel se fonde l'étude d'un phénomène linguistique.»[7]

În lingvistica românească, autorii articolului „*Argument pentru necesitatea lucrului pe corpus*" [1996: 163-168], Ana Maria Barbu, Dan Tufiș, Călin Diaconu, Lidia Diaconu definesc conceptul de *corpus* ca pe o colecție de texte scrise sau rostite, care este memorată și prelucrată pe calculator în scopul cercetării lingvistice:

> „Mai precis, un corpus este o colecție computerizată de texte autentice adecvată prelucrării sau analizei automate sau semi-automate. Textele sunt selectate potrivit unor criterii explicite cu scopul de a capta regularitățile unei limbi, ale unei varietăți a limbii sau ale unui sublimbaj. Textele conținute într-un corpus sunt păstrate într-un format comun, accesibile ca și cum ar alcătui un singur șir de caractere; totuși, se consideră important ca textele să nu-și piardă identitatea și ca, în orice moment, sursa originală a unei porțiuni de limbă date să fie accesibilă analistului" [Tufiș, 1996].

Având în vedere aceste definiții și elementele evocate în ele, precizăm că, pentru noi, „datele lingvistice" constituite în corpus sunt reprezentate de o colecție de texte ilustrative pentru genul discursiv al ghidului turistic și considerăm că studierea specificității ghidului turistic în calitate de gen discursiv poate să se producă utilizând cadrul teoretic și metode ale lingvisticii de corpus, pe care le detaliem în

---

[1] „o colecție de date de limbaj, un eșantion de limbaj " (tn)
[2] „ ansamblu de cuvinte (sau *sac de cuvinte*)" (tn)
[3] „un ansamblu de enunțuri (sau sac de fraze)" (tn)
[4] „un ansamblu de texte, de segmente de text, de fraze" (tn)
[5] „Un corpus este o regrupare structurată de texte integrale, documentate, eventual îmbogățite prin etichetare și adunate : (i) în manieră teoretică reflexivă ținând cont de discursuri și de genuri, și (ii) în manieră practică în vederea unei game de aplicații." (tn)
[6] „o colecție de date de limbaj care sînt selectate și organizate după criterii lingvistice explicite pentru a servi drept eșantion limbajului". (tn)
[7] „un corpus este un ansamblu de elemente pe care se fondează studiul unui fenomen lingvistic" (tn)

cele ce urmează. Cum vom vedea în continuarea articolului, corpusul nostru urmărește criteriile definite mai sus de autorii menționați, precum caracterul computerizat, omogeneitatea din punct de vedere al genului, pentru a permite studierea limbajului ghidului turistic. Vom observa însă și anumite limite derivând din caracterul comparativ al studiului nostru și anume, comparabilitatea limitată a datelor între limba română și limba franceză.

După cum precizează cercetătorii Marcel Cori, Sophie David și Jacqueline Léon [2008], sintagma «*linguistique(s) de corpus*» a fost împrumutată curentului britanic *Corpus Linguistics*, idee susținută și de denumirea dată de Benoît Habert et alii [1997] acestei lucrări importante pentru spațiul de expresie franceză, *Les linguistiques de corpus*.

În opinia lui Wolfgang Teubert [2009], lingvistica de corpus a fost dezvoltată pentru a extrage dintr-un corpus cunoștințele lingvistice necesare predării limbilor, iar în cadrul acestei discipline, care aduce un nou mod de a privi limbajul, semnificația este situată mai degrabă în discurs, în interacțiunea dintre oameni, decât în mintea locutorilor[8] : «Seuls les mots pris dans leur contexte ont du sens, et ce qu'ils signifient est déterminé par leurs collocats contextuels.»[9] [Teubert 2009]. Teubert precizează că lingvistica de corpus a fost creată pentru a furniza descrieri mai bune a limbilor naturale, a variantelor lor, și, ceea ce o deosebește de alte teorii lingvistice, este faptul de a propune o abordare de tipul *bottom-up* : «c'est à dire ascendante, ou encore de bas en haut, à l'inverse d'une approche *top-down,* analytique ou du haut vers le bas.» [10] [Teubert 2009]

Ne raliem opiniei lui Wolfgang Teubert [2009], privitoare la abordarea de tipul *bottom-up*, care ar fi proprie lingvisticii de corpus:

> «La linguistique de corpus ne tient aucune catégorie descriptive pour acquise, non plus que les mots, les parties du discours, les composants de la grammaire de constituants, ou les traits sémantiques. Elle commence par interroger l'évidence et la mettre à l'épreuve. Quels phénomènes linguistiques trouve-t-on dans un corpus, quels éléments présentent une tendance à la co-occurrence, quels conditionnements sont observables ? Alors seulement la linguistique de corpus se demande comment les données peuvent être catégorisées.»[11] [Teubert 2009]

Constatăm că funcția acordată corpusului, respectiv, un simplu rezervor de exemple sau materialul sursă pentru descriere, poate fi diferită, după distincția *corpus-driven* versus *corpus-based*, utilizată în tradiția anglo-saxonă. După cum remarcăm, Elena Tognini-Bonelli [2001] detaliază cele două tipuri de abordări : *corpus-based* (în care, cercetătorii utilizează corpusul pentru pentru a confirma sau a infirma o ipoteză) și *corpus-driven* (o abordare inductivă a corpusului, în care cercetătorii preferă explorarea datelor fără a avea idei preconcepute).

Ne raliem opiniei lui Damon Mayaffre [2005], care crede că cele două abordări ale corpusului, deductivă și inductivă, stau față în față și se pot, uneori, completa. Deși opțiunea sa metodologică în privința tratării corpusului este inductivă, cercetătorul, plecând de la o descriere exhaustivă și sistematică

---

[8] « La linguistique de corpus a été développé pour extraire d'un corpus les connaissances linguistiques nécessaires à l'enseignement des langues ; un corpus étant un ensemble collecté et ordonné de données langagières réelles. Faisant partie de la linguistique appliquée, la linguistique de corpus n'a jamais revendiqué réellement de cadre théorique propre. Cependant, elle a procuré une manière nouvelle de regarder le langage. La linguistique de corpus situe la signification dans le discours, dans l'interaction entre les gens, plutôt que dans l'esprit des locuteurs. » [Teubert 2009]

[9] „Doar cuvintele luate în contextul lor au sens, și ceea ce semnifică este determinat de colocațiile lor contextuale." (tn)

[10] „adică ascendentă, sau de jos în sus, în opoziție față de o abordare top-down, analitică sau de sus în jos." (tn)

[11] „Lingvistica de corpus nu are nicio categorie descriptivă dobândită, nu mai mult decât cuvinte, părți de vorbire, componente ale gramaticii elementelor constitutive sau trăsături semantice. Ea începe prin a pune la îndoială dovezile și a le testa. Ce fenomene lingvistice se găsesc într-un corpus, ce elemente au tendința de a se produce în coocurență, ce condiționări sunt observabile? Abia atunci lingvistica de corpus se întreabă cum pot fi clasificate datele." (tn)

a unităților materiale lingvistice ale corpusului, își organizează parcursul interpretativ de jos în sus, însă nu exclude o constantă și necesară întoarcere la text:

> «Pour les uns, le corpus est un observatoire d'une théorie *à priori*, pour les autres, le corpus est un observé qui permet de décrire puis d'élaborer des modèles *à posteriori*. Théorie et empirie, déduction et induction, linguistique de la langue et celle de la parole..., en ce moment, l'épistémologie fondamentale de la discipline se joue et se rejoue, parfois avec naïveté, parfois avec force, dans la réflexion sur les corpus.»[12]

În opinia noastră, cele două tipuri de abordări ale corpusului *corpus-based* și *corpus-driven* sunt complementare, precum și abordarea cantitativă și cea calitativă. Abordarea cantitativă ajută la definirea ipotezelor de cercetare, utilizând date măsurabile, iar abordarea calitativă sprijină formularea chestiunilor de cercetare specifice, utilizând analiza conținutului, de aceea, vom opta pentru o abordare mixtă.

În urma investigațiilor întreprinse, constatăm că, în spațiul cercetărilor lingvistice românești preocupările pentru exploatarea corpusurilor cu mijloace informatice nu reprezentau un fenomen de o amploare deosebită, la sfârșitul secolului al XX-lea și începutul secolului al XXI-lea [Cherata, Vușcan, Tămâianu, 1994; Barbu, Tufiș, Diaconu, Diaconu, 1996; Tufiș, Barbu, Pătrașcu, Rotariu, Popescu,1997; Tămâianu-Morița, Cherata, Vîlcu, 2006; Tufiș, Ion, Ceaușu, Ștefănescu, 2011; Mărănduc, 2011]. După anul 2000, se pune problema dezvoltării tehnologiilor informatice pentru a studia limba română [Cristea, Tufiș, 2002; Tufiș, Filip, 2002] și, în octombrie 2001, a fost creată în Academia Română, *Comisia de Informatizare pentru Limba Română* (CILR) și *Consorțiului de Informatizare pentru Limba Română* (ConsILR), cu scopul de a apăra identitatea limbii române prin promovarea studiilor dedicate ei, dintr-o perspectivă informațională.

După cum remarcăm, în perioada actuală, lingvistica românească este racordată la cea internațională și cercetătorii români [Coman, Mitrofan, Tufiș, 2019; Păiș, Ion, Tufiș, 2020] creează sau adaptează instrumente pentru studierea limbii române: aplicații informatice de adnotare automată, aplicații de extragere a informaților și de regăsire a informației în text etc, respectând standardele care reglementează la nivel internațional aceste activități, cu scopul de a conferi vizibilitate cercetării lingvistice românești.

3. Caracteristici ale corpusului. Prezentarea corpusului de lucru

Corpusul de lucru propus este reprezentat de ghiduri turistice generale despre România, originale, imprimate, în română și în franceză, publicate după anul 2000. Așa cum am constatat, în urma cercetărilor întreprinse, există un decalaj important între cele două spații culturale, în ceea ce privește evoluția ghidurilor turistice, în general, și a celor despre România, care reprezintă obiectul nostru de studiu, în particular. Precizăm că am selectat doar ghiduri originale în română și în franceză, traducerile nu fac obiectul studiului nostru. Decalajul se explică prin faptul că s-au scris și publicat mai puține ghiduri generale în România, deoarece cele mai multe publicații recente sunt despre regiuni sau orașe, obiective tematice.

Ghiduri în **limba română:**				
Titlul:	Cod:	Anul:	Editura și locul apariției:	Nr. pagini
România : ghid turistic	**RoGT 2007**	2007	Editura Ad Libri, București	96 p.
România	**RoGT 2015**	2015	Editura Ad Libri, București	184 p.
Ghiduri în **limba franceză:**				
Guides Bleus Évasion. Roumanie	**GB 2004**	2004	Hachette Tourisme, Paris	336 p.

[12] „Pentru unii, corpusul este un observator al unei teorii a priori, pentru alții, corpusul este un observat care face posibilă descrierea și apoi dezvoltarea modelelor a posteriori. Teorie și empirism, deducție și inducție, lingvistică a limbajului și cea a vorbirii ..., în acest moment, epistemologia fundamentală a disciplinei este jucată și rejucată, uneori cu naivitate, alteori cu forță, în reflecția asupra corpusurilor.” (tn)

Guide Vert Roumanie Michelin	**GV 2008**	2008	Michelin Propriétaires-Éditeurs, Clermont-Ferrand	360 p.
Le Petit Futé Roumanie	**GPF 2018**	2018-2019	Nouvelles Éditions de l'Université / Dominique Auzias &Associés, Paris	528 p.
Le guide du routard. Roumanie	**GR 2018**	2018	Hachette Tourisme, Paris	342 p.

Tabelul 1 – Prezentare corpus de lucru

Este de menționat în legătură cu cele două ghiduri românești, notabile pentru analiza noastră, publicate în 2007 și, respectiv, în 2015, că autoarea textului, pentru ambele ghiduri, este Mariana Pascaru. Considerăm că al doilea ghid al autoarei poate a fi privit ca o revizuire și o dezvoltare a primului ghid și, din aceste motive, trebuie să ținem cont în analize, de aceste limite importante în accepția conceptului de corpus, aplicate textelor în limba română.

Macrostructura ghidului din 2007, *România: ghid turistic*, cuprinde 3 capitole introductive, cu informații generale despre țară, cadrul natural și date istorice, după care urmează descrierea și prezentarea obiectivelor incluse în patrimoniul mondial al UNESCO. Observăm că itinerariile detaliate, propuse ca mod de a parcurge spațiul României, intitulate „*Trasee românești*", nu utilizează un criteriu unitar de clasificare, al regiunilor, după cum se poate remarca în denumirile date: *„Marea Neagră", „Valea Prahovei", „Maramureș", „Bucovina", „Culoarul Rucăr-Bran", „Castelul Bran", „Cazanele Dunării", „Oltenia de Nord", „Munții Apuseni", „Transfăgărășan", „Ținutul Neamțului".* Ghidul din 2007 se încheie cu un capitol intitulat *România - repere citadine*. Având mai puțin de 100 de pagini, cu ilustrații și harta țării, detaliată pe secțiuni la finalul ghidului, în acest ghid, nu găsim nicio informație practică referitoare la orarul de vizitare al siturilor și muzeelor, nicio indicație despre posibilitățile de a lua masa sau de a te caza.

Compararea celor două ghiduri în limba română ne oferă posibilitatea de a observa că ghidul turistic, *România,* publicat în 2015, vine cu o viziune nouă asupra structurării textuale a informațiilor referitoare la dimensiunea istorică, geografică și culturală: macrostructura textului este organizată diferit, în sensul că, în capitolul introductiv, intitulat *Bine ați venit în România!,* pe lângă reperele istorice și datele generale, au fost introduse, 2 capitole inedite pentru un ghid românesc, respectiv, *Cele mai frumoase 10 locuri din România* și *Cele mai interesante 10 experiențe românești*. Ghidul propune informații fundamentale și detaliate despre fiecare regiune, la care se adaugă și detaliile practice necesare pentru a vizita diferite obiective: telefon, orar de vizitare, prețul biletului, căi de acces, dar și informații de tip anecdotic, în secțiunea *Știați că?* Latura practică a ghidului este pusă în valoare și în ultimul capitol, *Recomandări și informații utile*.

Corpusul nostru de lucru conține cele patru ghiduri originale franceze contemporane, (altele care mai există, precum *Guide Gallimard* și *Lonely Planet* reprezintă traduceri și acestea nu fac obiectul studiului nostru), printre cele mai cunoscute și cumpărate în Franța și care, după cum propune Mariagrazia Margarito [2004], pot fi clasificate în două mari categorii: *culturale* (GB 2004 și GV 2008) și *practice* (GR 2018 și GPF 2018).

Ghidurile culturale propun, în viziunea autoarei menționate, cunoștințe mai detaliate despre istorie, istoria artei, etnologie, etnografie, literatură, cinema, iar cele practice diminuează cantitativ dimensiunea culturală în favoarea informațiilor practice, absolut necesare pentru călătorie. Însă, aceste frontiere stricte între *ghid cultural* și *ghid practic* se estompează după cum precizează Mariagrazia Margarito, deoarece s-a produs o evoluție după anul 2000, în sensul că și ghidurile culturale au introdus informații practice referitoare la adrese de cazare și pentru a mânca, iar ghidurile practice au îmbogățit secțiunile cu informații culturale.

3.1. Comparabilitatea. Criterii

Specialiștii în lingvistica de corpus dau o definiție destul de restrictivă a comparabilității unui corpus. Pentru unii, de exemplu, pentru Déjean și Gaussier [2002], două corpusuri de limbă L1 și L2

sunt comparabile dacă partajează o cantitate de lexeme care nu este neglijabilă, comparabilitatea fiind deci o proprietate legată de partajarea unui lexic comun.

Pentru Wolfgang Teubert [1996: 245] comparabilitatea se situează la nivelul criteriilor utilizare pentru eșantionarea corpusului și putem ține cont de domeniu de referință, într-o primă fază, la care se pot adăuga: tema, genul textual, perioada de publicare:

> « Des "corpus comparables" sont des corpus en deux langues ou plus composés de façon identique ou similaire. Les textes qu'ils contiennent peuvent être classés selon une variété de traits intralinguistiques et extralinguistiques. Le domaine, par exemple, peut être une caractéristique pertinente pour la composition du corpus. »[13]

În literatura de specialitate sunt prezentate mai multe criterii care pot fi utilizate pentru a calcula gradul de comparabilitate între două sau mai multe texte pe care intenționăm să le regrupăm într-un corpus. Putem efectua această regrupare după criteriile calitative, precizate de către Douglas Biber [1993] și John Sinclair [1996], utilizate în stilistică precum gen, autor, perioadă, mediu, precum și după un număr mare de măsuri cantitative bazate pe frecvența cuvintelor, în opinia lui Adam Kilgariff [2001]. Adesea, aceste criterii pot fi aplicate atât textelor monolingve, cât și celor bilingve.

Criteriul central pentru constituirea unui corpus comparativ este, în esență, credem noi, comparabilitatea dată de apartenența la același gen discursiv, pentru că genurile discursive pot fi supuse unei comparații între două sau chiar mai multe limbi și circumstanțele de comunicare pot fi similare. În plus, am adăuga noi, utilizarea de corpus comparabil permite accesul direct la folosirea reală a lexemelor în fiecare limbă. Acest criteriu nu este însă suficient, deoarece și alte elemente, evocate de John Sinclair [1996], precum perioada publicării și varietatea autorilor au impact asupra organizării textuale și asupra limbajului, constituind criterii de comparabilitate între două limbi.

În etapa de constituire a corpusului, alcătuit din texte originale comparabile, ne-am ghidat după criteriile propuse de către cercetătorul Stig Johansson [1998], specialist recunoscut în domeniul corpusurilor multilingve, care consideră pertinente următoarele aspecte de care trebuie să ținem cont pentru a grupa textele: data publicării, genul textual, domeniul de referință. **Aceste ghiduri sunt comparabile în ceea ce privește genul discursiv și perioada de publicare (după anul 2000), ceea ce ne va permite să observăm ce asemănări și diferențe se pot produce în contexte discursive similare.**

Corpusul nostru de lucru este un ansamblu de date lingvistice, construit, pe de o parte, în funcție de tipologia textelor, și, pe de altă parte, în funcție de coerența textelor. Coerența poate să fie determinată de diverși factori, precum contextul de producere al textelor (data publicării, genul discursiv, domeniul de referință) sau prezența fenomenului lingvistic studiat (particularitățile pragmatico-retorice care evidențiază dimensiunea publicitară a mesajului lingvistic actualizat în ghidurile turistice, de exemplu).

3.2. Limitele corpusului de lucru

În urma investigațiilor întreprinse pentru constituirea materială a corpusului în vederea exploatării automatizate, am constatat că există o disparitate între cele două spații culturale: român și francez, atât în ceea ce privește preocuparea românilor și francezilor pentru călătorii, cât și în ceea ce privește literatura turistică disponibilă. În spațiul românesc, am identificat foarte puține ghiduri turistice originale, generale, despre România, care au apărut după anul 2000 (doar 2 care să poată fi comparate cu ghidurile franceze), pe când, în spațiul francez, cititorii-călători beneficiază de o ofertă foarte largă și editarea de carte turistică reprezintă un segment foarte important pe piața de carte specializată, ghidurile turistice vânzându-se în 10 milioane de exemplare anual.

Dacă dorim să ne bazăm pe frecvența ocurențelor în ceea ce privește faptele de limbă studiate pe un corpus, este important să comparăm texte de talie comparabilă. În ceea ce privește corpusul nostru, suntem tributari faptului că textele în limba română sunt de talie mai mică și, din această cauză, ori de

[13] „"Corpusuri comparabile" sunt corpusuri în două sau mai multe limbi compuse în mod identic sau similar. Textele pe care le conțin pot fi clasificate în funcție de o varietate de caracteristici intralingvistice și extralingvistice. Domeniul, de exemplu, poate fi o caracteristică relevantă pentru compoziția corpusului." (tn)

câte ori va fi posibil, vom utiliza și date procentuale pentru a realiza comparații. Având în vedere acest aspect, rezultatele obținute și interpretarea lor trebuie văzute în limitele celor două corpusuri, francez și român. În ultimul caz, faptul că cele două texte au același autor, Mariana Pascaru, reduce și mai mult posibilitatea oricărei generalizări legate de ghidurile în limba română și ne vom baza pe cele două ghiduri pentru a stabili piste indicative asupra caracteristicilor textelor în română.

3.3. Comparabilitatea în interiorul corpusului francez

Se poate ușor constata că, ghidurile franceze sunt supuse unor constrângeri legate de genul ghidului turistic. Mai precis, observăm că există o schemă de organizare textuală internă, specifică genului ghid turistic, după tipologia propusă de Mariagrazia Margarito [2004]. Indiferent de categoria din care face parte, ghidul francez general, în privința unei destinații, are o *macrostructură* (în termenii lui Adam), [2008: 261] proprie, care ține și de clasificarea lor în colecții de ghiduri.

Macrostructura unui ghid cuprinde capitole și subcapitole care oferă informații generale despre destinație (istorie, geografie, gastronomie, viața rurală, religie, economie, credințe și tradiții etc), informații utile pentru organizarea călătoriei (adrese, mijloace de transport etc), itinerarii propuse, urmărindu-se, în general, un parcurs regional, detaliate în prezentarea localităților și a obiectivelor interesante de vizitat.

Indiscutabil, clasificarea ghidurilor franceze în „ghiduri culturale" și „ghiduri practice" poate fi nuanțată, dar, din punctul nostru de vedere, ea ilustrează foarte elocvent manifestările, mai mult sau mai puțin pregnante, în fiecare categorie, ale caracterului enciclopedic a acestui gen discursiv. Ghidurile culturale (*Guide Bleu* și *Guide Vert)* furnizează mult mai multe informații cu caracter istoric, geografic, cultural și antropologic, urmând un spirit enciclopedic și științific. Aceste ghiduri vizează un public cultivat, avid de erudiție și privilegiază informațiile cu caracter cultural care contribuie la înțelegerea țării pe care cititorul-călător o vizitează iar informațiile practice au o prezență relativ scăzută.

Investigarea comparativă a *Guide du Routard* și *Guide du Petit Futé,* care sunt considerate „ghiduri practice", ne oferă posibilitatea de a constata că aceste ghiduri propun informații generale despre țară, atracții turistice și itinerarii prin țară. În *Guide du Routard*, capitolul *Roumanie utile*, de exemplu, situat în arhitectura textuală a ghidului chiar înainte de prezentarea itinerariilor pe regiuni, demonstrează caracterul practic al ghidului, care se adresează, mai degrabă, călătorului independent, care are nevoie să găsească informații rapide și precise despre găzduire, transport și masă, pentru că acest tip de public nu apelează, de obicei, la serviciile unei agenții de turism.

4. Metodologia obținerii și prelucrării corpusului

- ***Culegerea datelor***

Constituirea materială a corpusului în vederea exploatării automatizate a necesitat o primă etapă de numerizare. Ghidurile turistice menționate au fost scanate și ocerizate cu programul *ABBYY FineReader PDF*[14]. Am considerat preferabil să comparăm doar texte originale iar cercetarea noastră vizează doar materialul textual, ceea ce ne-a ghidat în faza de editare și curățare a datelor lingvistice.

Fișierele scanate sunt supuse ocerizării, proces prin care recuperăm doar textul necesar investigării, salvat apoi în fișiere cu formatul rtf. Ocerizarea cu acest program ne-a permis păstrarea caracterelor cu diacritice, specifice limbii române, deoarece programul recunoaște 198 de limbi, printre care și limba română. Ceea ce este interesant în parametrii de ocerizare ai programului este că putem alege modul de prezentare al documentului final și am ales ca prezentarea textului în pagină să fie identică cu originalul pentru formatul rtf. Această alegere se justifică metodologic prin faptul că permite păstrarea structurii textului original care ne va ajuta în analiza elementelor lingvistice. După aceea,

[14] Programul *ABBY FineReader PDF* este disponibil pe site-ul https://pdf.abbyy.com/fr/ și este prezentat în acest mod: „FineReader PDF permite profesioniștilor să maximizeze eficiența în spațiul lor de lucru digital. Beneficiind de cea mai recentă tehnologie OCR de la ABBYY, alimentată de Inteligența Artificială, FineReader PDF facilitează scanarea, găsirea, modificarea, protejarea sau partajarea tuturor tipurilor de documente și lucrul în colaborare la acestea în același flux digital." (tn)

pentru tratarea automatizată a datelor cu ajutorul programelor specializate, acest format va fi modificat prin convertirea fișierelor în format *Word* și text.

- ***Editarea datelor (curățare, codare, formatare).***

Fișierele obținute (atât cele în limba română cât și cele în limba franceză) au fost curățate în format *Word* de imperfecțiunile care s-au produs în timpul recunoașterii optice a caracterelor și au fost înregistrate în formatul text, utilizat de programul ***IRaMuTeQ*** pentru textele în limba franceză și de concordanțierul ***Lucon*** pentru textele în limba română. Fișierele text au fost salvate cu Unicode UTF 8 pentru păstrarea caracterelor cu diacritice. Aceleași fișiere corespunzătoare fiecărui ghid în limba franceză au fost înregistrate și în format Word, pentru a putea fi investigate cu programul de lectură a corpusului ***Cordial Pro***, care funcționează prin integrarea profundă în suita *Microsoft Office* și înlocuiește corectorul *Word* original dar are posibilități de analiză mult mai performante, prin raportare la dicționarele integrate.

Trebuie să menționăm că, pentru fiecare ghid în limba franceză, fișierele obținute au fost editate cu programul ***Notepad ++,*** pentru a fi aduse în formatul utilizat de programul ***IRaMuTeQ***. Este un format de text brut dar care utilizează câteva convenții de codare simple ce presupun scrierea unei linii de cod la începutul fiecărui fișier global al ghidului sau al fiecărui fișier tematic, linie care trebuie să aibă următoarea formă: **** *Meta1_Val1 *Meta2_Val2 … *MetaN_ValN, în care *Meta_Val declară o metadată a textului, „Meta" este numele dat variabilei calitative, adică identitatea ghidului (GB, GV, GR și GPF), limba în care este scris textul și „Val" este valoarea descriptivă a corpusului GEN (general) sau a subcorpususlui tematic (*economie,* în cazul nostru).

Ghid:		Nr. total de cuvinte:
România : ghid turistic	RoGT 2007	16.070
România	RoGT 2015	70.345
Guides Bleus Évasion. Roumanie	GB 2004	121.294
Guide Vert Roumanie Michelin	GV 2008	203.846
Le guide du routard. Roumanie	GR 2018	162.677
Le Petit Futé Roumanie	GPF 2018	292.437

Tabel 2 – Dimensiunea corpusului scanat și ocerizat

4.1. Instrumente de lucru pe corpus. Programe de analiză

Studiul nostru cantitativ și calitativ se bazează pe exploatarea automatizată a corpusului, cu ajutorul concordanțierului ***Lucon***[15] (doar pentru textele în limba română) și, doar pentru textele în limba franceză, cu ajutorul programului informatic de lectură a corpusului, ***IRaMuTeQ***[16] și cu programul ***Cordial Pro***[17]. Prezentarea care urmează vizează explicitarea particularităților acestor programe și, în acest fel, metodologia de cercetare care va fi aplicată pe corpus.

4.1.1. Lucon

Concordanțierul ***Lucon*** (*Lucene based concordancer*), adaptat pentru limba română de cercetătorul Cătălin Mititelu, permite căutarea de termeni în baza unui index intern. Prin urmare, orice document trebuie mai întâi să fie indexat. Pentru a construi un index, trebuie accesat meniul *Concordanță*, din care se alege opțiunea *Creează index* și se pot seta parametrii incluși în concordanțier. În prezent, *Lucon* poate lucra cu text simplu sau fișiere XML, tipul acestora fiind recunoscut din extensia de fișier. Prin urmare, orice fișier final .xml este indexat ca fișier XML, altfel este indexat ca fișier text simplu (deocamdată, *Lucon* nu funcționează cu fișiere de tipul .doc, .docx. .xsl).

---

[15] Concordanțierul *Lucon* este disponibil la adresa : https://sourceforge.net/projects/lucon/

[16] *IRaMuTeQ (« Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires »*) poate fi accesat în varianta 0.7 alpha 2, de pe adresa de internet : https://sourceforge.net/projects/iramuteq/

[17] Programul ***Cordial Pro*** poate fi accesat contra cost, de pe adresa de internet: https://www.cordial.fr/cordial-pro

Concordanțierul permite căutarea termenilor în corpus și afișează toate formele flexionare ale termenului căutat și numărul de ocurențe pentru fiecare formă. În casetele din dreapta, putem avea acces și la contextele de apariție a termenilor.

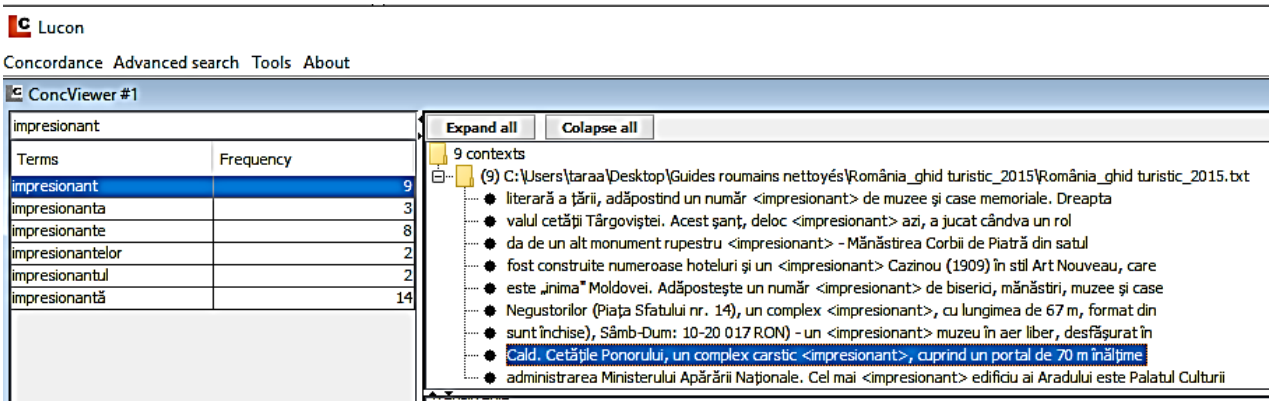

Figura 1 – Afișarea formelor flexionare ale termenului căutat și a numărului de ocurențe

4.1.2. IRaMuTeQ. Tipuri de analiză posibile cu *IRaMuTeQ*

IRaMuTeQ (« Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires ») este un program liber și deschis de analiză de date textuale sau de statistică textuală și se bazează pe limbajele de programare R și Python, dezvoltat în cadrul Laboratorului LÉRASS (Studii și cercetări aplicate în științe sociale), de la Universitatea din Toulouse.

Funcționarea programului constă în pregătirea datelor și în scrierea scripturilor care sunt apoi analizate în programul de statistică R. Rezultatele vor fi afișate, în final, de interfața grafică. Fișierele pe care le supunem analizei trebuie sa fie în format text brut (.txt) și să conțină semnele de punctuație. Prima linie a textului trebuie să fie codată cu variabile descriptive ale textului și modalități și nu trebuie să conțină alte caractere.

Programul ne dă posibilitatea de a alege parametrii de indexare ai textului și am optat pentru parametrul UTF 8 utilizat pentru codarea textului. ***IRaMuTeQ*** transformă tot textul în minuscule pentru a nu diferenția cuvântul scris cu minusculă, de același cuvânt scris cu majusculă, la început de propoziție.

Programul ***IRaMuTeQ*** propune diferite tipuri de analize (Figura 13) bazate pe : lexicometrie (*Statistici*), metodele statistice (*Calcularea Specificităților*, *Analiza factorială* sau *Clasificări*), vizualizarea datelor textuale (*Norul de cuvinte*) sau analiza rețelelor de cuvinte (*Analiza similitudinilor*). (tn) Dintre aceste tipuri de analize, am ales să ilustrăm cu propriile capturi de ecran care au rezultat în timpul testării programului și să prezentăm pe cele pe care le utilizăm în cercetarea noastră: Statistici, Calcularea specificităților.

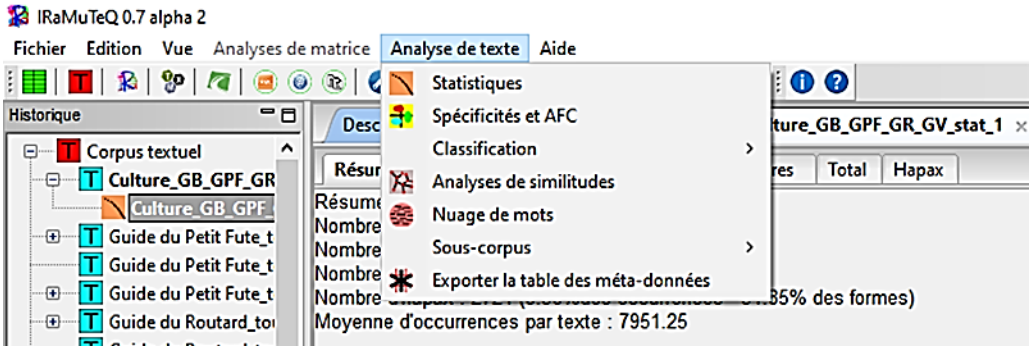

Figura 2 – Tipuri de analize posibile cu IRaMuTeQ

A) Statistici textuale

Această analiză propune statistici simple pe corpus textual : efectivul tuturor formelor (numărul total de ocurențe, care poate fi diferit, în funcție de faptul dacă textul este lematizat[18] sau nu) (Tufiș, Ceaușu, Ștefănescu, 2007), efectivul termenilor activi (cuvintele autosemantice) și suplimentari (cuvintele sinsemantice) dar și lista termenilor clasați în categoria hapax.

Lista de termenilor autosemantici, cu numărul de ocurențe pentru fiecare termen și categoria lor gramaticală, este disponibilă. Termenii autosemantici sunt numiți de *IRaMuTeQ termeni activi* și programul încadrează, în mod implicit, în această categorie: substantive, adjective, verbe și adverbe. Tot sub formă de tabel sunt disponibili și *termenii suplimentari* și termenii clasați în *hapax*. Cuvintele sinsemantice sunt numite de program *termeni suplimentari* și în această categorie se regăsesc, în mod implicit: articolul hotărât și cel nehotărât, adjectivele pronominale posesive și cele interogative, pronumele posesive, demonstrative și cele relative, conjuncțiile, onomatopeele și prepozițiile. Formele de *hapax* reprezintă cuvintele care nu apar decât o singură dată în corpusul textual.

Opțiunile suplimentare ale analizei statistice textuale ne permit, pentru fiecare formă lematizată (cu un click dreapta pe forma respectivă), afișarea formelor asociate lemei respective și a concordanțierului cu segmentele de text în care figurează forma respectivă.

B) **Calcularea *Specificităților* și *Analiza factorială a corespondențelor***

Această analiză permite identificarea cuvintelor specifice dintr-un fișier, clasate în subcategorii și realizează o analiză factorială a corespondențelor sub formă de tabel comparativ, definit cu variabilele și modalitățile specificate de noi (Figura 5). De exemplu, în tabelul încrucișat următor, din Figura 5, putem observa pe un subcorpus tematic, *cultură*, construit de către noi, originar din cele 4 ghiduri franceze, care sunt cuvintele cele mai frecvente în fiecare ghid, referitoare la acest subcorpus tematic: *«roumain»* (*român*) înregistrează numărul cel mai mare de ocurențe în *Guide du Petit Futé*, 109 și 78 de ocurențe în *Guide Vert*. Fără îndoială, în legătură cu arta religioasă și cu bogatul patrimoniu din acest domeniu, lexemul *«église»* (*biserică*) are maximum de ocurențe – 27, în *Guide Bleu* etc.

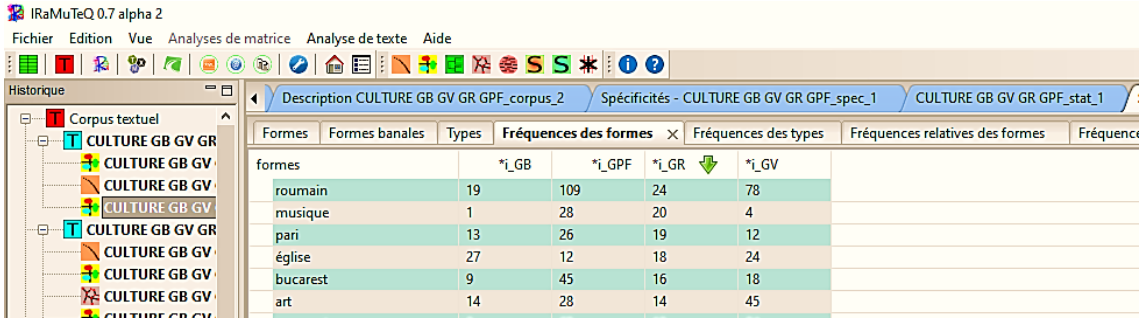

Figura 3 – *Specificități* și *Analiza factorială a corespondențelor*, tabel comparativ cu termenii activi și numărul de ocurențe specifice temei „*Cultură*” din cele 4 ghiduri franceze

Atunci când am codat fișierele pentru a putea fi investigate cu programul, am scris linia de cod necesară la începutul fiecărui text, care conține variabile calitative și modalități, precedate întotdeauna de 4 asteriscuri (****). Cele 3 variabile necesare, stabilite de către noi, au fost următoarele: ***i_GB**, de exemplu, pentru care **i reprezintă identitatea ghidului** (vom mai avea și GV, GR, GPF, adică cele 4 ghiduri franceze investigate), urmată de ***l_FR,** adică **l reprezentând variabila care specifică limba** în care este scris textul și cea de a treia variabilă ***t_GENERAL** sau ***t_ECONOMIE, t reprezentând tematica ghidului,** respectiv, un ghid general sau un subcorpus tematic. Modalitățile stabilite reprezintă identitatea ghidului: ***GB** (*Guide Bleu*), ***GV** (*Guide Vert*), ***GR** (*Guide du Routard*) și ***GPF** (*Guide du Petit Futé*). Aceste modalități și variabile calitative, precedate de asterisc și urmate de de numele variabilei, ne vor permite studierea corpusului în două feluri: fie studiem fiecare text individual, fie

[18] „Lematizarea cuvintelor necunoscute este un proces statistic, bazat pe reguli induse din lexicoane. Lema pentru un cuvânt necunoscut este aleasă dintr-un set de leme candidat generate cu aceste reguli. Mecanismul de selecție este bazat pe un Model Markov care a fost antrenat pe leme cu aceeași etichetă morfo-sintactică.”

putem crea un fișier global cu toate cele patru ghiduri, ceea ce ne va permite studierea în paralel a textelor și obținerea de tabele încrucișate cu datele supuse investigației, pentru a putea observa similitudini și diferențe pe baza lor. Vom putea selecta toate modalitățile unei variabile sau numai anumite modalități, după cum se observă în figura următoare, Figura 4.

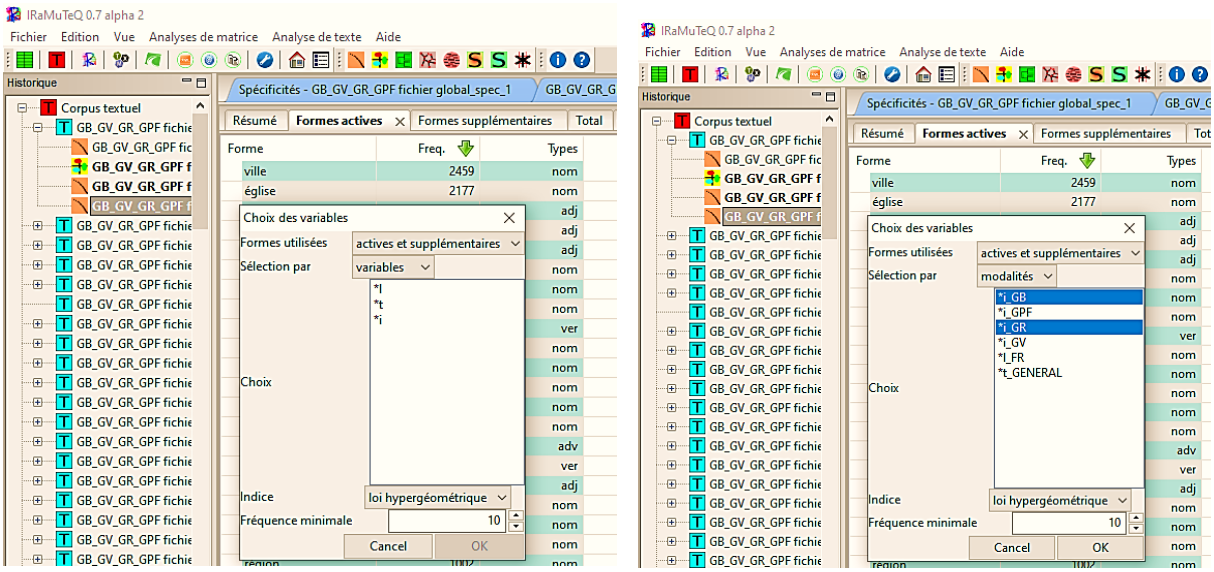

Figura 4 – Alegerea modalităților și a variabilelor pentru parametrarea analizei

De asemenea, putem realiza și reprezentări grafice cu termenii selectați (Figura 5) dintre termenii specifici sau să realizăm o analiză de tipul TGen (crearea de tipuri generalizate). Putem vizualiza specificitățile și putem afișa concordanțierul corespunzător. Precizăm că pentru a genera graficul următor am selectat cuvintele respective din tabelul de mai sus (Figura 3).

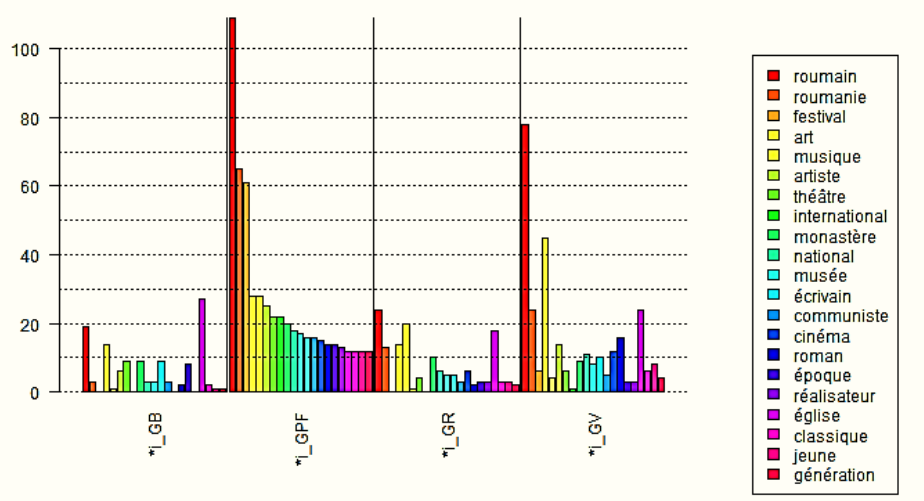

Figura 5 - Reprezentare grafică realizată cu o serie de termeni referitori la „Cultură”, prezenți în cele patru ghiduri

Menționăm un alt aspect foarte important al acestui tip de analize și anume faptul că putem alege din parametrii de lectură a corpusului cheile de lectură, respectiv, prezența sau eliminarea unei anumite clase gramaticale. Vom opta pentru a studia, de exemplu, doar substantivele, doar adjectivele din tot corpusul etc: 0 înseamnă eliminarea clasei respective, 1 = activarea în analiză a clasei gramaticale respective și 2 = forme suplimentare.

4.1.3. Cordial Pro

Programul ***Cordial Pro*** este proprietate a *Synapse Développement*, societate de la Toulouse, specializată în Inteligența Artificială aplicată textelor și tratării automatizate a limbajului. Principalul său punct forte constă în integrarea profundă în suita *Microsoft Office*, programul înlocuiește corectorul

Word original. În total, ***Cordial Pro*** are integrate 15 dicționare diferite, inclusiv dicționarul în limba franceză *Littré*, precum și antonime, abrevieri și semne, rime etc. Software-ul efectuează analize tematice, contextuale și semantice, furnizează statistici (duplicate, propoziții excesiv de lungi, repetări etc.) și semnalează erori tipografice. Menționăm că programul de lectură a corpusului ***Cordial Pro*** integrează și clasificarea textelor în funcție de genuri discursive și, plecând de la un fragment extras dintr-un text, permite precizarea apropierii față de alte genuri repertoriate în program. De exemplu, textul supus analizei, poate fi comparat de către program, în privința rezultatelor statistice afișate, cu texte literare, jurnalistice, tehnice, juridice și comerciale.

5. Concluzii

Prezentarea corpusului, a metodologiei și a instrumentelor de lucru, în acest articol, ne-a permis stabilirea coordonatelor pentru o viitoare cercetare pe care cadrul limitat al acestui studiu nu ne permite să o detaliem. Analiza statistică a corpusului de ghiduri turistice va permite, de exemplu, abordarea dimensiunii tematice a ghidurilor, a interdisciplinarității, precum și a adjectivelor și substantivelor evaluative, ca mărci ale subiectivității.

Considerăm că prezentarea diferitele accepții ale termenului corpus și a cadrului teoretic al lingvisticii de corpus ne vor permite să abordăm specificitatea ghidului turistic în calitate de gen discursiv, cu scopul de a studia limbajul utilizat.

Credem că utilizarea instrumentelor informatice ne permite să dăm mai multă vizibilitate rezultatelor cantitative, valabile și pertinente, pentru a le insera în abordarea calitativă. Trebuie să remarcăm că, programele informatice, care asistă munca cercetătorului cu analize automate și calcule statistice, au câștigat notorietate în cercetările lingvistice actuale, în general, însă, alegerile metodologice și interpretarea rezultatelor aparțin cercetătorului.

Semnalăm că, pentru a realiza o analiză cu mijloace informatice, trebuie să ținem seama și de gradul de intervenție al cercetătorului. De exemplu, programul ***IRaMuTeQ*** realizează o analiză automată, care presupune pregătirea de către cercetător a datelor brute, cu scopul de a realiza analize lexicometrice sau statistici textuale în vederea detectării regularităților, specificităților, a termenilor recurenți. Analiza automată cu acest program se interesează, în primul rând, de structura lexicală a textului și fișierele text trebuie formatate de către cercetător, pentru a corespunde exigențelor programului.

În opinia noastră, ceea ce este important a observa, este că programele de tip concordanțier (***Lucon***) au fost special concepute pentru a identifica termeni precișidintr-un ansamblu textual și de a-i localiza, raportând fiecare ocurență la un context de apariție. Ele permit și numărarea ocurențelor și situarea lor în cadrul discursului.

După cum putem remarca, chiar dacă recurgerea la aceste instrumente se dovedește a fi prețioasă pentru a observa mai bine proprietățile lingvistice ale corpusului supus analizei, rezultatele tratării informatice a corpusului nu reprezintă deloc rezultatele analizei. Tabelele lexicale obținute ne pot oferi grile de lectură, dar interpretarea acestor date ne aparține și este legată de problematica de cercetare pe care ne-am stabilit-o.

Se poate constata că instrumentele informatice permit navigarea rapidă între date și degajarea elementelor cantitative, însă, interpretarea lor adecvată revine cercetătorului, acesta este cel care produce rezultate tangibile ale analizei, făcând legătura între rezultatele informatice ale investigării corpusului și problematica de cercetare elaborată.

Bibliografie:

Barbu, Ana Maria, Tufiș, Dan, Diaconu, Călin, Diaconu, Lidia, 1996, *Argument pentru necesitatea lucrului pe corpus*, în Dan Tufiș, (ed), *Limbaj și tehnologie*, Editura Academiei Române, București, pp. 163-168.

Biber, Douglas, 1993, *Using Registered-diversified Corpora for General Language Studies*, Computational Linguistics, 19(2), consulté le 28 avril 2021, disponuble à l'adresse : https://www.aclweb.org/anthology/J93-2001.pdf

Bommier-Pincemin, Bénédicte, 1999, *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse de Doctorat en Linguistique, Université Paris IV Sorbonne, 6 avril 1999, chapitre VII : "*Caractérisation d'un texte dans un corpus : du quantitatif vers le qualitatif", § A "Définir un corpus*", pp. 415-427, consulté le 23 mai 2021, disponible à l'adresse : http://www.revue-texto.net/1996-2007/Corpus/Publications/pincemin_ad_1999.pdf

Bommier-Pincemin, Bénédicte, 2020, *La textométrie en question*, Le Français Moderne - Revue de linguistique Française, CILF (conseil international de la langue française), Linguistique et traitements quantitatifs, 88 (1), pp.26-43. ffhalshs-029020, consulté le 23 mai 2021, disponible à l'adresse : https://halshs.archives-ouvertes.fr/halshs-02902088/document

Cherata, Sanda, Vuşcan, Teodor, Tămâianu, Emma, 1994, *SILEX – Un sistem lexico-morfologic computerizat pentru analiza textelor româneşti*, în Dacoromania, serie nouă, I, 1994-1995, Cluj-Napoca, pp. 201-212, consultat în 27 iulie 2020, disponibil la adresa: http://www.diacronia.ro/ro/indexing/details/A1947/pdf

Coman, Andrei, Mitrofan, Maria, Tufiș, Dan, 2019, *Automatic Identification and Classification of Legal Terms in Romanian Law Texts,* In Proceedings of the 14th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing, Iași, Noiembrie 2019, pp.39-49.

Cori, Marcel, David, Sophie, Léon, Jacqueline, 2008, *Présentation : éléments de réflexion sur la place des corpus en linguistique*, Langages, 2008/3 (n° 171), p. 5-11. DOI : 10.3917/lang.171.0005, consultée le 29 décembre 2019, disponible à l'adresse : https://www-cairn-info.ressources.univ-poitiers.fr/revue-langages-2008-3-page-5.htm

Cristea, Dan, Tufiș, Dan, 2002, *Resurse lingvistice româneşti şi tehnologii informatice aplicate limbii române*, consultat în 10 noiembrie 2019, disponibil la adresa: http://www.philippide.ro/Identitatea%20limbii%202002/16_Cristea.pdf

Dalbera, Jean-Philippe, 2002, *Le corpus entre données, analyse et théorie*, Corpus [Online], 1 | 2002, Online since 15 December 2003, connection on 14 January 2020. URL : http://journals.openedition.org.ressources.univ-poitiers.fr/corpus/10

Déjean, Hervé, Gaussier, Éric, 2002, *Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables*, Lexicometrica, Alignement lexical dans les corpus multilingues, pages 1-22, consulté le 28 avril 2021, disponible à l'adresse : http://lexicometrica.univ-paris3.fr/thema/thema6.htm

Guilhaumou, Jacques, 2002, *Le corpus en analyse de discours : perspective historique*, Corpus [En ligne], 1 | 2002, mis en ligne le 15 décembre 2003, consulté le 29 décembre 2019. URL : http://journals.openedition.org.ressources.univ-poitiers.fr/corpus/8

Habert., Benoît, Nazarenko, Adeline, Salem, André, 1997, *Les linguistiques de corpus*, Paris, Armand Colin et Masson.

Habert, Benoît, 2000, *Des corpus représentatifs : de quoi, pour quoi, comment?* In M. Bilger (Ed), Cahiers de l'Université de Perpignan, 31, *Linguistiques sur corpus. Études et réflexions*, (pp. 11-58), Perpignan, Presses universitaires de Perpignan.

Johansson, Stig, 1998, *On the role of corpora in cross-linguistic research*, in Stig Johansson, Signe Oksefjell, eds., Corpora and Cross-Linguistic Research: Theory, Method, and Case Studies, Amsterdam-Atlanta, Rodopi.

Kilgariff, Adam, 2001, *Comparing Corpora, International Journal of Corpus Linguistics*, lu le 28 avril 2021, disponible à l'adresse : https://www.researchgate.net/publication/263252975_Comparing_Corpora

Margarito, Mariagrazia, 2004, *Quelques configurations de stéréotypes dans les textes touristiques*, dans Baider, Fabienne, *La Communication touristique : approches discursives de l'identité et de l'altérité*, L'Harmattan, pp 117-132.

Mayaffre, Damon, 2005, *Rôle et place du corpus en linguistique. Réflexions introductives* in Pascale Vergely, Actes du colloque JETOU'2005, Université de Toulouse-Le Mirail, pp.5-17, 2005. ffhal-00553742, Consultat în 24.01.2020, disponibil la adresa : https://hal.archives-ouvertes.fr/hal-00553742/document

Mărănduc, Cătălina, 2011, *Continuitate şi sincronizare terminologică în Gramatica Academiei, ediţiile din 2005−2008*, consultat în 10 noiembrie 2019, disponibil la adresa: http://www.philippide.ro/Metafore%20ale%20devenirii_2011/07.%20C.%20Maranduc.pdf

Păiș, Vasile, Ion, Radu, Tufiș, Dan, 2020, *A Processing Platform Relating Data and Tools for Romanian Language, în Proceedings of the LREC 2020 Workshop IWLTP 2020 – 1st International Workshop on Language Technology Platforms*, consultat în 27 iulie 2020, disponibil la adresa: https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/IWLTP2020book.pdf

Rastier, François, 2004, *Enjeux épistémologiques de la linguistique de corpus.* Texto ! [en ligne], Rubrique Dits et inédits, consultée le 29 décembre 2019, disponible sur : http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html

Sinclair, John, 1996, *EAGLES. Preliminary Recommendations on Corpus Typology*, consultat în 24 ianuarie 2020, disponibil la adresa http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html

Tămâianu-Morița, Emma, Cherata, Sanda, Vîlcu, Cornel, 2006, *Analiza sintagmatică a textelor românești prin mijloace informatice: Proiectul Siastro*, consultat în 27 iulie 2020, disponibil la adresa: http://www.dacoromania.inst-puscariu.ro/articole/2006-2007_2.pdf

Teubert, Wolfgang, 1996, *Comparable or Parallel Corpora?* In *International Journal of Lexicography*, 9 (3): 238-264.

Teubert, Wolfgang, 2009, *La linguistique de corpus : une alternative [version abrégée],* Semen [Online], 27 | 2009, Online since 01 April 2009, connection on 23 January 2020. URL : http://journals.openedition.org.ressources.univ-poitiers.fr/semen/8914

Tognini Bonelli, Elena, 2001, *Corpus Linguistics at Work,* Amsterdam, Benjamin, pp. 65-83 și pp. 84-100.

Tufiș, Dan, Barbu, Ana Maria, Pătrașcu, Vasile, Rotariu, Georgiana, Popescu, Camelia, 1997, *Corpora and Corpus-Based Morpho-Lexical Processing*, in Tufiș, D., Andersen, P. (eds.), *Recent Advances in Romanian Language Technology*, București, Editura Academiei, pp. 35–56, consultat în 27 iulie 2020, disponibil la adresa: https://www.academia.edu/29488769/Corpora_and_Corpus-Based_Morpho-Lexical_Processing

Tufiș, Dan, Filip, Florin (coordonatori), 2002, *Limba Româna în Societatea Informațională - Societatea Cunoașterii*, București, Editura Expert, consultată în 15 august 2020, disponibilă la adresa: https://www.researchgate.net/profile/Dan_Tufis/publication/228382091_Limba_Romana_in_Societatea_Informationala-Societatea_Cunoasterii/links/0fcfd50fa496d56936000000.pdf

Tufiș, Dan, Ion, Radu, Ceaușu, Alexandru, Ștefănescu, Dan, 2007, *Servicii Web lingvistice ale ICIA*, în *Lucrările atelierului Resurse lingvistice și instrumente pentru prelucrarea limbii române,* Iași, 14-15 decembrie 2007, accesat în 30 ianuarie 2020, disponibil la adresa : https://hobbydocbox.com/Sci_Fi_and_Fantasy/65981442-Lucrarile-atelierului-resurse-lingvistice-si-instrumente-pentru-prelucrarea-limbii-romane-iasi-decembrie-2007.html