ȘCOALA DOCTORALĂ INTERDISCIPLINARĂ
Facultatea de Litere

Andreea GHIȚĂ

Calitatea produsului semantic al inteligenței artificiale în traducerea textelor de vulgarizare științifică

Rezumat

Coordonatori de Doctorat

Prof. univ. habil. Alexandru MATEI – Universitatea Transilvania Brașov
Prof. univ. habil. în cotutelă Sonia BERBINSKI – Universitatea București

BRAȘOV, 2025

Cuprins

1. Introducere

Traducerea automată (TA), mecanizată sau mecanică care există încă din 1950 a încercat dintotdeauna să « înțeleagă » limba scrisă sub toate aspectele sale cu ajutorul instrumentelor informatice. În zilele noastre, asistăm la nașterea de programe de prelucrare a limbajului extrem de puternice și fiabile în ceea ce privește calitatea. Acestea merg dincolo de simpla încorporare a regulilor de limbă sau a altor modele de procesare care realizează calcule la nivel de unități lexicale, sintagme și chiar propoziții.

Noutatea în acest domeniu este reprezentată de traducerea automată neuronală realizată de o mașină de traducere neuronală (MTN) la baza căreia stă modul de funcționare al rețelelor neuronale proprii creierului uman, după cum ne sugerează însăși denumirea sa. Mecanismul se servește de numeroase procesoare ce creează legături conform modelului rețelelor neuronale umane. Prin punerea în aplicare a unui proces de învățare independent, cu ajutorul inteligenței artificiale, mașina generează traducerea.

Această tehnologie nu se mai rezumă doar la decuparea propozițiilor pentru a ajunge la decriptarea limbii sursă, după modelul predecesorilor săi. Ea ia în considerare propoziția în integralitatea sa, putând astfel să genereze o traducere îmbunătățită din punct de vedere al semnificației. Ce este, însă, și mai important este faptul că acest software de traducere se consolidează autonom, trasând căi neuronale artificiale care guvernează procesul de învățare. Astfel că, în timp, el devine din ce în ce mai puternic și performant pe măsură ce îi sunt furnizate date.

Este prin urmare, imperativ să se stabilească dacă există deja indicii care să permită anticiparea unui scenariu conform căruia într-o zi, acest tip de programe nu vor mai avea nevoie de supraveghere și intervenție umană. Analiza lingvistică a gradului de calitate semantică a textului livrat în limba țintă este maniera prin care putem respinge sau admite valoarea de adevăr a acestei ipoteze. Acest lucru ne-a determinat să punem sub lupă greșelile semantice comise de Google Translate, respectiv de mașina sa neuronală. Apreciem că realizarea acestei cercetări este esențială, dat fiind faptul că această MTN a ieșit pe piață la finalul anului 2016, lucru care semnifică că ea se află deocamdată la un stadiu precoce de evoluție ; mașina avea vârsta de 3 ani în momentul în care corpusul nostru a fost alimentat în sistemul său.

Așadar, cercetarea noastră se încadrează printre primele cercetări lingvistice ce au drept subiect calitatea traducerii neuronale. Mai mult, abordarea pe care o vom urma pentru a răspunde la obiectivele de cercetare ale tezei sunt de inspirație cognitivă. În plus, trebuie să notăm că tendința predominantă este ca orice metodologie cognitivă, indiferent de domeniu, să răspundă la nevoile IA.

În consecință, am încercat să exploatăm acest tip de traducere automată dotată cu inteligență artificială prin intermediul unor instrumente care să se ghideze după principii similare. Sub acest aspect, teza noastră poate deveni o lucrare de referință atât pentru subiectul abordat, cât și pentru metoda de analiză întrebuințată.

2. Contextul actual al cercetării

În primul rând, ținem să precizăm cu mândrie că teza noastră se înscrie într-un context actual de cercetare și acest lucru se explică prin două axe principale. Prima are legătură cu traducerea obținută prin intermediul GTNM și a doua este legată de instrumentele lingvistice ce vor fi utilizate pentru atingerea scopului cercetării noastre.

Primul argument în ceea ce privește relevanța cercetării noastre constă deci, în faptul că tehnologia de TA a Google pe care ne propunem să o evaluăm este echipată cu inteligență artificială. De fapt, acest apelativ de « inteligență artificială » pare că ne-a invadat cotidianul. Orice tehnologie nouă încearcă să adopte mecanisme inteligente.

Al doilea aspect care face ca teza noastră să fie de actualitate este folosirea unor teorii și principii specifice lingvisticii cognitive, mai ales din semantica cognitivă. Este recunoscut la scară largă faptul că semantica cognitivă este de câțiva ani în prim plan în materie de metodă de analiză a limbajului, în defavoarea semanticii structuraliste.

Aceste două domenii principale care se intersectează în cercetarea noastră, și anume traducerea automată neuronală (TAN) și semantica cognitivă (SC) împărtășesc accentul pus pe a) sensul lexical și b) contextul lexemelor. Acest lucru ne reamintește de perspectiva contextualistă din lingvistica corpusului care postulează că :

> « Fiecare apariție a unei unități lexicale transmite propria sa istorie textuală, un mediu colocațional particular care a fost construit în momentul în care textul a fost creat și care oferă contextul în care unitatea este materializată pentru acea ocazie specială. Acest mediu determină sensul instanțiat, sau sensul textual, al unității, care este unic pentru fiecare utilizare specifică. » (traducerea noastra cf. Halliday & Hasan, 1976, p. 289 în Williams, 2003, p. 35).

Dubreil (2008, p. 13) adaugă ideea conform căreia « cuvântul își derivă sensul din context și influențează simultan acest context pentru a crea mediul textual ». Din acest motiv, analiza sensului unităților lexicale și polilexicale traduse eronat de MTN poate fi efectuată numai prin luarea în calcul a contextului acestora.

Dat fiind faptul că mașina de traducere neuronală își propune să extragă sensul cuvintelor și frazelor pentru a facilita înțelegerea globală a textului, suntem de acord cu ipoteza lui Jakobson (1959) care spune că semnificația trebuie să aibă întâietate în procesul de traducere. Această teorie a fost susținută și de către alți lingviști, încă din momentul apariției TA. Acest aspect ne face să credem că în definitiv, latura semantică și conceptuală a textului ar trebui să fie considerată obiectul cheie al traducerii automate. Și la urma urmei, cum poate fi ea mai bine analizată dacă nu prin teorii conceptuale semantice ce decurg din lingvistica cognitivă ?

3. Ipoteză, întrebări de cercetare și planul tezei

Mereu asediați de știri despre performanțele incontestabile ale inteligenței artificiale în numeroase domenii de activitate, nu putem să nu formulăm o întrebare destul de tulburătoare : Va fi oare posibil ca tot ce înseamnă amprentă umană, îndeosebi comportamentul uman și lingvistic să fie modelat și înglobat într-un algoritm ?

În ceea ce privește domeniul de cercetare studiat de noi, și anume TA neuronală, există temeri că aceasta va înlocui definitiv atât profesia de traducător, cât și pe cea de interpret. Dezvoltatorii de MTN și lingviștii-informaticieni prevăd că în următorii ani nu va mai fie nevoie să angajăm experți lingviști pentru a livra traduceri specializate calitative. Se estimează că într-o zi, traducătorii nu vor mai trebui să intervină în traducerile scrise pentru a le rafina ori perfecționa. Deși astfel de scenarii nu sunt greu de imaginat, nivelul de complexitate al limbii vorbite, limbajului literar și tehnic este destul de ridicat și exact acesta este pariul pe care va trebui să-l câștige IA, asta în cazul în care nu a făcut-o deja…

Conjugând ipoteza de mai sus cu faptul că se dorește să credem că TAN este investită cu capacități semantice de amploare, am construit o întrebare de cercetare pivot : Care este nivelul de calitate al produsului semantic al inteligenței artificiale în traducerea textelor de vulgarizare științifică ?

Conform acestei întrebări legitime, ce se subînțelege încă din titlul tezei, este ușor de dedus faptul că obiectivul nostru principal este acela de a descoperi cât de calitativă este traducerea neuronală pe plan semantic în cazul unui corpus de specialitate. Pentru a ne atinge obiectivul, vom examina în profunzime transferul semantic din franceză în română al lexemelor simple și compuse și al sintagmelor frazeologice.

Întrebarea noastră de cercetare principală încearcă în primul rând să obțină un rezultat de ordin cantitativ, pe care în mod cert îl vom asigura. De fapt, prima etapă a cercetării noastre presupune inventarierea și gruparea tuturor greșelilor de traducere pe categorii, în fișierul Excel de pre-analiză a corpusului aliniat (i.e. Pré-analyse Corpus aligné.xls). Acest lucru va favoriza extragerea unor statistici care să indice procentajul de fiabilitate al sistemului de TAN *Transformer* aparținând Google Translate.

În etapa ulterioară, incontestabil mai importantă, vom încerca să înțelegem ce aspecte au pus dificultate neuronilor artificiali la nivelul traducerii :

1. termenilor și colocațiilor de specialite din corpus ;
2. lexemelor simple și compuse ce fac parte din limbajul general și care pot face parte din construcții mai mult sau mai puțin extinse și fixe ;
3. expresii *encoding* și *decoding* din corpus, a căror construcție depinde în mare parte de fenomenul de metaforizare și de tropii conceptuali.

De asemenea, toată comunitatea lingvistică recunoaște că distingerea sensului, precum și înțelegerea/producerea unui nou lexem, termen, colocație și expresie, presupune folosirea abilităților cognitive umane responsabile de conceptualizare. Ba chiar, lingviștii cognitiviști afirmă că aceasta din urmă este influențată de factori contextuali diverși și de fiziologia umană.

În consecință, pentru a putea descoperi în ce măsură « abilitățile cognitive » ale mașinii sunt similare celor umane pe durata execuției traducerii elementelor lingvistice de la 1 la 3 de mai sus, am conceput întrebările următoare de cercetare :

1. În ce mod și în ce proporție influențează polisemia producerea de greșeli la nivelurile LexS, LexC, Co și Loc ?
2. În ce mod și în ce proporție influențează contextul producerea de greșeli la nivelurile LexS, LexC, Co și Loc ?
3. În ce mod și în ce proporție influențează tropii conceptuali producerea de greșeli la nivelurile LexS, LexC, Co și Loc ?
4. În ce mod și în ce proporție influențează engleza, limba de antrenare a datelor, producerea de greșeli la nivelurile LexS, LexC, Co și Loc ?

După cum ipoteza inițială a servit drept fir conductor pentru formularea întrebărilor de cercetare, aceste întrebări de mai sus au ajutat la elaborarea planului tezei care urmează după acest capitol introductiv :

- Cap. 2 : descrie inițial studiile anterioare în raport cu semantica cognitivă și pune în lumină mai apoi, conceptele și teoriile ce decurg din semantica cognitivă și sunt folosite în executarea analizei (i.e. polisemia în context, metafora, metonimia și integrarea conceptuală), precum și noile perspective asupra construcțiilor frazeologice inspirate de curentul cognitivist ;

  Cap. 3 : detaliază mai întâi contextul istoric al traducerii automate și într-o etapă secundară, situează apariția mașinii neuronale a Google Translate pe scena inteligenței artificiale ;

- Cap. 4 : justifică alegerile metodologice și prezintă totodată corpusul și metodele de analiză ;

- Cap. 5 : interpretează sursele greșelilor care aparțin limbajului general și specializat, atât la nivel de unități lexicale simple și compuse, cât și la nivel de colocații și expresii idiomatice, prin prisma teoriilor menționate la Cap. 2 ;

- Cap. 6 : face cunoscute contribuțiile originale ale tezei și rezultatele obținute în baza tabelelor, figurilor și explicațiilor de la Cap. 5, conectându-le cu întrebările de cercetare ;

- Cap. 7 : explorează posibilitatea realizării unor studii viitoare, în continuitatea celui prezent.

4. Metodologia de cercetare

4.1 Descrierea și fiabilitatea corpusului

Corpusul nostru conține 100 de articole ce au ca temă tehnologii emergente ori a căror apariție este așteptată într-un viitor apropiat la scară globală. Întâi de toate, se impune evidențierea faptului că teza noastră beneficiază de un « corpus paralel și multilingv » deoarece este compus din texte scrise în franceză și echivalentul acestora în română și engleză. Totodată, corpusul este « omogen » datorită faptului că toate textele aparțin aceluiași stil de scriere, și anume celui de vulgarizare științifică (VS). De aceea, articolele conțin detalii cu privire la părțile constitutive ale noilor tehnologii, procesele lor de operare, precum și alte specificații în raport cu ele. Așasar, conform unor reprezentanți ai lingvisticii corpusului, precum Bowker și Pearson (2002, pp.11-13), putem afirma că articolele noastre compun un « corpus de specialitate ». Mai mult, dat fiind faptul că nu vom adăuga alte texte corpusului înainte de finalizarea cercetării, el poate fi considerat de asemenea un « corpus închis ».

Având în vedere că studiul nostru tratează « traducerea automată specializată », suntem nevoiți să aducem în această secțiune detalii despre : a) traducerea specializată (TS) și provocările impuse de aceasta ; b) încrucișarea conceptelor de TS și VS ; c) definiția, rolul și caracteristicile VS ; d) rolul vulgarizatorului și al publicului ; e) cele mai recente schimbări cu privire la modul în care se face VS.

Pentru început, « limba de specialite » sau « limbajul specializat » este supus uei analize profunde încă din anii 60, o dată cu apariția lingvisticii aplicate, în interiorul căreia s-au dezvoltat ariile de expertiză denumite terminologie și lingvistica corpusului. Este important să subliniem că limbajul de specialitate este privit ca un limbaj restricționat, destinat exclusiv experților dintr-un anumit domeniu în opoziție cu limba curentă cu care co-există.

Chiar dacă inițial, se credea că doar profesioniștii puteau înțelege acest limbaj, ca urmare a lexicului său specializat, Müller (1985, p. 187) a introdus noțiunea de « grade de specializare ». Atâta timp cât mesajul, destinatarul și cultura celor ce sunt beneficiarii acestui transfer de cunoaștere diferă, vor exista diferite grade de specializare. Datorită progresului tehnologic, îndeosebi din sfera digitală, accesul la cunoaștere, ce până nu de mult era realizat doar de comunitatea științifică, va fi disponibil tuturor.

În consecință, publicul general, indiferent de subiectul de interes, începe să acceadă la limbajul de specialitate cu ajutorul vulgarizării, limba sa de diseminare. Vulgarizarea științifică poate, de asemenea, purta titulatura de « traducere intralimbaj », datorită tehnicii sale principale, ce constă în reformularea conținutului semantic și a terminologiei, care sunt uneori greoaie pentru cititor. Că vorbim despre TS sau VS, ambele au aceeași funcție, și anume, aceea de a sensibiliza marele public la informații emise de universul științific. În ceea ce privește textele de VS, în mod cert, ele vor prezenta anumite lacune semantice dar care sunt absolut necesare pentru construcția textelor.

După spusele lui Bruno Dufay (2005, p. 33), există patru obiective ale VS : a informa, a explica, a stârni simțul critic și curiozitatea. Lucrul cel mai dificil de făcut de către jurnalistul științific este acela de a-și

atrage cititorul, râmânând obiectiv și rezistând tentației de a produce efectul de senzațional. În zilele noastre, există tendința de a miza pe senzațional pentru a deveni comercial și pentru a vinde. Dufay (2005) enumeră tehnicile ce trebuie aplicate pentru a elabora o vulgarizare calitativă : comparația și metafora, asocierea de subiecte, rezumatul și sinteza, istoria și anecdota, raționamentul logic, dezbaterea contradictorie și punerea sub semnul întrebării, metoda pointillismului, jocul și stratagema, un vocabular accesibil, captarea atenției și surpriza.

În ceea ce privește rolul său, vulgarizarea este esențială în țările din lumea a treia sau în curs de dezvoltare, acolo unde informația științifică lipsește. Știința este cheia dezvoltării și drept urmare, obiectivul central al elitelor care i s-au dedicat ar trebui să fie acela de a vulgariza. Pentru oamenii de știință, această activitate ar trebui să fie o obligație morală, o datorie pe care trebuie să o plătească pentru simplul fapt că au fost înzestrați cu cunoaștere. Ei sunt singurii care pot asigura rectitudinea și rigoarea informației științifice. Cu toate acestea, întrebarea care se pune este dacă cei care fac transmiterea sunt savanți, experți autentici sau cel puțin, persoane cu experiență în domeniul pe care îl relevă și altora. Cu titlu de exemplu, îl putem aduce îm discuție pe William Herschel care credea că există « motive întemeiate că credem că Soarele putea fi locuit » (Meadows, 1986, p. 397). Chiar dacă era lipsită de fundament, această idee absurdă s-a răspândit.

Acest lucru dovedește faptul că « vulgarizatorul » poate avea un rol decisiv în lume. Ar putea chiar să schimbe cursul unei crize umanitare și modul în care ea este percepută de oameni ca urmare a modului de prezentare a faptelor. Să luăm exemple concrete, provocările stringente cu care s-a confruntat omenirea recent : Covid, gripa aviară, gripa porcină, boala vacii nebune, încălzirea climatică, clonarea. Dacă persoana care conduce dezbaterea pe unul dintre aceste chestiuni sau redactează un discurs scris pentru publicul larg este neavizată, acest lucru poate face diferența între supraviețuirea și moartea unei comunități întregi sau a unui popor. Să ne imaginăm că tocmai ar fi ieșit pe piață un nou medicament pentru a trata Covidul sau pentru a eradica cancerul. Pe lângă instituțiile guvernamentale și internaționale de reglementare în domeniul sănătății și farmaceutic, și vulgarizatorul are rolul de a prezenta atât avantajele cât și efectele sale secundare, primejdiile la care se expune populația administrându-i-l. Deținerea unei astfel de expertize este vitală, dat fiind faptul că există din ce în ce mai multe reviste, site-uri Internet, bloguri științifice care pretind că fac vulgarizare științifică. Deducem prin urmare, că nnu este ușor să deosebești grâul de neghină.

Mai mult, există suspiciuni conform cărora un vulgarizator nu mai este capabil să abordeze un subiect în mod imparțial, întrucât acum, accentul nu mai este pus pe « « explicarea » științei, ci pe « mediatizarea faptelor științifice și tehnice care are la bază punerea în scenă de controverse între sfere de activitate ce nu împărtășesc aceleași interese » (trad. noastră cf. Moirand, Reboul-Touré, Pordeus Ribeiro, 2016, p. 144).

Totodată, autorii de mai sus ne informează că asistăm la o modificare a rolului publicului vulgarizării științifice. Dacă la începutul VS, publicul deținea un rol « interpretativ » și « pasiv » prin excelență, astăzi, publicul este un « grup productiv ». Acest fapt se datorează progresului tehnologic și apariției

rețelelor de socializare întrucât în zilele noastre, publicul poate răspunde la articolele scrise pe bloguri științifice, deținute uneori de amatori, sau pe site-uri Internet de vulgarizare trimițând tweet-uri, mesaje pe Facebook sau alte mesaje la poșta cititorilor. Ne aflăm deci, în prezența unui public care a devenit reactiv în relație cu informația care îi este aruncată în față, aaspect favorabil pentru că el îl poate depista pe vulgarizatorul profan.

În ceea ce privește articolele ce compun corpusul notru, putem spune că ne aflăm departe de astfel de suspiciuni. Acest lucru decurge din faptul că vulgarizatorii care scriu pentru revista Science et Vie sunt jurnaliști științifici cu dublă specializare. De fapt, majoritatea au urmat studii de licență și/sau master în domeniul căruia i s-au dedicat. Pentru a completa aceste competențe, aceștia au absolvit, în egală măsură, studii de jurnalism științific, lucru care îi califică, fără echivoc, drept specialiști în materie de VS. Prin urmare, nu putem decât să recunoaștem caracterul veridic și calitativ al textelor acestora. Și categoric, nu putem decât să îi privim ca pe niște adevărați promotori ai descoperirilor științifice. În concluzie, estimăm că articolele acestora valorează aur pentru cercetarea noastră pe planul calității, inventivității, obiectivității, fiabilității surselor ș.am.d.

4.2 Modalități de colectare, format și pregătirea corpusului

Articolele care constituie corpusul nostru au fost extrase atât de pe site-ul revistei Science et Vie cât și din versiunea digitală a acesteia, mulțumită acordului oferit de conducerea revistei al cărei deziderat a fost în mod evident, acela de a sprijini progresul științific din domeniul lingvisticii. Cele 100 de articole pe care le-am colectat, au fost incluse în anexă (i.e. Annexe A– Corpus français) în vederea utilizării acestora în cadrul cercetării noastre.

Articolele care compun Anexa A sunt însoțite în primul rând, de un număr de la 1 la 100, lucru care semnalează ordinea colectării lor. Această notație este urmată de literele « FR » care indică faptul că este vorba despre un text în franceză. Vin dupa aceea, numele autorului, luna precum și anul de publicare al articolului și mai apoi, numărul de caractere cu spații deținut de articolul în cauză. Toate aceste caracteristici se regăsesc în partea stângă a articolului care precede titlul. În ceea ce privește numele vulgarizatorului, acesta poate fi menționat fie în mod integral, fie sub formă de inițiale (i.e. par L.B.). Alegerea îi aparține în mod evident autorului, iar noi suntem nevoiți să informăm cititorul tezei noastre că anumiți vulgarizatori preferă să publice în revistă și pe site sub numele real, inițialele numelui real, un pseudonim sau pur și simplu, sub inițialele pseudonimului lor. Indiferent că publică sub un pseudonim sau nu, toți sunt, indiscutabil, adevărați maeștri ai artei de vulgarizare.

De cealaltă parte, în Anexa B – Corpus român, textele care reprezintă traducerea automată în română a textelor sursă au același număr de ordine, însoțit, de această dată, de abrevierea « RO », o prescurtare de la cuvântul român. Cititorul tezei noastre va putea constata de asemenea, că cele trei elemente de identificare, despre care am discutat în paragraful anterior, nu se regăsesc în celelalte două anexe de corpus. Am considerat că este potrivit să nu includem aceste detalii în relație cu *tertia comparationis* de teamă ca informația să nu devină redundantă.

Dintre cele trei elemente indicate mai sus, unul este extrem de important, și anume, numărul de caractere și spații care funcționează drept indicator al lungimii articolului. Menționarea acestuia contează în mare măsură în cadrul cercetării noastre, fiind în strânsă legătură cu modul de funcționare al mecanismului de traducere al GTNM. Interfața Google Translate permite traducerea unui segment de text ce are 5000 de caractere și spații, aceasta constituind limita superioară. Plecând de la această regulă impusă și într-un efort de a menține coeziunea și omogeneitatea sensului, am decis să nu fragmentăm niciun text de intrare. Așadar, chiar dacă inițial colectasem texte a căror lungime depășea 5000 de caractere și spații, acestea au fost într-un final înlăturate. Pentru ca acest atribut de integralitate să fie transferat și semnificației globale a textului tradus, MTN a procesat texte integrale. Trebuie să ne amintim că GTNM creează și restituie semantica ansamblului frazelor ce alcătuiesc un articol, lucru care justifică alegerea noastră. În ceea ce privește limita inferioară, am estimat că un text cu mai puțin de 500 de caractere și spații nu s-ar putea califica în mod legitim, drept articol. Această opțiune s-a tradus într-o gamă de articole ce pornesc de la 546 caractere și spații (i.e. articolul 1FR) și ajung până la 4 919 (i.e. articolul 78FR).

Un ultim punct, dar la fel de important, este faptul că deoarece scopul nostru este examinarea calității traducerii automate a acestor articole și pentru că GTNM nu poate procesa paratext (i.e., sigle, scheme, diagrame, desene, poze) orice element de acest ordin este ignorat. Este adevărat cu toate acestea, că elementele paratextuale asigură caracterul complet al textului într-un articol de vulgarizare științifică și că el reprezintă un criteriu de evaluare a calității popularizării științifice. Dar faptul că nu putem analiza niciun element constitutiv al paratextului nu ni-l putem reproșa. Nu este o deficiență ce ne poate fi atribuită. Din păcate, cercetarea noastră este limitată de însăși mașina neuronală, care ne obligă, pentru moment, să ne mulțumim cu analiza corpului articolului.

În afara acestor două anexe principale, corpusul nostru a fost tradus și în engleză, fapt ce corespunde Anexei C – Corpus englez. Am făcut acest lucru deoarece una dintre întrebările noastre de cercetare vizează să descopere dacă engleza, care rămâne limba principală de antrenare a rețelelor de neuroni artificiali, joacă un rol în producerea de greșeli de traducere din franceză în română. Motivul principal pentru care engleza are un rol important pentru traducerea automată neuronală este existența unui volum mare de date digitalizate și de corpusuri (traduse) paralele din și în engleză.

Engleza se comportă ca o « limbă pivot » a TAN deținută de Google, dar fără a fi de fapt asta, în realitate ; cel puțin din perspectiva modului de funcționare al mecanismul cu *interlingua*, și anume, perechi de limbi individuale precum FR →/EN/→RO.

Cu toate acestea, în cadrul GTNM, un sistem care traduce din și către mai multe limbi, engleza poate fi văzută drept « limbă intermediară nec plus ultra ». Mașina folosește mai multe codificatoare și decodificatoare. Orice informație lingvistică ce aparține tuturor perechilor de limbi și pe care sistemul o găsește relevantă este utilizată de către mașină pentru a învăța și pentru a ajunge de la limba sursă la limba țintă. Echipa Google Brain ia în calcul o arhitectură neuronală *Zero-Shot*, aptă să facă o traducere directă între două limbi pe care sistemul nu le-a văzut niciodată.

Conform unui articol ce datează din septembrie 2021 publicat de Wang et al. (p. 4321), realizarea unei traduceri de înaltă calitate în modul « zero » este o provocare de calibru mare ce nu reprezintă deocamdată decât o funcție promițătoare a mașinii de traducere automată neuronală multilingve. Chiar dacă arhitectura care se bazează pe engleză ca legătură implicită sau *implicit bridging* (i.e. le *Zero-Shot translation*), nu este încă performantă, cea care folosește engleza drept legătură explicită sau *explicit bridging* este în funcțiune. Acest lucru ne obligă să îi examinăm implicarea în contextul traducerii neadecvate a unităților semantice de dimensiuni mici și mari.

4.3 Exploatarea corpusului

Cuvântul « exploatare » din sintagma care constituie titlul prezentului subcapitol ar putea să-l facă pe cititor să-și imagineze un capitol ce prezintă explicații cu privire la analiza lingvistică propriu-zisă. În realitate, el privește etapele de la 1 la 3 de mai jos, realizate în etapa anterioară analizei și care sunt elemente ale fișierului Excel « Pré-analyse_Corpus aligné » :

1. alinierea corpusului multilingv

2. anotarea pe corpus

3. structurarea nivelurilor semantice

Această muncă preanalitică a fost realizată cu ajutorul unor programe și s-a coagulat într-un fișier Excel, o metodă obligatorie, dată fiind natura obiectivelor noastre. Doar după alinierea, anotarea și strcturarea planurilor semantice într-un fișier Excel am putut pune în aplicare metode de analiză cantitativă.

Generarea de statistici este indispensabilă, întrucât ele dovedesc faptul că rezultatele noastre nu sunt simple supoziții. De asemenea, ele ne ajută să organizăm mai bine datele, să înțelegem mai bine rezultatele de cercetare, să descoperim modele și să facem previziuni.

4.3.1 Aliniere automată – metodă și instrumente

Primul stadiu de pregătire al corpusului în vederea analizei a fost utilizarea instrumentului LF Aligner[1] (Farkas, 2015), o aplicație ce poate alinia texte în mai multe limbi simultan cu precizie, fără să facă apel la un instrument de traducere asistată pe calculator. Efectiv, ea ne-a permis să generăm un fișier XLS plecând de la trei documente .docx (i.e. documentul cu discursurile sursă și alte două documente cu

[1] LF Aligner este *a priori* o aplicație care permite traducătorilor să genereze memorii de traducere pornind de la mai multe tipuri de documente și traducerile acestora. Ea reușește să genereze fișiere TXT, TMX sau XLS. Aplicația poate fi descărcată de la adresa următoare https://sourceforge.net/projects/aligner/.

traducerile automate ale acestora în engleză și franceză). Software-ul, în versiunea 4.1 bazat pe Hunalign[2] a efectuat automat corespondența propozițiilor în celulele foii de calcul tabelar.

După ce am verificat și corectat inadvertențele reieșite în urma alinierii, am obținut o bază de date Excel, lucru care ne-a permis ulterior să identificăm și să anotăm fiecare greșeală manual. Acest lucru a facilitat utilizarea programului AntpConc[3], instrument de analiză lingvistică pentru corpusul paralel. Făcând acest lucru, am reușit să tratăm corespunzător colecția noastră de documente. Printre numeroasele feluri în care acest program ne-a ajutat, putem menționa detectarea tuturor aparițiilor unui lexem în corpusul francez, român sau englez și identificarea co-textului său.

4.3.2 Anotare manuală – procedură și convenții

Formatul greșelii este de tip « erreur_nombre1.nombre2 » (ro. : eroare_număr1.număr2). Eroarea este evidențiată în roșu și poartă același număr de referință în cele trei limbi care intră în componența corpusului nostru. Mai mult, formatul « ERREUR _nombre1.nombre2 » (ro. : EROARE_număr1.număr2) oferă indicații adiționale în ceea ce privește amplasarea greșelii. El ne indică că greșeala face parte fie dintr-un titlu, fie dintr-un subtitlu scris în majuscule. În consecință, orice item lingvistic ce apare în majuscule în filele fișierului nostru Excel « Pré-analyse_Corpus aligné » se situează la nivelul titlului sau subtitlului de text.

Atunci când nici cea mai mică alterare a sensului care ar putea fi pusă pe seama limbjului intermediar (LI) nu a fost depistată, elementul lingvistic în cauză nu a fost nici numerotat și nici evidențiat în roșu. Cât despre absența interferenței LI în producerea greșelilor, ea este semnalată printr-o linie oblică, în timp ce netraducerea unui element lingvistic este indicată prin numărul « 0 ».

Din rațiuni de rapiditate și concizie, în partea de analiză vor fi folosite des abrevierile întrebuințate în fișierul Excel « Pré-analyse_Corpus Aligné ». Numărul greșelii poate fi uneori urmat de FR, RO sau EN în funcție de limba corpusului din care face parte (ex. : 48.11_FR ; 48.11_RO ; 48.11_EN). Totodată, litera T sau G poate apărea după greșeală, lucru care indică apartenența unității lingvistice respective la limbajul tehnic ori general (ex. : 48.11_G ; 48.11_EN_G). Mai apoi, notațiile LexS, LexC, Co și Loc intervin în completarea celor menționate în prealabil cu scopul de a preciza care este nivelul semantic la care se situează greșeala (ex : 1.2_RO_G_LexS ; 12.8_EN_T_Co ; 5.7_Loc ;6.2_G_LexC). Trebuie, în același timp, precizat că cele trei tipuri de notații care oferă informații complementare pot apărea simultan, individual sau sub diferite combinații.

---

[2] Varga D., Németh L., P. Halácsy P., A. Kornai A., V. Trón V., V. Nagy V. (2005). Parallel corpora for medium density languages. *Proceedings of the RANLP 2005*, pp. 590-596. https://kornai.com/Papers/ranlp05parallel.pdf ;

[3] Anthony, L. (2013). AntPConc (Version 1.0.2) [Computer Software]. Tokyo, Japan : Waseda University. Disponibil la http://www.antlab.sci.waseda.ac.jp/

4.3.3 Structurare – criterii specifice de analiză

Cercetarea noastră se articulează în jurul a patru niveluri semantice, fiecare repertoriat într-o filă Excel distinctă (i.e. LexS, Lex C, Co, Loc). Fiecare filă, cu excepția filei Excel Loc, prezintă un ansamblu uniform de 10 coloane, fiecare scoțând în evidență caracteristici specifice. Este indicat să specificăm că ultima coloană « Référent dans la réalité extra-linguistique » (ro : Referent în realitatea extra-lingvistică) nu este pertinent pentru categoria Loc astfel că locuțiunile nu sunt asociate unor referenți din lumea reală.

Acest lucru nu înseamnă că fiecare lexem din celelalte trei categorii afișează și referentului din lumea reală. Această corelare se aplică doar unui număr limitat de lexeme, care pot crea dificultate atunci când încercăm să le asociem unor obiecte fizice pe care le putem percepe în mediul înconjurător. Mai mult, referentul este reprezentat de o fotografie ce ilustrează obiectul în cauză, imagine preluată din articolul ce face parte din corpusul nostru și în care se vorbește despre el.

În ceea ce privește numărul de rânduri, acesta variază în funcție de numărul de greșeli identificate pentru fiecare grup în parte. Prima categorie de erori în care am inclus lexeme individuale este redată de fila Excel « LexS ». Pentru fiecare membru al acestei categorii, am optat pentru o descriere preanalitică ce conține specificități de tipul : număr de referință, apartenență la limbajul general sau tehnic, tipologie sintaxică, cauza greșelii, scurtă pre-analiză, formă grafică în limba sursă și limba țintă, formă în LI, dacă și numai dacă ea trădează o participare la traducerea greșită și la final, o reparație semantică care este construită doar pentru contextul propriu greșelii.

Această reparație, ce poartă numele de « remediere contextuală », este rezultatul a numeroase consultări de surse fiabile mai ales pentru unitățile de limbă ce provin din limbajul specializat. Atunci când am considerat relevant, aceste surse au fost incluse în « brève pré-analyse ». (ro : scurta preanaliză). În cazul în care am descoperit existența mai multor variante de traducere în uz, echivalentul român selecționat a fost acela al cărui frecvență de utilizare este net superioară. În momentul în care anumite variante prezintă grade de utilizare aproape egale în limba țintă, ele sunt toate incluse în fișa unității în cauză. Acest ultim caz este mai frecvent în rândul unitățile lexicale sau polilexicale care constituie elemente ale limbii tehnice. După cum se va constata în cazul greșelii 41.9, există două variante de echivalare principale în română, și anume *învățare prin consolidare* și *învățare prin întărire*. Incidența 41.9 reflectă o noțiune neologică și este motivul pentru care nu există încă un termen precis pentru a o numi. Sunt multe alte cazuri ca acesta, inclusiv cazuri în care conceptele sunt atât de noi încât limba română nu a avut timp să le denumească. În marea lor majoritate, denumirile sunt împrumutate din engleză în întregime, sau cel puțin, unul dintre constituenți este împrumutat, atunci când vine vorba de colocații din domeniul specializat. Putem da drept exemplu, colocația românească *eoliană offshore*, al cărei colocativ este împrumutat din engleză.

Dar, vom insista pe subiectul remedierilor și al reglementării termenilor străini în română în capitolul dedicat prelungirilor tezei. Acolo, vom arăta în ce mod cercetarea noastră reprezintă o sursă de date valoroasa pentru alcătuirea unui glosar terminologic dedicat denumirilor pe care le poartă conceptele

inovatoare în funcție de uzanța curentă. Mai mult, munca noastră poate sta la baza creării unui glosar cu propuneri lingvistice de substituire care să poată furniza alternative românești la neoterminologia englezească.

Mai apoi, greșelile care implică unități lingvistice compuse se reunesc în fila Excel « LexC ». Fișele lor descriptive conțin același tip de informație, cu excepția încadrării sintaxice. Este mai relevant ca aceste lexeme să fie însoțite mai curând de un criteriu de tipul « typologique morphologique » (ro. : tipolgie morfologică) pentru că acest lucru permite observarea descompunerii în morfeme. Astfel, accesăm o schemă brută a construcției de sens, a împletirii straturilor de semnificației. Punerea în evidență a acestei întrepătrunderi de sensuri facilitează înțelegerea prelucrării lor incorecte lor de către MTN.

Greșelile aparținând domeniului combinării de lexeme care se asociază natural în limbajul general și specializat sunt prevăzute cu aceleași filtre de caracterizare ca în cazul lexemelor simple. Am adăugat în cadrul acestei file, construcțiile cu verb cauzativ precum și pe cele cu verb suport (i.e. colligations). Acestea beneficiază de un semantism actualizat fie de un verb la infinitiv fie de un nume predicativ. Și din cauza faptului că unii din membrii de formare ai acestor structuri sunt lipsiți de sens în întregime sau parțial, aceste structuri fac obiectul unei subliste.

Două alte file Excel sont de asemenea, încorporate în fișierul dedicat preanalizei datelor. Ele sunt concepute special pentru legendele care explică sursele de producere a greșelilor de traducere. Una din file este dedicată legendei pentru lexeme simple și compuse, în timp ce cealaltă, deservește nivelului sintagmelor frazeologice. Fără aceste file Excel suplimentare, dobândirea unora dintre datele nostre statistice ar fi imposibilă.

5. Concluzii generale

Această teză interdisciplinară a fost concepută pentru a constitui înainte de toate, un model polivalent de studiu semantic de inspirație cognitivă aplicată unei tehnologii de TA cu IA pe un corpus de specialitate, multilingv. În acest scop, 100 de articole de VS au fost aliniate automat prin intermediul LF Aligner în timp ce rezultatele de traducere incorecte livrate de GTNM au fost anotate și inventariate manual în funcție de mai mulți factori (i.e. dimensiune sintaxică, morfologie, limbaj tehnic sau general, cauza greșelii, interferența LI, remediere contextuală etc.). Acest lucru ne-a permis convertirea acestora în date statistice, capabile să indice cu exactitate, coeficientul de calitate semantică ce le corespunde.

De fapt, teza noastră a depistat toate greșelile de traducere furnizate în 2018-2019 de către tehnologia debutantă și consolidată a Google în materie de *deep learning* (i.e. sistemul *Transformer* cu acces gratuit) la nivel de construcții semantice reduse și extinse. Aceste construcții au fost mai apoi explorate în profunzime datorită utilizării unor instrumente de analiză ce izvorăsc din semantica cognitivă. Potrivită particularităților TAN, semantica cognitivă s-a dovedit a fi alegerea adecvată în ceea ce privește strategia de analiză a datelor. Amintim faptul că mecanismul de atenție al lui Badhanau implementat în 2017 în mașina de TAN a Google pune accentul pe context și pe înțelegerea textului la nivel global. Mai mult, această reprezentare dinamică a sensului efectuată de rețelele de neuroni artificiali este un model comparabil cu abordarea dinamică propusă de SC în raport cu construcția semnificației.

Prin urmare, pentru a putea optimiza integrarea fenomenelor semantice complexe în arhitectura traducătoarelor automate neuronale, cercetările viitoare în legătură cu prelucrarea limbajului natural și IA se pot folosi de rezultatele noastre de cercetare. Pentru a obține aceste rezultate, am formulat întrebări de cercetare ce pot arăta măsura în care alegerile efectuate de TA neuronală se diferențiază de construcțiile mentale umane și derivă din limba de antrenare a rețelei. Ca urmare a acestui lucru, prezentarea rezultatelor generice ale cercetării are o structură ce gravitează în jurul întrebărilor de cercetare și este compusă din secțiunile I– IV care urmează :

I. INFLUENȚA POLISEMIEI ÎN PRODUCEREA DE GREȘELI

Un principiu care guvernează SC stabilește că reprezentarea sensului este enciclopedică, lucru care echivalează cu o respingere a abordării definiționale. Ținând cont de faptul că în cadrul SC, semantica și pragmatica sunt inseparabile, selectarea unui sens anume nu se poate face în afara contextului. Dat fiind faptul că rețeaua neuronală artificială a Google împărtășește această perspectivă cu privire la dezambiguizarea lexicală, provocările pe care ea trebuie să le depășească sunt alegerea unei variante din alternativele sens propriu/figurat/uzual/specializat/modulat, oricare dintre ele dezvăluindu-se numai în context.

La nivel LexS, acolo unde predomină lexemele nominale cu 181 de apariții, polisemia a fost resposabilă de 1/3 din rupturile de sens, fapt ce s-a soldat cu 86 de greșeli dintr-un total de 288. Această grupă include 47 lexeme ce provin din limbajul general și 39 de lexeme specializate. Polisemia s-a manifestat

în mai multe moduri, dar mai bine de jumătate din cazuri sunt în corelație cu LI. Pe deasupra, 6 greșeli din această categorie sunt împrumuturi din engleză în limba sursă. De fapt, polisemia din LI este prima cauza de apariție a greșelilor în cazul lexemelor franceze împrumutate din engleză. Din cele trei împrumuturi din engleză al căror transfer deformat este provocat de LI, record_56.2/93.3 este un exemplu de problemă de TA ce are legătură cu o polisemie convențională în timp ce împrumuturile integrale flash_38.6/38.10 și brownie_64.6 dobândesc un microsens contextual care nu a fost corect decodat de către mașină.

La nivel LexC, acolo unde găsim aproape numai lexeme nominale, am reperat 5 situații de traducere deficientă din 57 a căror sursă este polisemia. În acestă categorie intră doi termeni, și anume împrumutul din engleză X-plane_18.4/18.6 și prise de vue_37.6/94.3. Ultimul lexem compus este supus unei modulații în cadrul contextului sursă.

La nivel Co, acolo unde greșelile clasate sub tipologiile N Adj et N prép N prevalează, polisemia intervine doar rareori, și anume de 19 ori din 152 :

- 11 cazuri de polisemie care se manifestă fie la nivel de bază a colocației, fie la nivel de colocativ sau la nivelul întregului ansamblu frazeologic livrat în LI (a se vedea 11A+11B+11C) ;
- 8 cazuri de polisemie fie la nivel de bază a colocației, fie la nivel de colocativ din limba sursă (voir 10A+10B).

În ceea ce privește polisemia din LI, ea se observă în cazul a cinci G_Co_std, 4 T_Co_std, 1 T_Collig de tip « faire-vb support » și 1 G_Collig de tip « avoir-vb support ». Cât despre polisemia care nu implică LI, ea apare în cazul a patru colocații ce aparțin lexicului general, trei colocații terminologice și 1 T_Collig. Fiecare tipologie sintaxică din sfera Co_std ce se află pe podiumul frecvenței este reprezentată. Li se alătură totodată și o construcție de tip « faire causatif ».

La nivel Loc, acolo unde majoritatea erorilor sunt locuțiuni verbale, polisemia constituie un factor determinant doar pentru o greșeală. Aceasta se observă în cazul unității polilexicale fixe « mettre en service_8.1 », în cazul căreia polisemia lexemului rezultat în LI generează eroarea de ieșire.

II. INFLUENȚA CONTEXTULUI ÎN PRODUCEREA DE GREȘELI

Prima secțiune care oferă o priveliște de ansamblu asupra influenței polisemiei în producerea de sensuri compromise ne-a făcut să înțelegem că în definitiv, în cazul tuturor greșelilor, avem de-a face cu o tratare necorespunzătoare a contextului de către arhitectura neuronală a Google. S-a observat că semnificația construcțiilor care fac parte din corpusul nostru în limbă sursă este reprodusă fidel de către creierul artificial Google în limbă țintă doar dacă acesta stăpânește aspectele de mai jos :

- sensurile instituționalizate ale unei construcții lexicale de dimensiune mică sau mare din limba sursă (i.e. sens conotativ, denotativ, specializat) ;
- sensurile noi ce sunt adăugate construcțiilor preexistente din limba sursă ;
- neotermenii și neofrazemele din limba sursă ;

- semnificația siglelor și abrevierilor ;
- forma și particularitățile unui titlu/subtitlu ;
- particularitățile numelor proprii.

Ori, faptul că mașina nu poate încă să trateze corespunzător toate aspectele de mai sus a generat :

- greșeli de traducere la nivel de lexeme și sintagme frazeologice provenite din LG și LT, altele față de cele care sunt legate de LI sau polisemie ;
- o alterare a sensului titlurilor/subtitlurilor care conțin jocuri de cuvinte sau subînțelesuri fondate pe tropi conceptuali ;
- transferuri incorecte la nivel de nume proprii, majuscule, sigle și abrevieri ;
- segmente vide de traducere ;
- greșeli recurente de traducere ;
- fabricare de referenți fictivi ;
- iinterpretare distorsionată a construcțiilor cauzative și cu verb suport.

Traducerile eronate care nu sunt în relație cu LI sau cu polisemia sunt grupate sub 6A pentru nivelul LexS și LexC și 14A/B/C pentru nivelul sintagmelor frazeologice. Cauza 6A totalizează 36 de greșeli la nivel LexS din care 25 sunt lexeme generale. Mai apoi, nivelul Co_std însumează 22 de greșeli de acest tip care se fac simțite îndeosebi în cazul colocațiilor specializate în timp ce nivelul Collig. înregistrează 5. La nivel LexC, cauza 6A înglobează 11 greșeli, dintre care 9 sunt termeni compuși. În materie de construcții idiomatice, printre acestea au fost identificate doar trei cazuri de traducere lipsite de orice legătură cu LI sau cu polisemia, și anume două locuțiuni nominale (i.e. rappel des faits_96.3 și coup d'éclat_49.5) și o locuțiune verbală (i.e. monter le son_83.1). Transferul lor incorect în limba sursă are în schimb de-a face cu tropii conceptuali pe care îi găzduiesc.

Deformarea sensului titlurilor și subtitlurilor este cauzată în principal, de contextul limitat ce le caracterizează. Traducătoarele neuronale tratează uneori titlurile și subtitlurile în mod izolat față de contextul global, lucru care conduce bineînțeles, la greșeli în interpretarea implicitului. La nivelul titlurilor și subtitlurilor, vulgarizatorul inserează uneori construcții idiomatice ale căror nuanțe nu sunt surprinse de mecanismul de TA neuronală (i.e. prendre le large_5.1, viser le large_7.1). Din cauza dimensiunii reduse a titlurilor și subtilurilor, survin traducerile deformate. Lăsând la o parte construcțiile verbale fixe (CVF) precedente, iată celelalte construcții care participă la semnificația titlurilor și care au fost înțelese greșit : LexS (i.e. chargé_10.2, aller_ 10.3, supersonique_18.1, arriver_24.1, duvet_33.1, bourdon_43.1, feu_44.1, embouteillage_54.1, drone_55.1, mollusque_60.1, durer_60.2, crécelle_63.1, fractale_75.1_végétal_79.1, tenir_82.1, skate_85.1, connecté_81.3, s'enrouler_86.1, se lacer_88.1, métro_92.1, colorant_99.1), LexC (valise accordéon_10.1, monoroue_72.2, éthylotest_74.1, autonettoyante_81.2), Co (i.e. lave-linge à pédale_11.1, partie mobile_36.1, essaim de minirobots_47.1, journal télévisé_68.1, véhicule tout-terrain_70.1, trottinette électrique_72.1, litière pour chat_81.1, salto arrière_90.1) și Collig (faire bang_18.2, faire des gâteaux_64.1). O observație care se impune în egală măsură, este faptul că există cazuri în care itemii lingvistici sunt

tratați corect de către mașină în corpul articolului, dar în mod defectuous în titlu ori subtitlu (ex. métro). Un alt aspect de subliniat ar fi acela că unele dintre aceste unități lingvistice generează de mai multe ori aceeași greșeală sau greșeli diferite de-a lungul textului (ex. bourdon, mollusque, duvet, crécelle, monoroue, salto arrière).

Greșelile de traducere recurente ce aparțin primului nivel de analiză se regăsesc în sub-capitolele 5.1.2.3_LexS și 5.1.3.3_LexC. Dintre cele 23 de greșeli care se repetă și care au generat fiecare între 2 și 10 greșeli la nivel LexS, 12 sunt lexeme specializate și 9 fac parte din limbajul uzual. Acestea au avut cauze multiple de producere, dar interferența LI și-a pus amprenta asupra majorității (i.e. 5C et 4A). În ceea ce privește nivelul LexC, acesta reunește 13 lexeme, mai exact 7 care livrează același rezultat de ieșire (X-plane, memristor, tout-en-un, prise de vue, feu rouge, mille-pattes, éthylotest) și 6 a căror traducere finală variază. Deși greșelile recurente care apar la nivelul structurilor colocative nu fac obiectul unui subcapitol distinct în cadrul tezei, ele vor fi discutate aici. Acestea se subdivizează în două grupe, una care include asocierile lexicale care produc același rezultat în limba țintă și una care conține construcții care dau naștere la două rezultate diferite. Prima grupă cuprinde 6 colocații specializate, dintre care 3 includ o siglă, precum și o colocație de limbaj general care generează trei traduceri identice (a se vedea première mondiale_42.1/68.2/94.4). Cealaltă grupă afișează 5 colocații terminologice și o colocație din LG care sunt, în marea lor majoritate, transferate în limba sursă în mod viciat din cauza interferenței LI ; 4 cazuri de traducere literală din LI (a se vedea apprentissage par renforcement_41.8, station émettrice_43.13 & 43,11, trajet travail-domicile_72.6), 3 cazuri de reproducere din LI (a se vedea affichage tête haute_80.2, salto arrière_90.1/90.3 & 90,6) și două cazuri de polisemie din LI la nivel de colocativ (a se vedea apprentissage par renforcement_41.9, essaim de minirobots_47.1). O remarcă finală care se impune, este faptul că, dintre toate aceste erori repetitive, unele dintre ele constituie neologisme, împrumuturi din engleză, elemente de semnificație din titluri și subtitluri și generatoare de lexeme independente.

Segmentele vide de traducere sunt repertoriate sub cauza 7 și 17 și sunt observabile doar la nivelurile LexS și Co. Absența traducerii la nivel de lexeme simple este analizată mai în profunzime în cadrul subcapitolului 5.1.2.4. După cum am arătat, aceasta poate apărea doar în limba sursă sau atât în LI cât și în limba sursă și ea poate decurge din faptul că :

- lexemele ce trebuie traduse fac parte din domeniul specializat (a se vedea matériau _28.11, laser_38.8, houle_96.14) ;
- sensul lexemelor nu este cunoscut de către mașină (a se vedea intémperie_26.3, adapté_48.12, lacé_88.3) ;
- lexemele dețin un sens metaforic (a se vedea ligne_86.7).

S-ar părea și că mecanismul de limitare a pierderilor de traducere dictează uneori reproduceri ale segmentelor vecine la nivel LexS astfel încât coerența textului de ieșire să nu fie ruinată (a se vedea intémperie_26.3, lacé_88.3).

De cealaltă parte, coligațiile oferă doar 4 situații în care poate fi observată o non-traducere. Trei dintre acestea patru sunt colocații terminologice al căror colocativ nu este deloc transpus în limba țintă (a se vedea éolienne offshore_7.2, avion à réaction_36.13, courant continuu_93.2).

Traducerile prin lexeme inventate se reunesc sub cauza 6B pentru nivelul LexS et LexC și sub cauza 18A/B pentru nivelul Co & Loc. Unitățile lexicale simple și polilexicale născocite de GTNM în LT și LG sunt listate mai jos :

- G_LexS : OBSTACULURI_15.1, trântitor_18.5, ștafari_50.1, tobacconiști_50.5, tacamici_50.8, schiț_51.2, RETIREA_62.5, indisociat_72.19, Andele_76.4, Oița_91.1 ;
- T_LexS : etanșitate_5.5, PROZETA_16.3, moluscul_60.1/60.3/60.10, cubă_73.2, reeditǎ_97.1 ;
- T_LexC : monoplaz_31.3, CUPURI DE BARBE_50.4, difenilamino-clorarsină_69.6 ;
- G_Co : haina *zarbată*_47.6
- T_Co : orezuri în terase_6.5, tentativă de intruzare_45.3, timp de decenare_49.10.

Atunci când modelele neuronale se servesc de un corpus de antrenare limitat/bruiat și nu dispune de suficient context, acestea pot inventa soluții plauzibile care respectă regulile morfologice și fonetice ale limbi țintă sau creează aproximări. Alteori, o segmentare incorectă în *subwords* poate duce la o reasamblare eronată care produce, la rândul său, o construcție inexistentă (i.e. zarbată_47.6). Mai mult, uun sistem de TAN multilingv poate amesteca structuri din două sau mai multe limbi, lucru care dă naștere la cuvinte hibride (i.e tobacconiști_50.5). Uneori, în cazul traducerilor din franceză în română, absența diacriticelor din fragmentul de tradus poate obliga mașina să ghicească lexemul în cauză, generând astfel, un lexem arbitrar.

Transferurile incorecte la nivel de nume proprii iau naștere ca urmare a unui algoritm ce nu satisface nivelul așteptărilor în materie de performanță. În cazul a două nume proprii la nivel LexS, modelul neuronal creează aproximări care țin cont de bazele morfologice și fonetice ale limbii române.(i.e. Andele_76.4, Oița_91.1). După aceea, cazurile în care un nume propriu este în egală măsură nume comun creează ambiguitate, acesta fiind motivul pentru care creierul artificiel decide deseori să le traducă în loc să le păastreze sub aceeași formă (i.e LexS_Zefir_26.1). Regăsim acest fenomen și în cazul numelor proprii compuse X-plane_18.4/18.6, Bulk Jupiter_96.6, Cake Factory_64.3, Emerald Star_96.5, Hunday Kite_12.3 care au primit următoarele traduceri literare : plan X, Jupiterul Bulk, fabricare a prăjiturilor, steaua de smarald, Hunday Zmeul.

Interpretarea distorsionată care vizează construcțiile în majuscule face obiectul unei analize în conexiune cu titlurile și subtitlurile dat fiind faptul că acestea au fost detectate exclusiv în aceste părți de text. Drept urmare, ele decurg din cotextul condensat inerent titlurilor și subtitlurilor redactate cu litere mari de la nivelul cărora uneori lipsesc și accentele diacritice. Majoritatea unităților scrise cu litere majuscule sunt LexS, și anume șase G_LexS și cinci T_LexS. Cât despre devierile semantice pe care acestea le declanșează, pot fi exemplificate următoarele :

- lexeme simple inventate : OBSTACUL, DRAFĂ, PROZETA, ALTE, RETIREA pentru OBSTACLE_15.1, DRONE_15.2, PROTHÈSE_16.3, ETHER_50.2, RÉTICENCE_62.5 ;
- lexeme compuse fictive : CUPURI DE BARBE pentru CODE-BARRES_50.4 ;
- reproduceri din LI : BIRD pentru OISEAU_15.3, ROUTE pentru TRAJET_17.5 ;
- alegerea sensului specializat inadecvat din cauza unei polisemii din LI : nervură pentru NERF_16.1 ;
- traducerea literală din LI : emisie zero pentru ZÉRO ÉMISSION_17.3, CERERE pentru JUSTIFICATIF_50.6.

Trebuie să notăm faptul că mașina încearcă să convertească literele majuscule în minuscule în cazul termenului 16.1 și a binomului specializat 17.3 în vederea ameliorării calității traducerii.

Inexactitățile din traducerea siglelor și a trunchierilor apar la nivel T_Co. Modelul neuronal este tributar greșelilor de decodare a conceptelor vehiculate de sigle și forme trunchiate. De fapt, neofrazemele siglice simple (i.e. EDP, MSRR, RNA, TER) sau hibride (i.e. brins d'ARN, e-EDP, dalle OLED) întâlnite în corpusul nostru, sunt situații metonimice de tip FORMĂ A - CONCEPT A PENTRU FORMĂ B - CONCEPT A, explicate în detaliu în cadrul analizei noastre.

Construcțiile cauzative și cu verb suport sunt interpretate deficient pentru că ele reprezintă structuri mai complexe ce prezintă dificultăți și particularități de ordin atât sintaxic cât și semantic. Mai mult, verbele « faire » și « avoir » care intră în compoziția a 15 coligații din 20 ridică probleme de contextualizare, întrucât acestea pot beneficia de mai multe traduceri. Cu toate acestea « faire causatif » este construcția care generează cele mai multe greșeli, și anume 5 T_Collig și 3 G_Collig, pentru că limba română nu utilizează verbul « faire » urmat de infinitiv pentru a exprima cauzalitatea. În ceea ce privește cauzele traducerii eronate ale acestei construcții cauzative, am observat că jumătatea cazurilor sunt determinate de o traducere literală din LI în timp ce cealaltă jumătate nu este contaminată de LI

III. IINFLUENȚA TROPILOR CONCEPTUALI ÎN PRODUCEREA DE GREȘELI

Datorită analizei noastre, am putut remarca că idiomaticitatea este prezentă în mod transversal la orice nivel lingvistic. Ea are la bază tropi conceptuali, fie metaforici, fie metonimici care structurează gândirea și contribuie la naturalețea limbajului. Locutorii nativi utilizează tropii conceptuali intuitiv după ce i-au asimilat prin prisma corporalității și experienței lor din lumea reală în timp ce pentru MTN, înțelegerea și utilizarea lor în mod eficient implică un proces de învățare într-o lume virtuală, unde nu există corp conectat la percepții senzoriale. Mai mult, sensul metaforic este înrădăcinat într-un cadru cultural specific și este dificil pentru creierul artificial să integreze toate aceste scheme cognitive din cauza limitărilor sale tehnologice și a unei existențe într-o lume fragmentată și mediată.

La nivel LexS și LexC, trebuie verificat inițial dacă contextul lexemelor în cauză lasă să se întrevadă o relație de asociere între un domeniu concret și unul abstract sau de contiguitate în același domeniu pentru a putea detecta astfel sensurile figurate. În mod evident, corpusul nostru conține mai multe

lexeme simple cu sens conotativ în limba sursă. Totuși, greșelile în materie de figurativitate iau forme multiple, cele mai frecvente fiind: :

- lexeme simple folosite metaforic în limba sursă, dar care au fost traduse printr-un sens propriu în limba țintă, fără interferență LI, de exemplu, engin_9.2 & 17,1 & 48,5 & 36,12/36,15 (*mașină* în loc de *dispozitiv & vehicul & robot/aeronavă*), bridage_58.4 (*prindere* în loc de *limită*), tomber_96.1 (*a cădea* în loc de *a publica*) ;
- lexeme simple redate în limba țintă printr-un sens figurat în loc de sens propriu fără interferență LI, de exemplu, carrefour_44.3/44.5/44.10 (*răscruce* în loc de *intersecție*) ; embouteillage_54.1 (*blocaj* în loc de *ambuteiaj*), acheminer_5.10 (*a transporta* în loc de *a orienta*) ;
- lexeme simple folosite metaforic, dar care au fost reproduse în limba țintă fără interferență LI, de exemplu, las_95.1 (las în loc de *bătrân/învechit*) ;
- lexeme simple folosite metaforic în limba sursă, dar a căror traducere în limba țintă a fost alterată de LI, de exemplu, engin_36.6/36.17/36.18/72.26 (*ambarcațiune* în loc de *aeronavă*), logé_9.4/24.3 (*adăpostit* în loc de *amplasat*), surenchère_3.4 (*escaladare* în loc de *concurs/competiție*), alléger_91.3 (*a lumina* în loc de *a reduce sarcina/greutatea*), arriver_18.8/24.1 (*a ajunge* în loc de *a apărea*) ;
- lexeme simple folosite metaforic în limba sursă care au fost traduse în mod eronat prin construcții idiomatice în LI, lucru ce a influențat greșeala finală din limba țintă, de exemplu rogner_78.6 (a se vedea Blend-ul conceptual_COUPER LES COINS EST OBTENIR UN RÉSULTAT DE MOINDRE QUALITÉ) ;
- lexeme simple folosite metaforic în limba sursă care nu au primit traducere nici în LI și nici în limba țintă, de exemplu, ligne_86.7 (a se vedea Metafora conceptuală_ LA LIGNE ÉCRITE EST UNE TRAJECTOIRE) ;

La nivel Co, metaforizarea este de asemenea prezentă, dar mai rar ca în cazul lexemelor simple sau a expresiilor idiomatice (EI). În cazul colocațiilor generale din corpus în care apare metaforizarea, colocativul este de obicei cel care permite stabilirea de corespondențe între două domenii diferite de semnificație. Greșelile în raport cu metaforizarea reprezintă aproximatic 10% din numărul total al colocațiilor și afectează colocațiile generale precum și pe cele specializate în raporturi aproape egale. Putem da drept exemplu următoarele colocații generale :

- utilisation fluide _17.7 = Metafora conceptuală_LA SIMPLICITÉ EST LA FLUIDITIÉ ;
- calories dépensées_25.2 = Metafora conceptuală_ L'ARGENT EST DE L'ÉNERGIE CORPORELLE ;
- gros pic_45.2 = Metafora conceptuală_L'IMPORTANCE EST LA HAUTEUR ;
- gâteau au cœur coulant_64.7 = Metafora conceptuală_ LE CŒUR EST UN CENTRE ;
- franchir un obstacle_40.2 = Metafora conceptuală_ UN OBSTACLE/DÉFI EST UNE BARRIÈRE ;

În cazul coocurențelor tehnice, constatăm adesea că termenii-pivot care adăpostesc conceptul central joacă un rol cheie în construirea referinței metaforice (ex. dalle OLED, réservoir computing). Pe de altă parte, efectul de metaforizare dat de colocativ poate fi sesizat în cazul următoarelor colocații :

- flotte captive_ 1.6 ;
- essaim de minirobots_47.1 (a se vedea Amalgamul L'ESSAIM DE MINIROBOTS EST UNE STRUCTURE QUI RÉPOND COLLECTIVEMENT).

La nivel Loc, printre tropii conceptuali care stau la baza locuțiunilor a căror semnificație a fost deteriorată, putem enumera :

a) G_Loc verbale : metaforele conceptuale LE PROGRÈS EST UNE MARQUE PHYSIQUE LAISSÉE (creuser un sillon_5.11), LA CLARTÉ DE L'INFORMATION EST LA LUMIÈRE (laisser dans l'ombre_49.2), UN POINT CULMINANT EST UN COURONNEMENT (pour couronner le tout_95.13), UNE IDÉE EST UN OBJECT FIXÉ (enfoncer le clou_ 5.6), LE PROGRÈS EST LE DÉPASSEMENT D'UNE LIGNE DE DÉMARCATION PHYSIQUE (franchir un cap_23.1 & 76.1), LA VICTOIRE EST L'ARRIVÉE À UNE LIGNE DE DÉMARCATION PHYSIQUE ET LA TÊTE EST LA DOMINANCE (coiffer au poteau _49.12), LE DÉNONCIATEUR EST UN DOIGT (pointer du doigt_78.3) ;
b) G_Loc nominale : metafora conceptuală UN ÉVÉNEMENT NOTABLE EST UN COUP D'ÉCLAT (coup d'éclat_49.5) și metafora conceptuală FAIRE DU VÉLO EST UN SPORT NOBLE care pleacă de la baza metonimică LE VÉLO EST LA REINE (la petite reine_49.5) ;
c) T_Loc verbale : L'OBJECTIF EST UN ENDROIT (viser le large_7.1), L'IMPORTANCE EST LA HAUTER (tenir le haut de l'affiche 28.2) ;
d) T_Loc nominale : UN ÉVÉNEMENT SOUDAIN ET DÉSÉQUILIBRANT EST UN COUP (coup de roulis_96.12) ;
e) T_Loc Adjectivale : LA CONSOMMATION RAPIDE DE RESSOURCES ÉNERGÉTIQUES EST UNE DÉVORATION DE NOURRITURE (être gourmand en energie_43.4 & 43,8/43,10).

Mai mult decât atât, evaluând greșelile la nivel de construcții idiomatice, am observat în ce fel contextul le adaugă uneori microsensuri ce au la origine lanțuri metaforice sau blend-uri conceptuale (découvrir le pot-aux-roses_54.5, faire figure d'Arlésienne_1.4, poids plume_26.2).

În concluzie, tropii conceptuali se regăsesc la toate nivelurile semantice cu toate că discursul specific diseminării științifice este esențialmente interesat de informația tehnică expusă și mai puțin de maniera literară în care aceasta este transmisă. Constatarea generală care se impune este că atunci când se confruntă cu proiecții metaforice, modelul de TA neuronală al Google livrează sistematic greșeli de traducere la nivel LexS, LexC et Co și aproape întotdeauna greșeli la nivel de EI, care rămân marginale din punct de vedere al frecvenței.

IV. INFLUENȚA LI ÎN PRODUCEREA DE GREȘELI

Conform analizelor noastre, totul indică că rezultate improprii de TA imputabile TAN decurg într-o proporție covârșitoare de limba de antrenare a rețelei. Procentajul efectiv atins de anomaliile de traducere care sunt provocate, în totalitate sau parțial, de către LI la fiecare nivel semantic este specificat mai jos :

- ✓ LexS : ~75 % (i.e. 214 cazuri din 288)
- ✓ LexC : ~ 64 % (i.e. 36 cazuri din 57)
- ✓ Co : ~ 60 % (i.e. 91 cazuri din 152)
- ✓ Loc : ~ 68 % (i.e. 17 cazuri din 25)

La nivel LexS, modurile în care LI a intervenit în deformarea sensului de plecare sunt reluate în lista de mai jos care pornește de la cea mai mare și merge către cea mai mică valoare :

1. Polisemia din LI : 103 erori (4A – 57 er., 2A - 28 er., 2B – 12 er., 1A – 6 er.) ;
2. Trad. literală din LI : 71 erori (5C – 34 er., 2C – 18 er., 4C – 13 er., 1C – 4 er., 1E – 2er.) ;
3. Reproducere din LI : 19 erori (4B – 9 er., 5B – 6 er., 1B – 3 er., 1D – 1er.) ;
4. Interpretare din LI : 4 erori (5A) ;
5. Trad. eronată din LI : 2 erori (4D).

La nivel LexC, erorile în legătură cu LI sunt atribuite cauzelor următoare, clasate în ordine descrescătoare a frecvenței :

1. Trad. literală din LI : 25 er. (1C - 10 er., 5C – 8 er., 4C – 6 er., 1E – 1er.) ;
2. Reproducere din LI : 5 er. (1B/4B – 2 er., 5B – 1 er.) ;
3. Polisemie din LI : 4 er. (1A/2A/2B/4A – 1 er.) ;
4. Trad. eronată din LI : 2 er. (4D).

La nivel Co, traducerea literală din LI este cauza principală a greșelilor, urmată de polisemia din LI și de reproducerea din LI :

1. Trad. literală din LI : 72 er. (16A/B/C) ;
2. Polisemie din LI : 11 er. (11A/B/C) ;
3. Reproducere din LI : 8 er. (13A/B/C).

La nivel Loc, tot traducerea literală se face vinovată de majoritatea greșelilor :

1. Trad. literală din LI : 16 er. (16A) ;
2. Polisemie din LI : 1 er. (11A).

În concluzie, toate limitările TAN Google pe care le-am detaliat de la I la IV decurg dintr-o insuficiență de date, context restrâns sau izolat, probleme de decupaj în *tokens* (subunități), suprainterpretare a relațiilor dintre lexeme, ambiguitate semantică, necunoaștere a domeniului, valoarea metaforică a construcțiilor, variații culturale și idiomatice ale *decoding idioms*.

Prin exploatarea limitelor TAN Google din 2019, sperăm că am reușit să propunem piste de ameliorare în ceea ce privește mecanismele de gestionare a polisemiei în context, neonimiei, majusculelor, siglelor și a valorii metaforice a oricărei construcții. Ținând cont de rezultatele noastre de cercetare, nu putem decât să spunem că ipoteza temută de traducătorii umani în jurul căreia ne-am construit teza, și anume că IA îi va înlătura, era doar o viziune de roman, demnă de greii ficțiunii în 2019. Analiza noastră a punctat faptul că TA neuronală aplicată pe texte semi-specializate nu este ireproșabilă și că are un drum lung de parcurs pentru ca într-o zi să poată devansa cunoașterea umană de pe acest teren. Acest rezultat al studiului nostru le cere totodată actorilor de pe scena traducerii să consolideze sau să optimizeze hibridul curent traducător – mașină în loc să privească viitoarele schimbări tehnologice cu teamă. Mai bine spus, orice « traducere asistată pe calculator » ar trebuie să se transforme în « traducere asistată de IA » sau cel puțin, să faciliteze participarea TAN în procesul de traducere. De fapt, putem afirma chiar că această graniță a fost depășită deoarece în ultimii 3-4 ani, software-urile de traducere asistată pe calculator precum TRADOS, unii din liderii de pe piața de programe de traducere incearcă să ridice acest nou sistem la rang de normă curentă. În cadrul acestei noi formule, traducătorul uman îndeplinește rolul de depanator al iregularităților lingvistice, ce emană de această dată din automatizarea inteligentă.

Observația finală care se profilează pentru stadiul tehnologiei TAN Google din 2019 este că nu ne putem debarasa de traducătorul uman experimentat. El este singurul ce poate reface legăturile de sens și de stil rupte de TAN iar textele noastre de VS abundă în exemple în această privință. Cu toate acestea, TAN poate furniza rezultate bune atunci când textul nu are o valoare extrem de tehnică sau figurată. Acest lucru a fost observat în cazul unor articole și paragrafe din corpusul nostru. Astfel că, nu ne rămâne decât să urmărim îndeaproape inovațiile tehnologice în acest domeniu al TA neuronale pentru a determina dacă acest gen de probleme vor fi eliminate definitiv.

6. Prelungiri posibile

Acest capitol este dedicat posibilităților de prelungire a acestui studiu, scoțând în evidență provocările persistente de pe cele două axe de cercetare principale care se află în centrul tezei noastre : TAN și SC.

După cum am observat, studiul nostru a exploatat pe deplin rezultatul calitativ al traducerii automatizate neuronale a Google pe plan semantic atunci când aceasta beneficia de mecanismul introdus în 2017. Cu toate acestea, de la acel moment încoace, Google continuă să facă cercetări, să amelioreze și să înregistreze progrese atunci când vine vorba de rețelele sale neuronale, modelele sale de antrenare, gestionarea limbilor ș.a.m.d. Pentru verificarea stadiului cercetării, pot fi consultate blogurile sale și publicațiile oficiale, așa cum am făcut-o și noi pentru teza noastră în ultimii ani.

Prin urmare, am putea realiza o analiză comparativă între modelul de TA neuronal al Google din 2018/2019 și cel mai recent model al său pornind de la același corpus și folosind aceleași instrumente de analiză semantică. În mod evident, considerăm că pentru a putea atribui un « procentaj de conformitate semantică » unei traduceri transculturale, toate planurile pe care le-am supus analizei ar trebui reexaminate. Totuși, se poate lua în calcul doar tratarea polisemiei în context de către TAN sau doar tropii conceptuali, întrucât aceste teorii cognitive sunt la rândul lor, supuse constant schimbărilor, modificărilor și evoluției. În ambele domenii de cercetare care se împletesc în teza noastră, chiar dacă se fac descoperiri în mod regulat, sunt încă multe de perfecționat. Cercetătorii TAN fac eforturi să depășească limitele traducerii în ceea ce privește sensul gramatical, vocabularul terminologic existent sau emergent sau dependența de limba de antrenare, în timp ce lingviștii de orientare cognitivistă încearcă să-și facă teoriile cât de fiabile posibil pe plan științific și să se pună de acord cu privire la metodele de analiză cele mai valide. Atât în cadrul TAN cât și SC, se implementează metrici din ce în ce mai complexe pentru a construi o arhitectură/cadru de analiză optim(ă).

O altă idee de prelungire ar putea fi realizarea unei analize comparative între GTNM și modelul de traducere al DeepL pe același corpus sau alt corpus de specialitate. Pe pagina lor online[4], DeepL prezintă cu mândrie recenzia TechCrunch, USA care pretinde că această companie, deși mică comparativ cu gigantul Google, l-a depășit pe acesta din urmă și a ridicat ștacheta în materie de traducere. Această comparație este cu atât mai validă științific cu cât perioada de lansare a DeepL corespunde cu cea a GTNM. Compania germană laudă meritele tehnologiei sale în materie de rețele neuronale și garantează că va continua să inoveze pentru a face comunicarea mai rapidă și eficientă. Mai mult, acesta ar fi un studiu ce se pliază perfect pe un corpus cu limbaj tehnic pentru că DeepL Translate declară că a introdus în 2020-2021 noi modele capabile să restituie mai fidel sensul frazelor traduse și chiar să trateze cu succes jargonul specific anumitor sectoare de activitate.

[4] https://www.deepl.com/de/whydeepl

O altă prelungire ar putea fi analiza nivelului de fidelitate semantică asigurată de către mașina neuronală Google doar în cazul titlurilor și subtitlurilor de articole. Am remarcat că în anumite cazuri, atunci când titlul sau subtitlul este scris în majuscule, traducerea este mai eronată decât în situația în care acestea sunt scrise cu litere mici. După acestă constatare, am introdus în traducătorul automat titlurile și subtitlurile scrise cu majuscule sub o formă transformată în litere mici pentru a putea vedea dacă apar schimbări pe planul calității. Trecerea de la un text scris cu majuscule la unul cu minuscule a ameliorat exactitatea transferului semantic realizat de GTNM. Prin deducție, un factor cu o enormă pondere în controlul conformității traducerii sunt « accentele limbii franceze ». Acestea ajută la discriminarea lexicală și la îndepărtarea ambiguității lexicale. Dacă luăm ca exemplu cuvintele franceze « ou » și » où », putem avea certitudinea că acestea vor fi tratate corespunzător de către mașina de TAN doar dacă se regăsesc în titlu/subtitlu în forma lor grafică corectă.

În ceea ce privește corpusul nostru, acesta cuprinde multe exemple de itemi lexicali ce întregesc componența unui titlu sau subtitlu și care nu au fost bine traduși din cauza faptului că au fost redactați cu litere mari. Efectuând extragerea tuturor greșelilor de acest fel, am remarcat că unele dintre ele au fost :

- traduse prin copierea rezultatului din LI (i.e. 15,3, 17,5) ;

-traduse prin lexeme inexistente în limba țintă (i.e. 15,1, 15,2, 16,3, 50,2, 62,5, 50,4).

Mai mult, dintr-un total de 13 greșeli de acest fel, 11 se găsesc la nivelul lexemelor simple (6 x G_LexS și 5 x T_LexS). Celelalte două greșeli fac parte din limbajul tehnic. Derivă de aici faptul că există o cvasi-egalitate de greșeli când ne raportăm la LT vs LG.

O altă constatare adiacentă a fost aceea că titlurile și subtitlurile articolelor demască o altă slăbiciune a TA neuronale, și anume traducerea cuvintelor ce au un context limitat și o referință ambiguă. Cuvintele din titluri și subtitluri introduc uneori nuanțe metaforice pentru a atrage atenția cititorului și nu au o referință foarte exactă. Referința inițială este de cele mai multe ori clarificată în corpul articolului, care îl ajută pe cititor în procesul de dezambiguizare. Frecvența greșelilor la nivel de titlu și subtitlu este destul de crescută pe fondul acestor factori. De fapt, mai bine de o treime din totalitatea titlurilor din corpusul nostru conțin cel puțin o greșeală.

Prin urmare, o altă dezvoltare ulterioară a cercetării ar putea fi analizarea nivelului de fidelitate semantică asigurat de mașina neuronală a Google în cazul articolelor integrale comparativ cu versiunea lor fragmentată sau împărțită pe paragrafe. Aceasta posibilă extindere s-a prefigurat în urma constatării anterioare. Ar fi interesant de pus în evidență schimbările de sens ivite în urma unui text împărțit în secțiuni în contrast cu un text integral, sau cel puțin, de descoperit în ce măsură traducerea unui text divizat este mai deficitară. Această idee poate servi drept complement de cercetare sau temă de cercetare autonomă.

Un studiu care își propune să testeze traducerea abrevierilor (i.e. sigle, acronime, etc.) și numelor proprii ce figurează în articole ar fi important, dat fiind faptul că acestea pun dificultăți MTN. Fie că este vorba

despre nume de locuri, persoane sau denumiri de aparate, dispozitive, tehnologii precum și altele, abrevierile și numele proprii ridică destul de multe probleme pentru un traducător neuronal. Teza noastră a pus în valoare faptul că sensul acestor unități de limbă trebuie păstrat, iar asta implică de multe ori, o adaptare subtilă sau păstrarea identică a formei din limba sursă în vederea reducerii interpretărilor culturale greșite sau a înlesnirii identificării corecte. Analiza noastră a dezvăluit că toate numele proprii din FR care au avut parte de o traducere incorectă au aceeași formă în engleză. O cercetare care să se concentreze pe capacitatea actuală din 2025 a traducătorului neuronal Google în comparație cu capacitatea sa inițială din 2018/2019 ar indica, fără echivoc, nivelul de ameliorare al calității traducerii efectuată de mașină. Mai mult, ea ar putea contribui la îmbunătățirea mecanismelor neuronale de detecție a numelor proprii pentru a putea astfel depăși stadiul actual al acestora care nu este unul satisfător. După cum reiese din teza noastră, traducătorul neuronal al Google din 2018/2019 are tendința să reducă numele proprii la nume comune și să livreze ca urmare, traduceri literale sau adaptări incorecte.

În plus, am putea examina nivelul de fidelitate semantică asigurat de mașina neuronală Google în cazul textelor de VS care abordează sectorul de inovație într-un domeniu specific, ca medicina, de exemplu. Este important de văzut cum este tratat limbajul specializat în acest caz. În corpusul nostru, ce cuprinde multiple domenii de specializare în care inovația este prezentă, am remarcat că un termen dobândește mai multe variante de traducere în funcție de domeniul specializat din care fac parte. Un exemplu concret este termenul « enceinte » care este prezent în articolul N°2 în care se vorbește despre un asistent vocal, în articolul N°5 în care subiectul principal este centrala nucleară, și în articolul N° 45 în care este vorba de un stetoscop pentru Internet. În articolul N°2, acest lexem ar trebui să primească traducerea de *boxă* (fr. : haut-parleur) pentru că face parte din domeniul acusticii. Cu toate acestea, GTNM optează în cazul 2.1 pentru o echivalare a termenului prin *carcasă* (fr. : boîtier). Pe de cealaltă parte, în cazul ocurenței 5.6, lexemul este tradus prin *incintă* (fr. : espace clos, fermé à l'intérieur d'une construction) care restituie în mod potrivit sensul contextual.

Totodată, fișierul Excel în care am introdus greșelile ce provin de la TAN a Google poate deservi construcției unui glosar terminologic FR → RO în sectorul noilor tehnologii care va putea fi actualizat în mod regulat, chiar și după finalizarea studiului. Mai mult, suntem convinși că misiunea noastră în calitate de cercetător care a ales acest gen de teză consistă în propunerea celei mai bune soluții terminologice pentru conceptele care au fost deja introduse în română sub forma unui împrumut parțial sau total din engleză, o cale rapidă larg răspândită. Motivul nu este deloc surprinzător pentru că majoritatea inovațiilor au avut loc în țări anglofone, lucru care face ca engleza să domine lumea comunicării în materie de progres științific. Faptul că engleza a devenit limba principală de diseminare științifică a noilor tehnologii este un aspect ce nu poate fi negat.

Fiecare țară ar trebui totuși să țină la depărtare această tendință de suprimare a semnăturii naționale pusă pe un concept pentru că ea conduce la o pierdere a identității culturale și asta nu e decât un alt

efect al globalizării. Din păcate, influența culturii anglofone nu este resimțită doar în cazul denumirii conceptelor recente.

Academia română, un organism care poate influența acest fenomen datorită misiunii sale de protecție a identității lingvistice române nu are o influență prea mare când vine vorba de stoparea lui. Este greu de oprit valul de propagare a împrumuturilor în practica profesioniștilor care preferă uniformitatea și utilizează termenii în forma lor originală pentru e evita orice confuzie. Din acest motiv, credem că Academia română ar trebui să găsească căi de a rămâne la curent cu ultimele inovații din fiecare arie specializată pentru a putea publica în permanență propuneri lingvistice pe care să le transmită populației, dar mai ales profesioniștilor din sectorul de activitate în cauză. Acest lucru este indispensabil deoarece lumea tehnologică avansează cu o viteză halucinantă de la an la an. Putem menționa ca strategii care ar putea ajuta în acest sens, implementarea unor programe de sensibilizare cu privire la importanța conservării limbii române prin utilizarea de termeni români. Mai apoi, stabilirea de parteneriate cu media în vederea promovării acestui uz ar putea aduce rezultate bune. Cât despre măsuri un pic mai drastice, pot fi concepute legi de reglementare a terminologiei utilizate în documentele oficiale, publicitare ș.a.m.d. Subvențiile ar putea stimula companiile din domeniul științific să utilizeze o terminologie românească.

În ceea ce mă privește, pentru că sunt implicată în cercetarea lingvistică, mă simt obligată să mă angajez în lupta împotriva împrumuturilor engleze. Prin urmare, am creat un glosar - eșantion intitulat « Propositions linguistiques de substitution à la néo-terminologie technique anglaise », în care propun echivalenți 100% românești pentru termenii simpli și frazeologici moderni din corpusul nostru.

În concluzie, conform traducerilor furnizate în 2018/2019, s-ar părea că inteligența artificială nu va duce la dispariția profesiei de traducător, ci mai degrabă, o va schimba din nou. În ce-l privește pe traducător, pe acesta îl va echipa cu noi competențe tehnologice, transformându-l astfel într-un inginer al comunicării bilingve sau multilingve. Urmând această cale de dezvoltare, traducerea automată va continua să profite din ce în ce mai inteligent de glosare terminologice, de corpusuri de referință și mai ales, de contribuția traducătorilor umani. *In fine*, rămânem optimiști chiar și atunci când contextul pare sumbru și considerăm că aducerea pe scena IA a TAN este un omagiu adus traducătorului care nu va sfârși prin a dispărea.

Bibliografie

Dufay, B. (2005). *Apprendre à expliquer : l'art de vulgariser*. Eyrolles ;

Halliday, M. A. K., Hasan, R. (1976). Cohesion in English. Longmans ;

Jakobson, R. (1959). On linguistic aspects of translation. R. A. Brower (Éd.), *On translation.* Harvard University Press ;

Müller, B. (1985). *Le français d'aujourd'hui*. Éditions Klincksieck ;

Dubreil, E. (2008). Collocations : définitions et problématiques. *Texto !, Vol. XIII* (1). http://www.revue-texto.net/docannexe/file/126/dubreil_collocations.pdf ;

Meadows, J. (1986). Histoire succincte de la vulgarisation scientifique. *Impact : science et société, n° 144*, pp. 395-401. https://unesdoc.unesco.org/ark:/48223/pf0000071156_fre ;

Moirand, S., Reboul-Touré, S. Pordeus Ribeiro, M. (2016). La vulgarisation scientifique au croisement de nouvelles sphères d'activité langagière. *Bakhtiniana, Vol. 11* (2), pp.137-161. http://dx.doi.org/10.1590/2176-45732387 ;

Williams, G. (2003). Les collocations et l'école contextualiste britannique. *Travaux et recherches en linguistique appliquée, série E,* n°1, pp. 33-45 ;

Wang, W., Zhang, Z., Du, Y., Chen, B., Xie, J., Luo, W. (2021). Rethinking Zero-shot Neural Machine Translation:From a Perspective of Latent Variables. *Findings of the Association for Computational Linguistics : EMNLP 2021*, pp. 4321– 4327. doi : 10.18653/v1/2021.findings-emnlp.366 ;