(19) OFICIUL DE STAT
PENTRU INVENŢII ŞI MĂRCI
Bucureşti

ROMÂNIA

(11) **RO 134414 A0**
(51) Int.Cl.
G06F 40/205 (2020.01);
G06F 40/284 (2020.01);
G06F 40/30 (2020.01)

(12) # CERERE DE BREVET DE INVENŢIE

(21) Nr. cerere: **a 2019 00567**

(22) Data de depozit: **13/09/2019**

(41) Data publicării cererii:
28/08/2020 BOPI nr. **8/2020**

(71) Solicitant:
• **COGNOS BUSINESS CONSULTING S.R.L.**, *BD.IULIU MANIU NR.7, SECTOR 6, BUCUREŞTI, B, RO*

(72) Inventatori:
• **DASCĂLU MIHAI**, *BD.ION MIHALACHE, NR.126, BL.1, SC.A, ET.5, AP.37, SECTOR 1, BUCUREŞTI, B, RO;*
• **TOMA IRINA**, *STR.ION HELIADE RĂDULESCU NR.64, CĂLĂRAŞI, CL, RO;*
• **COTEŢ TEODOR-MIHAI**, *STR.ELENA DOMNEASCA NR.15, SLATINA, OT, RO;*
• **TRAUSAN-MATU ŞTEFAN**, *STR.PROF.DR.MIHAIL GEORGESCU NR.6, ET.1, AP.6, SECTOR 2, BUCUREŞTI, B, RO*

(74) Mandatar:
ROMPATENT DESIGN S.R.L., *STR.ŢEPEŞ VODĂ NR.130, ET.1, AP.C1, SECTOR 2, BUCUREŞTI*

(54) # METODĂ ŞI SISTEM DE ÎMBUNĂTĂŢIRE A STILULUI DE SCRIERE

(57) **Rezumat:**

Invenţia se referă la o metodă şi la un sistem de analiză automată a textului introdus de un utilizator într-un dispozitiv informatic, cu scopul de a oferi sugestii de corectare gramaticală a textului, precum şi de a îmbunătăţi stilul de scris din punct de vedere morfologic, sintactic şi semantic. Metoda conform invenţiei cuprinde: o fază de antrenare în care se folosesc două tipuri de corpusuri pentru crearea de modele specifice, respectiv, colecţii de documente de referinţă pentru un anumit domeniu, în vederea stabilirii de valori admisibile pentru diverşi indecşi de complexitate a textului, şi o colecţie de fraze greşite şi sugestii de corectare, pe care se aplică, într-o primă etapă, un algoritm de augmentare a corpusului, urmat într-o a doua etapă, de antrenarea unui model bazat pe reţele neuronale, axat pe corectarea automată la nivel morfo-sintactic a frazelor, şi o fază de testare în care, într-o primă etapă, se introduce un text pe dispozitivul utilizatorului, într-o a doua etapă se preprocesează textul folosind tehnici de procesare a limbajului natural, într-o a treia etapă se calculează indecşi de complexitate a textului care pot fi utilizaţi în caracterizarea stilului descris şi pot fi grupaţi prin intermediul unei descompuneri de tip PCA (Principal Component Analysis) în componente ce caracterizează diferite dimensiuni de scriere; într-o a patra etapă se aplică reguli ajustate fiecărui domeniu, ce verifică dacă valorile aferente indecşilor/componentelor sunt admisibile, raportat la colecţiile de documente de referinţă utilizate în etapa de antrenare; într-o a cincea etapă se aplică modelul de corectare antrenat pe colecţia de fraze greşite, iar într-o a şasea etapă se generează un feedback privind textul introdus, luând în considerare recomandările rezultate în etapele patru şi cinci. Sistemul conform invenţiei este integrat într-un dispozitiv informatic în care utilizatorul scrie un text sau încarcă un document de analizat.

Revendicări: 8
Figuri: 6

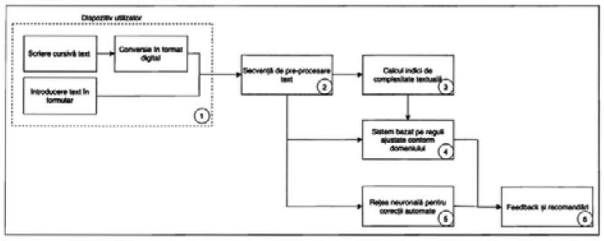

Fig. 1

RO 134414 A0

102

Metodă și Sistem de Îmbunătățire a Stilului de Scriere

Descriere

Prezentarea domeniului de aplicare

Invenția se referă la o metodă și un sistem de analiză automată a textului introdus de utilizator într-un dispozitiv (spre exemplu, tabletă, telefon mobil sau orice dispozitiv ce permite introducerea de text într-un formular), mecanism destinat utilizării în domenii variate, în funcție de specificitatea textelor procesate (spre exemplu, beletristică, jurnalistică, documente științifice). Sistemul oferă sugestii de corecție și de îmbunătățire a stilului de scris din punct de vedere morfologic, sintactic și semantic, respectiv de organizare a discursului și coeziune textuală, precum și corecții specifice la nivel de limbă.

Stadiul tehnicii

Brevetul US9465793B2 (http://patft1.uspto.gov/netacgi/nph-Parser?patentnumber=9465793) prezintă două etape de evaluare a greșelilor gramaticale. Prima etapă este analiza gramaticală a textului folosind diverși algoritmi sau metode automate de corecție, iar în a doua etapă textul este corectat de către un evaluator uman. Prima etapă oferă feedback personalizat ce cuprinde și textul inițial din care a fost identificată eroarea. Dezavantajul metodei este că evaluarea este făcută doar din punct de vedere gramatical, fără a evalua coeziunea textului.

Brevetul US5678053A (http://patft1.uspto.gov/netacgi/nph-Parser?patentnumber=5678053) prezintă o modalitate de a evidenția greșelile gramaticale dintr-un text prin sublinierea acestora și afișarea unui simbol „V" întors pentru a marca lipsa unui cuvânt.

Dezavantajul acestei reprezentări este faptul că nu se pot evidenția sugestiile de înlocuire a cuvintelor și că sunt considerate exclusiv greșeli gramaticale.

Brevetul WO1997049043A1 (https://worldwide.espacenet.com/publicationDetails/originalDocument?CC=WO&NR=9749043A1&KC=A1&FT=D&ND=4&date=19971224&DB=) prezintă o metodă care sugerează corecții gramaticale și de scriere pentru documente electronice. Dezavantajul acestei metode este că evaluarea nu oferă sugestii de corecții semantice.

Brevetul US6988063B2 (http://patft1.uspto.gov/netacgi/nph-Parser?patentnumber=6988063) descrie reprezentarea gramaticală a propozițiilor sub forma unui arbore, construit utilizând un parser de tip POST (en. part-of-speech tagged parser). Arborele obținut este comparat cu structuri preexistente în sistem, identificându-se astfel posibile greșeli gramaticale. Similar cu brevetele anterioare, dezavantajul acestei propuneri este faptul că soluția este limitată la limba engleză și nu oferă sugestii de corecții semantice.

Brevetul US20070271510A1 (http://patft1.uspto.gov/netacgi/nph-Parser?patentnumber=20070271510) identifică greșelile din textele conținute de pagini web prin aplicarea unui set de instrucțiuni și consultarea unui modul de identificare de erori. Dezavantajul soluției este faptul că nu acceptă alt format în afară de pagini web.

Brevetul US20090192787A1 (http://patft1.uspto.gov/netacgi/nph-Parser?patentnumber=20090192787) prezintă o metodă de identificare a greșelilor gramaticale pe baza unui set de reguli bine definite. Regulile sunt definite pentru toate legăturile de text posibile. Metoda include și un parser ce se bazează pe un set de reguli logice pentru identificarea părților de vorbire, și folosește legăturile dintre cuvinte pentru a marca dacă legăturile gramaticale sunt corecte sau nu. Dezavantajul acestei metode este că aceasta presupune generarea tuturor combinațiilor gramaticale. Mai mult,

metoda nu oferă sugestii de corecţie din punct de vedere semantic, al coeziunii textului sau informaţii statistice despre tipul părţilor de vorbire din text.

Brevetul US20150019207A1 (http://patft1.uspto.gov/netacgi/nph-Parser?patent number=20150019207) propune folosirea ontologiilor pentru a determina contradicţiile logice la nivel gramatical şi oferirea de feedback conform axiomelor ontologiei. Dezavantajul acestei metode este că se identifică doar erorile gramaticale.

Brevetul US6012075A (http://patft1.uspto.gov/netacgi/nph-Parser?patentnumber= 6012075) presupune verificarea gramaticală a câte unei propoziţii din inputul introdus de către un utilizator, în timp ce acesta face o pauză în scris. Verificarea se face propoziţie cu propoziţie, iar sugestiile de corecţie sunt afişate utilizatorului la selectarea erorii, într-o fereastră pop-up, sub propoziţia respectivă. În fereastră sunt afişate primele 3 sugestii de corecţie, plus toată propoziţia corectată. Dezavantajul acestei metode este că propoziţiile se verifică doar dacă există o pauză în scriere, iar ca intrare este acceptat doar formatul text.

Brevetul RO127582A2 prezintă o metodă de inserare automată a diacriticelor în texte în limba română scrise fără diacritice. Lista de cuvinte ce pot accepta diacritice este trecută prin trei nivele de filtrare: unigrame, bigrame şi trigrame, fiecare nivel eliminând variantele incorecte ale cuvintelor. Dezavantajul sistemului este că nu se ia în considerare toată fraza în care se află cuvântul care trebuie restaurat, şi se folosesc exclusiv apariţii statistice, fără a folosi reprezentarea contextuală a fragmentelor de text. Totodată, brevetul adresează doar o componentă foarte specifică din cadrul invenţiei curente.

Brevetul US6115683A (http://patft1.uspto.gov/netacgi/nph-Parser?patentnumber= 6115683) prezintă o metodă de notare automată a eseurilor. Iniţial, eseul este trecut printr-un pas de simplificare morfologică, în urma căruia este salvat într-un arbore de

parsare. Pe baza arborelui de parsare și a unui lexicon definit într-un fișier extern se identifică noduri frazale asociate cu eseul. Notarea se face prin numărarea potrivirilor între nodurile frazale și un set de reguli predefinite de sistem. Dezavantajul acestui sistem este că regulile definite nu au o pondere asociată, astfel că toate greșelile din eseu sunt tratate la fel. În plus, sistemul consideră doar arborele de parsare, fără a utiliza informații de semantică și discurs aferente.

Brevetul US6181909B1 (http://patft1.uspto.gov/netacgi/nph-Parser?patentnumber=6181909) propune o soluție pentru notarea automată a eseurilor. Arhitectura sistemului cuprinde un parser, un sistem pentru extragerea unui vector de caracteristici sintactice, un sistem pentru extragerea unui vector de caracteristici retorice, două sisteme pentru extragerea caracteristicilor de scor, și un sistem final ce generează nota finală pe baza caracteristicilor anterioare. Dezavantajul acestei abordări constă în lipsa profunzimii generate de analizele semantice și la nivelul discursului, precum și lipsa suportului multilingv.

Brevetul US8472861B2 (http://patft1.uspto.gov/netacgi/nph-Parser?patentnumber=8472861) prezintă o metodă de evaluare automată a eseurilor. Etapele metodei sunt: derivarea unui set de caracteristici prestabilite din eseu, notarea și ponderarea acestor caracteristici folosind diverse ecuații, generarea unei note brute pe baza ecuațiilor, și interpretarea notei folosind un algoritm adaptiv. Caracteristicile identificate de sistem sunt: numărul de cuvinte al eseului, procentul de erori gramaticale/sintactice/de stil raportat la lungimea eseului, numărul optim de elemente de discurs, numărul mediu de cuvinte ale elementelor de discurs, categoria din care face parte eseul pe baza similarității semantice între textul eseului și un vocabular, scorul de similaritate între eseu și cel mai bine notat eseu din categoria identificată anterior, raportul între numărul de cuvinte și numărul de token-uri din eseu, nivelul vocabularului și dimensiunea medie

a cuvintelor. Dezavantajul acestei abordări este determinat de lipsa analizei la nivelul discursului, elemente succinte la nivelul secvenței de pre-procesare a textului, precum și lipsa suportului multilingv.

Prezentarea sintetică a invenției

Problema tehnică pe care invenția a propus să o rezolve poate fi formulată pornind de la constatarea că, odată cu evoluția mediilor online, comunicarea scrisă devine principalul mod prin care oamenii se exprimă; prin urmare, textele transmise trebuie să fie coerente și corecte din punct de vedere gramatical, morfologic și sintactic, precum și adaptate domeniului și audienței specifice. Metoda care face obiectul prezentei invenții adresează problema tehnică menționată anterior prin faptul că analizează automat textul introdus de către utilizator direct în dispozitivul acestuia (spre exemplu, tabletă, telefon mobil, sau orice dispozitiv care permite introducerea de text într-un formular) și oferă sugestii de corecție și de îmbunătățire a stilului de scris considerând: a) elemente de suprafață (spre exemplu, abundență sau utilizare incorectă de semne de punctuație, fraze excesiv de scurte sau lungi ca număr de cuvinte), b) erori gramaticale, c) greșeli sintactice (spre exemplu, densitate prea mare de anumite părți de vorbire), d) probleme semantice (spre exemplu, propoziții cu coeziunea locală mică raportat la contextul semantic, repetiții), e) elemente de discurs (spre exemplu, conectori de discurs utilizați excesiv, probleme de coeziunea globală care reflectă o incoerență a ideilor), precum și f) corecții specifice unei anumite limbi (spre exemplu, pentru limba română, lipsa diacriticelor, disonanțe, probleme de acord între diverse părți de vorbire, utilizarea excesivă a diatezei pasive sau a modului gerunziu).

Avantaje

Sistemul și metoda conform invenției prezintă următoarele avantaje:

- Sistemul acceptă un format divers de date de intrare, de exemplu, text introdus într-un formular, text în format digital scanat, text în format cursiv convertit în text în format digital, sau orice imagine care conține text.
- Sistemul este furnizat înglobat în cadrul dispozitivului utilizatorului, oferind feedback în timp real în momentul scrierii textelor.
- Metoda propusă oferă suport multilingv.
- Nu există un sistem similar disponibil pentru limba română.
- Sistemul încadrează textul introdus de utilizator într-o anumită categorie în funcție de stilul utilizat (de exemplu, științific, oficial, publicistic, beletristic sau colocvial), analiza textuală fiind specifică domeniului identificat.
- Sistemul oferă feedback utilizatorului la patru niveluri diferite de granularitate, și anume: cuvânt, propoziție/frază, paragraf și întregul document.
- Sistemul prezintă rezultatul analizei automate într-un mod vizual, folosind ca reprezentare o hartă de culori. Fiecare element de text (de exemplu, cuvânt, propoziție, paragraf, document) este identificat printr-un dreptunghi colorat conform nivelului de severitate al greșelilor identificate. O culoare cu luminozitate scăzută reprezintă un nivel ridicat de severitate, iar o culoare cu un grad ridicat de luminozitate, nivel scăzut de severitate.
- Sistemul de restaurare de diacritice integrat în cadrul invenției prezintă o mai bună contextualizare. Spre deosebire de un sistem bazat pe n-grame, care poate lua în considerare contexte de maxim 4-grame, 5-grame din cauza limitărilor de memorie, sistemul propus ia în considerare toată frază în care se află cuvântul care se dorește a fi restaurat. De asemenea, lipsa n-gramelor face ca sistemul să poată fi folosit pe

dispozitive cu memorie limitată, întrucât modelul nu necesită decât stocarea ponderilor modelului și a vectorilor cuvintelor din dicționar.

- Sistemul de corecții bazat pe rețele neuronale conform invenției prezintă principalul avantaj față de metodele prezentate anterior, prin faptul că poate oferi sugestii de corecții mult mai complexe, implicând schimbarea, adăugarea sau ștergerea unui număr extins de cuvinte, fiind capabil de reformulări complexe ale frazelor inițiale. Acest avantaj este datorat naturii generative a modelului, acesta generând fraza corectă pornind de la fraza greșită, asigurând totodată o libertate extinsă în schimbările pe care le poate aplica. Niciun brevet prezentat anterior nu folosește metode generative pentru corecții, limitând astfel schimbările care pot fi aplicate.
- Sistemul propus combină atât metode statistice bazate pe indici desprinși din textul inițial care reflectă stilul de scriere al autorului, cât și rețele neuronale utilizate pentru generarea de corecții subtile.

Prezentarea figurilor

În continuare, invenția va fi descrisă în detaliu, cu referire la Figura 1, care prezintă schema bloc a metodei propuse ce preia textul de analizat de la utilizator prin diverse modalități: scrierea cursivă de text și conversia acestuia în format digital (spre exemplu, tabletă sau dispozitiv mobil pe care se scrie text de mână și care este convertit în format digital), respectiv încărcarea textului propriu-zis într-un formular specific. Figura 2 detaliază secvența de pre-procesare a textului folosind tehnici de procesare a limbajului natural. Ulterior, Figura 3 introduce exemple de indicii de complexitate textuală utilizați în furnizarea de feedback. Figurile 4 și 5 detaliază arhitectura de rețele neuronale pentru generarea automată iterativă de greșeli gramaticale și corecții aferente. Ulterior, rezultatele modulelor anterioare sunt

prezentate în interfața prin intermediul modulului de feedback și recomandări (Figura 6).

Descrierea detaliată a invenției

Metoda propusă necesită primirea textului de analizat de la utilizator. Aceasta se poate face prin diverse variante, de exemplu: încărcarea unei imagini sau a unui document scanat ce conține text scris de mână sau tipărit, respectiv prin scrierea directă a unui text într-un formular pus la dispoziție de sistem. În cazul încărcării unui document, acesta se trece printr-un proces de recunoaștere optică a caracterelor pentru a transforma textul într-un format utilizat de calculator. În continuare, textul trece prin fazele de procesare descrise mai jos.

Secvența de pre-procesare a textului

Secvența de pre-procesare a limbajului natural din Figura 2 include o multitudine de procesări specifice limbajului natural, grupate în următoarele categorii: restaurarea diacriticelor (dacă este aplicabilă pentru o anumită limbă) (2.1), prelucrări preliminare (2.2), prelucrări independente (2.3), precum și corecții suplimentare (2.4).

Restaurarea diacriticelor (2.1) este un pas obligatoriu pentru prelucrarea adecvată a textelor în limba română, întrucât diacriticele modifică morfologia, valoarea gramaticală a unui cuvânt („cană" și „cana") sau chiar sensul cuvântului („fată" și „față"). Sistemul conform invenției folosește un model ce combină informațiile lexicale cu cele semantice, folosind rețele neuronale cu scopul de a capta contexte complexe. Modelul este antrenat folosind un corpus artificial, format din texte bine-formatate, din care au fost scoase diacriticele. Predicția se face pentru fiecare caracter ce poate accepta diacritice. Arhitectura sistemului este formată din trei ramuri, fiecare primind date de intrare diferite [*Ruseti, S., Cotet, T.-M., & Dascalu, M. (2018). Romanian Diactrics Restoration using Recurrent Neural Networks. In V. Pais, D. Gifu, D. Trandabat, D.*

Cristea & D. Tufis (Eds.), 13th Int. Conf. on Linguistic Resources and Tools for Processing Romanian Language (ConsILR 2018) (pp. 61–68). Iasi, Romania]. Prima cale din cadrul rețelei neuronale este reprezentată de un codificator de memorie bidirecțional - Bidirectional Long Short-Term Memory (BiLSTM) [*Graves, Alex, & Schmidhuber, Jürgen. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural networks, 18(5-6), 602–610*] utilizând reprezentări vectoriale ale caracterelor (character embeddings). O fereastră de dimensiune fixă este utilizată pentru a reprezenta contextul caracterului curent. Rețeaua poate învăța diferite asemănări între litere prin utilizarea acestor reprezentări vectoriale. De asemenea, aplicarea codificatorului la nivelul caracterelor permite rețelei să generalizeze diferite forme ale aceluiași cuvânt, care vor avea reprezentări foarte asemănătoare și, de asemenea, vor generaliza pentru cuvinte neîntâlnite, dacă arată similar cu alte concepte cunoscute. Cu toate acestea, modelul poate beneficia și de informații semantice. Pentru a realiza acest lucru, a fost adăugată cea de-a doua cale a arhitecturii, reprezentată de un codificator (spre exemplu, BiLSTM) aplicat pe propoziția actuală. Cuvintele din propoziție sunt reprezentate de reprezentări vectoriale FastText pre-antrenate [*Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135–146*], care mapează cuvintele cu puncte într-un spațiu vectorial multi-dimensional, bazat pe contextul în care acestea apar. Presupunerea este că cei doi codificatori din cadrul BiLSTM (de la începutul textului, respectiv de la final) sunt complementari unul cu celălalt, iar modelul poate învăța cum să combine informațiile din ambele sensuri. Cea de-a treia cale este reprezentată de utilizarea reprezentării vectoriale a cuvântului, respectiv a propoziției actuale. Aceasta

este adăugată pentru a ajuta rețeaua să selecteze care parte din contextul codificat este utilă pentru intrarea curentă.

Tokenizarea (2.2.1) stabilește segmentarea textului în entități unitare, folosindu-se, în general, delimitarea după spații. Pentru excepții, cum ar fi „let's go" pentru limba engleză, despărțirea se face în funcție de sufixe și prefixe, caracterul „ ' " este considerat sufix și se face o delimitare suplimentară.

Următoarea etapă, *eliminarea cuvintelor de tipul "stop-words"* (2.2.2) este importantă, întrucât aceste cuvinte nu au conținut semantic și, prin urmare, nu ajută la definirea contextului, spre exemplu: pronume (acea, aceasta, acești, acest, acele, toți etc.), conectori (și, sau, între etc.), numerale (opt, șase, nouăsprezece etc.). O strategie care poate fi folosită este sortarea termenilor după frecvența totală de apariție și eliminarea selectivă a celor mai frecvenți.

În a treia etapă se realizează *adnotarea cu părți de vorbire* (2.2.3), reprezentând clasificarea fiecărui cuvânt conform părții de vorbire corespunzătoare. Adnotarea este bazată pe o rețea neuronală, formată din 3 straturi ascunse dense cu aproximativ 500 neuroni, folosind ca funcție de activare ReLU (Rectified Linear Units) [*Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th Int. Conf. on Machine Learning (ICML-10) (pp. 807–814)*]. Pentru transformarea valorilor finale în probabilități (care reprezintă probabilitatea ca un cuvânt să fie o anumită parte de vorbire) se poate folosi funcția softmax, iar pentru calcularea erorii, cross-entropia care măsoară similaritatea între două distribuții de probabilitate. Pentru antrenarea rețelei se poate folosi algoritmul de optimizare Adam, iar pentru eliminarea overfitting-ului (problemă care poate apărea la programele de învățare automată), tehnica „dropout" [*Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural*

networks from overfitting. The Journal of Machine Learning Research, 15(1), 1929–1958] care în timpul antrenării va ignora o parte dintre neuroni la fiecare actualizare.

Lematizarea (2.2.4) reprezintă gruparea formelor flexionare ale unui cuvânt, astfel încât acestea să poată fi analizate ca o singură entitate, care reprezintă lema cuvântului. Uneori prelucrarea este dificilă deoarece cuvinte care se scriu diferit („fi", „sunt", „ești") au aceeași lemă. Lematizarea propusă în cadrul acestui sistem este statică, bazată pe dicționar.

Analiza sintactică bazată pe dependențe (2.2.5) are ca scop identificarea structurii gramaticale (sintactică) a propozițiilor. De exemplu, pentru o propoziție se va determina cine este subiectul sau obiectul unui verb. În continuare, este prezentată o posibilă metodă de implementare pentru determinarea acestor dependențe bazată pe tranziții. Metoda scanează fiecare propoziție și construiește arborele de dependențe cu o complexitate computațională în timp liniar. La fiecare pas păstrează o stivă de cuvinte care sunt încă în procesare și un buffer (tampon) de cuvinte care urmează a fi procesate. Buffer-ul împreună cu stiva definesc starea curentă. În starea inițială, toate cuvintele se află în buffer, iar stiva conține un singur nod și anume nodul de tip rădăcină („root"). Pe orice stare se pot realiza trei tranziții posibile: 1) arc-stânga: marchează al doilea element de pe stivă ca dependent al primului și îl elimină de pe stivă (se poate aplica doar dacă stiva conține minimum două elemente); 2) arc-dreapta: marchează primul element de pe stivă ca dependent al celui de-al doilea și îl elimină (se poate aplica doar dacă stiva conține minimum două elemente); 3) deplasare (shift): elimină un cuvânt din buffer și îl pune pe stivă. Parserul decide ce tranziție să aleagă, folosind un clasificator bazat pe o rețea neuronală care va avea ca intrare cuvintele din propoziție reprezentate ca vectori.

Ulterior sunt efectuate prelucrări de text care nu sunt interdependente. *Recunoașterea entităților cu nume* (2.3.1) este o etapă centrată pe identificarea grupurilor de cuvinte care reprezintă o entitate reală, cu nume (de exemplu, nume de persoane, de orașe sau țări, de instituții sau firme etc.). În general acest proces îmbunătățește alte prelucrări cum ar fi parsarea, adnotarea etc. De multe ori entitățile în cauză sunt formate din mai multe cuvinte, au mai multe sensuri sau definesc aceeași entitate. Spre exemplu, „Universitatea de Vest din Timișoara" trebuie tratată ca o singură entitate. Așadar, un model de predicție este necesar pentru a decide dacă un grup de token-uri face parte din aceeași entitate. Pe lângă acest lucru, trebuie să subliniem că unele entități au mai multe sensuri. „Universitatea Craiova" este și o echipă de fotbal, dar și o denumire posibilă pentru Universitatea din Craiova, așadar este necesară și o eventuală dezambiguizare. Grupuri diferite de token-uri ("Președintele Barack Obama" vs. "Obama") trebuie însă recunoscute ca fiind aceeași entitate.

Un alt tip de procesare este *dezambiguizarea sensurilor cuvintelor* (2.3.3), mai precis asocierea unui singur sens fiecărui cuvânt. Pentru aceasta se poate folosi WordNet (un dicționar digital care grupează cuvintele pe bază de sinonime) [*Miller, G.A. (1998). Foreword. In C. Fellbaum (Ed.), WordNet: An electronic lexical database (pp. xv–xxii). Cambridge, MA: MIT Press*], construindu-se un graf cu toate posibilele interpretări ale cuvintelor din text, și alegându-se ulterior sensul cel mai probabil pentru fiecare cuvânt. Ultimul pas din cadrul categoriei de prelucrări independente este *identificarea coreferințelor* (2.3.3). O coreferință reprezintă un grup nominal (persoană, obiect) care este menționat într-o anumită parte a textului și este referit (în general mai târziu în text) prin intermediul unui pronume, sau sub o altă formă de grup nominal, în altă parte a textului. Un lanț de coreferințe reprezintă mulțimea acestor forme care referă același element, într-un text existând mai multe astfel de lanțuri. Pentru coreferințe sunt

descrișii următorii indici: numărul mediu de cuvinte per lanț de coreferințe, media sau maximul întinderii lanțurilor de coreferințe (distanța între pronume și mențiunea corespunzătoare).

Pentru *corecții suplimentare* (2.4) se folosește drept corpus de referință un set de fișiere PDF ce conțin greșeli întâlnite în diverse canale media, TV și radio. Regulile generate pe baza analizei corpusului adresează următoarele categorii: disonanțe, repetiții și greșeli de punctuație. Disonanțele sunt cauzate de alăturarea unor silabe asemănătoare, cele mai frecvente în limba română fiind cauzate de: „la/la", „sa/sa", „ca/ca", „ca/ce", „ca/ci", „că/ca", „cu/co", „că/co", „că/cu", „că/când". Pentru a evita disonanțele, multe persoane folosesc construcția „ca și", numită „și parazitar". Există cazuri însă, când construcția „ca și" este utilizată corect, anume ca sinonim pentru „la fel ca". Disonanța și construcția „și parazitar" se identifică folosind șabloane descrise prin expresii regulate.

Repetițiile se împart în două categorii: a) repetarea aceluiași cuvânt, indiferent de forma morfologică și b) repetarea sinonimelor. Pentru primul caz de repetiție se consideră deranjantă apariția aceluiași cuvânt într-o frază, într-o fereastră de o anumită dimensiune. Pentru a putea analiza fraza, aceasta se aduce la un format unitar prin eliminarea cuvintele de tip stop-words, lematizarea și extragerea rădăcinii acestora. Având fraza redusă la un format unitar, se alege o fereastră de lungime fixă (cuvinte pentru a marca repetițiile). De exemplu, într-o fereastră de 3 cuvinte ce conține „Copac*ul* este copac", cuvântul „copacul" se reduce la forma de dicționar, „copac", care este apoi identificată drept repetiție. Al doilea caz se recunoaște prin repetarea sinonimelor aceluiași cuvânt într-o fereastră similară, din care au fost extrase cuvintele de tip stop-words, dar nu s-au aplicat tehnici de lematizare. Pentru a detecta dacă două cuvinte sunt sinonime, se poate folosi WordNet (disponibil în mai multe limbi),

sau resurse specifice fiecărei limbi (spre exemplu, Dex Online – un dicționar pentru limba română, din care se extrage lexiconul). Două cuvinte sunt considerate sinonime dacă intersecția lexiconului acestora are cel puțin un element. Astfel, în propoziția „Pomul este un copac, sau un arbore" se va detecta o repetiție.

Greșelile de punctuație se identifică folosind expresii regulate. Sunt detectate greșeli rezultate din nerespectarea următoarelor reguli: 1) semnele de punctuație se alipesc de cuvântul precedent și sunt separate de cuvântul următor printr-un spațiu; 2) dacă o frază este formată din propoziții coordonate, detectate prin folosirea conjuncțiilor „și", „sau", „ori", acestea nu sunt urmate de virgulă; 3) cele mai multe conjuncții adversative („dar", „nici", „iar", „însă", „ci și", „dar și", „precum și", „deci", „prin urmare" și „așadar") sunt succedate de virgulă; altele (spre exemplu, „în concluzie") sunt precedate de virgulă; 4) adverbele intercalate în propoziție sunt precedate și succedate de virgulă: „desigur", „firește", „așadar", „bineînțeles", „în concluzie", „în realitate", „de exemplu".

În plus față de aceste greșeli, pentru limba română, se identifică greșeli legate de acordul între subiect și predicat, corelația între substantive și alte părți de vorbire (adjectivul, numeralul și articolul nehotărât).

Calcularea indicilor de complexitate textuală

Indicii de complexitate textuală acoperă multiple categorii, fără a se limita la: metrici de suprafață (3.1), indici de sintaxă și morfologie (3.2), respectiv indici de semantică, coeziune și structura discursului (3.3). În scop demonstrativ, includem în paragrafele următoare exemple relevante de indici care pot fi integrați în sistem în vederea furnizării de feedback.

Metricile de suprafață (3.1) includ *elemente de suprafață* (3.1.1) ce oferă atribute statistice raportat la distribuția diverselor elemente textuale (spre exemplu, cuvinte, propoziții, paragrafe), respectiv semne de punctuație prezente în text. De exemplu:

numărul mediu de cuvinte la nivel de propoziție sau paragraf, numărul mediu de virgule la nivel de propoziție sau paragraf, lungimea medie de caractere din fiecare propoziție sau paragraf etc.

Entropia la nivel de cuvânt, caracter și n-grame (3.1.2) oferă o perspectivă relevantă asupra complexității textuale la nivel de literă, cuvânt și n-grame, prin evaluarea diversității între elementele analizei. Astfel, un text mai complex conține mai multe informații și necesită mai multă memorie și mai mult timp pentru ca un cititor să îl proceseze. Entropia se măsoară la nivel de caracter, cuvânt și n-grame de caractere (3, 4 sau 5 caractere).

Complexitatea cuvintelor (3.1.3) poate include informații cu privire la: numărul silabelor, distanța dintre forma flexionată, lemă și rădăcină, în timp ce specificitatea se reflectă în frecvența inversă a documentelor din corpurile de antrenare, numărul de sensuri din ontologie. Indicii asociați cuvintelor sunt calculați într-o manieră simplă, prin medierea valorilor relevante pentru toate cuvintele de conținut din text (numai cuvintele din dicționar, lematizate, neincluse în listele de cuvinte de tip stop-words, având ca parte de vorbire: substantiv, verb, adverb sau adjectiv). În ceea ce privește polisemia medie a unui cuvânt, se ia în calcul presupunerea potrivit căreia, cu cât un cuvânt are mai multe sensuri posibile, cu atât este mai dificil de utilizat într-un text și de identificat sensul corect. Prin urmare, textele mai simple vor conține cuvinte mai puțin ambigue, în timp ce textele complexe vor conține mai multe cuvinte polisemantice. Suplimentar, distanța dintre poziția unui cuvânt în arborele de hipernime, care leagă cuvintele de concepte cu sensuri mai generale, și rădăcina acestui arbore, poate fi văzută ca o măsură a specializării și specificității cuvântului. Cu alte cuvinte, cu cât calea către rădăcina ierarhiei din ontologie este mai lungă, cu atât cuvântul este mai specific.

Din categoria *sintaxă și morfologie* (3.2) menționăm indicii care reflectă *statistici la nivel de densități de părți de vorbire* prezente în text (3.2.1). Aceștia sunt importanți pentru analiza sintactică a textelor, deoarece frecvența diverselor părți de vorbire, cum ar fi prepozițiile, adjectivele și adverbele, dictează o structură complexă și profundă a textului. Din această categorie se amintesc: numărul mediu de cuvinte care aparțin unei anumite părți de vorbire (inclusiv substantive, verbe la gerunziu, adjective, pronume la nivel de propoziție sau paragraf, la persoanele I, a II-a, a III-a, indefinite sau interogative).

Pentru determinarea *statisticilor la nivelul unor dependențe sintactice specifice* (3.2.2) se pot folosi modele bazate pe rețele neuronale antrenate pe corpusuri disponibile în frameworkul Universal Dependencies (spre exemplu, RoRefTrees [*Barbu Mititelu, V., Ion, R., Simionescu, R., Irimia, E., & Perez, C.-A. (2016). The Romanian Treebank Annotated According to Universal Dependencies. In Proceedings of the Tenth Int. Conf. on Natural Language Processing (HrTAL2016). Dubrovnik, Croatia*]) care conține texte adnotate manual din diverse domenii: literatură, medicină, scrieri academice, etc. Pe baza acestor adnotări se învață 50 de tipuri de dependențe, dintre care 11 sunt specifice limbii române. Exemple de dependențe: obiectul verbului, obiectul indirect, modificator adjectival, conjuncție coordonatoare etc.

Indicii de complexitate la nivel de n-grame (3.2.3) cuprind numărul celor mai frecvente bigrame și trigrame (secvențe continue de 2 sau 3 elemente din text) de cuvinte la nivel de frază și paragraf, precum și indici ce reprezintă gradul de coeziune între cuvinte pe baza unui corpus din care se vor învăța statistic respectivele *n*-grame.

Coeziunea locală și globală (3.3.1) este o măsură medie a similarității semantice, a apropierii dintre segmentele textuale, care pot fi cuvinte, fraze, paragrafe sau întregul text. Această similaritate semantică ia în considerare, pe deoparte, proximitatea

lexicală, identificată ca distanțe semantice în WordNet, și pe de altă parte, similaritatea semantică. Similaritatea semantică poate fi evaluată folosind diverse modele semantice, spre exemplu LSA [*Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. Discourse Processes, 25(2/3), 259–284*], LDA [*Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3(4-5), 993–1022*], word2vec [*Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representation in Vector Space. In Workshop at ICLR. Scottsdale, AZ*] sau Glove [*Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. In The 2014 Conference on Empirical Methods on Natural Language Processing (EMNLP 2014). Doha, Qatar: ACL*]. Din această categorie fac parte: coeziunea adiacentă sau ierarhică între elementele de text, coeziunea între primul și ultimul element de text (introducere și concluzie), coeziunea între primul element de text și celelalte elemente etc.

Pentru structura discursului se pot folosi expresii regulate pentru a recunoaște utilizarea *conectorilor de discurs specifici* (3.3.2). Metricile identificate sunt numărul mediu de conectori la nivel de paragraf din categoriile: adăugire, concesii, condiții, conectori temporali, conectori de coordonare, conectori logici, conjuncții, conjuncții coordonatoare, contraste, disjuncții, opoziții, ordine etc.

Statisticile cu privire la rezoluțiile de coreferințe (3.3.3) referă indici de complexitate textuală definiți pe lanțuri de coreferințe, de exemplu: numărul mediu de cuvinte per lanț de coreferințe, media și maximul întinderii lanțurilor de coreferințe (distanța între pronume și mențiunea corespunzătoare).

Sistem bazat pe reguli ajustate conform domeniului

Ca urmare a folosirii unui număr mare de indici, interpretarea acestora este destul de dificilă, prin urmare se folosește o metodă care grupează indicii în componente care

caracterizează diverse dimensiuni ale scrisului, Principal Component Analysis (PCA) [*Jolliffe, I. (2011). Principal Component Analysis International encyclopedia of statistical science (pp. 1094–1096). Springer Berlin*]. Înainte de gruparea în componente se aplică următoarele filtre: eliminarea indicilor de localitate și a documentelor de tip outliner, verificarea pentru normalitate și eliminarea indicilor puternic corelați între ei. Astfel, se elimină: 1) indicii care au acoperire lingvistică scăzută, adică mai mult de 80% dintre scorurile pentru toate documentele sunt 0 sau 1; 2) documentele pentru care mai mult de 10% din indici sunt de tip outliner; 3) indicii care corespund distribuțiilor statistice anormale raportate la nivelul aplatizării și simetriei și 4) indicii care sunt puternici corelați cu alți indici, pe baza coeficientului de corelație Pearson.

Toții indicii rămași după etapele de filtrare se grupează în componente ortogonale. Fiecare componentă generată primește un nume pe baza indicilor de complexitate textuală din cadrul acesteia. Componentele rezultate sunt folosite pentru generarea de feedback personalizat prin adăugarea lor în sistemul bazat pe reguli. Sistemul conține două tipuri de reguli: reguli bazate pe scorurile componentelor PCA (care conțin mai mulți indici agregați) rezultate din analiza anterioară, și reguli bazate pe scorurile individuale ale unor indici. Pentru fiecare regulă se stabilește un interval [minim; maxim], ajustat conform domeniului, precum și mai multe mesaje corespunzătoare pentru cazul în care valoarea aferentă documentului nu se află în acest interval, cu scopul de a evita monotonia.

Rețeaua neuronală pentru efectuarea de corecții automate

Componenta (5) din Figura 1 se referă la detectarea și corectarea automată a textelor prin intermediul unei rețelelor neuronale artificiale de tipul codificator-decodificator. O parte netrivială a acestei componente o reprezintă și generarea corpusului pentru

antrenarea rețelei neuronale; întrucât modelul folosit este unul supervizat, acesta necesită un corpus adnotat. Astfel, pornind de la un număr limitat de exemple de tipul sursă-țintă (ordin de mărime de minim 10 de mii de perechi), unde sursa reprezintă o frază incorectă gramatical, iar ținta reprezintă aceeași frază corectată, se antrenează o rețea neuronală care învață să modifice fraze corect gramaticale în fraze incorecte, în mod asemănător cu greșelile prezente în corpusul inițial. Arhitectura acestei rețele (prezentată în Figura 4) este una de tipul codificator-decodificator, bazată pe rețele neuronale recurente (RNR, spre exemplu folosind celule de tip GRU [*Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint, arXiv:1406.1078*], LSTM [*Hochreiter, Sepp, & Schmidhuber, Jürgen. (1997). Long short-term memory. Neural computation, 9(8), 1735–1780*] sau Transformer [*Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Paper presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*]), la care se adaugă un mecanism de atenție (5.3) pentru construirea la fiecare pas a vectorului contextual (5.4). În plus, pentru a spori zgomotul (diversitatea) frazelor generate de către rețea, algoritmul de *beam-search* din cadrul decodificatorului (5.5) este modificat prin adăugarea la fiecare pas a unei valori $k * \beta$, distribuții de probabilitate ale frazelor candidat. Variabila k este ales aleator din intervalul [0, 1], iar β este o valoare fixată. Pentru un β suficient de mare, această modificare are efectul de a reordona aleator candidații algoritmului de *beam-search*. Astfel, această metodă sporește corpusul inițial cu greșeli gramaticale, care ulterior îmbunătățesc performanțele modelului final. Fără această modificare algoritmul clasic

are tendința de a genera doar fraze corecte, pentru că majoritatea exemplelor văzute la antrenare conțin sintagme corecte.

Componenta de corecții este similară cu arhitectura prezentată în Figura 5, cu mențiunea că aceasta este antrenată folosind exemple având ca sursă fraza greșită, și ca țintă fraza corectă, spre deosebire de rețeaua neuronală folosită pentru generare de fraze greșite. De asemenea, algoritmul de *beam-search* rămâne în varianta lui clasică.

Furnizarea de feedback și recomandări

Modulul de feedback și recomandări prezentat în Figura 6 evidențiază o metodă de afișare a feedback-ului și a recomandărilor către utilizatorul final. Acesta are posibilitatea de a selecta gradul de granularitate pentru care se afișează feedback-ul și recomandările, din următoarele variante: cuvânt, propoziție/frază, paragraf sau document (6.1). Odată aleasă o opțiune, partea de jos a interfeței se modifică în consecință. A doua parte a interfeței este împărțită în trei părți, anume: textul original (6.2), feedback-ul generat la nivelul de granularitate selectat (6.3), și reprezentarea grafică a greșelilor identificate (6.4). Severitatea greșelilor din textul original este marcată prin sublinierea textului, conform granularității selectate (de exemplu la nivel de propoziție), folosind linii de dimensiune și culoare diferite. Grosimea liniilor scade în dimensiune, iar luminozitatea crește cu cât severitatea greșelii este mai mică. Pentru a avea o privire de ansamblu asupra greșelilor din text, se reprezintă grafic greșelile identificate sub forma unei hărți de culori, folosind același cod de culoare ca la textul subliniat. Astfel, fiecare element de text este reprezentat printr-un dreptunghi de culoare corespunzătoare.

Feedback-ul și recomandările rezultate de la modulele descrise anterior sunt integrate în cadrul sistemului într-o serie de pași succesivi: restaurare de diacritice (pentru limba română), analiză gramaticală, analiză morfo-sintactică și analiză semantică.

În continuare se vor exemplifica greșelile găsite la fiecare pas de analiză, precum și sugestiile de corecție pentru textul: *„Eu, ca si cadru didactic regret ca colegii mei nu sunt de acord cu această hotarare. Deși consiliul a hotărât că această hotărâre este necesară și indispensabilă, ei nu o vroiau implementată. Totusi ei zicea că directorul are putere mai mare ca cea a consiliului, și nu se merită să continuăm discuția.*". Se consideră că rezultatul fiecărei etape reprezintă datele de intrare pentru etapa următoare. De exemplu, în etapa de analiză morfo-sintactică se va găsi construcția „Totuși ei ziceau", în loc de „Totusi ei zicea", pentru că diacriticele și dezacordul dintre subiect și predicat au fost identificate și corectate în etapa de restaurare de diacritice, respectiv de analiză gramaticală.

Etapă de analiză	**Fragment incorect**	**Tipul greșelii**	**Rezultat**
Restaurare de diacritice	**ca si cadru**	Lipsă diacritice	ca și cadru
	hotarare	Lipsă diacritice	hotărâre
	Totusi ei zicea că	Lipsă diacritice	Totuși ei zicea că
Analiză gramaticală	Totuși **ei zicea** că	Dezacord subiect - predicat	Totuși ei ziceau că

Etapă de analiză	Fragment incorect	Tipul greșelii	Rezultat
Analiză morfo-sintactică	**ca și** cadru didactic	Folosirea construcției "și parazitar"	Reformulare folosind una din expresiile: „în calitate de" sau „drept"
	că colegii	Cacofonie	Reformulare. Expresia „că colegii" este o cacofonie.
	a **hotărât** că această **hotărâre**	Repetiție	Reformulare. Repetiția cuvintelor „hotărât", „hotărâre"
	necesară și **indispensabilă**	Repetiția sinonimelor	Reformulare. Repetiția cuvintelor „necesară" și „indispensabilă". Aceste cuvinte sunt sinonime
	nu o **vroiau** implementată	imperfect "vroiam"	Schimbat în: „nu o voiau implementată"
	Totuși ei ziceau că.	Adverb la început de propoziție	Schimbat în: „Totuși, ei ziceau că ..."
	putere mai mare **ca cea** a consiliului	Comparare	Schimbat în: „putere mai mare decât cea a consiliului"

Etapă de analiză	**Fragment incorect**	**Tipul greșelii**	**Rezultat**
	nu **se merită**	Reflexiv parazitar	Schimbat în: „nu merită". Verbul „a merita" se folosește doar la diateza activă

Pentru a exemplifica analiza semantică, adăugăm următorul paragraf textului inițial: *„Extenuată, am plecat de la serviciu. Mă gândeam că ajunsă acasă o să găsesc totul pregătit dar, am observat că nu se făcuse cumpăraturi, casa nu se curățase și nici mâncare nu se facuse. M-am supărat foarte tare. Ei își inchipuie că chiar dacă eu sunt plecată toata ziua la servici, trebuie când mă întorc să fac cumpărături, să fac mâncare, și să fac curățenie? Oare eu nu merit să mă bucur de timpul liber, să mă mândresc de munca mea, să fiu respectată?"*

Pe baza acestui text, sistemul calculează indicii de complexitate textuală și oferă feedback personalizat, la patru nivele de granularitate: cuvânt, propoziție, paragraf și întregul document. Astfel, se identifică următoarele greșeli:

- Numărul de verbe (nivel de propoziție) - „În această propoziție sunt prea multe verbe. Este recomandată reducerea lor pentru simplificarea propoziției "
- Numărul de cuvinte (nivel de paragraf) - „În acest paragraf sunt prea puține cuvinte. Vă recomandăm o descriere mai amplă a anumitor concepte."
- Diateza pasivă (nivel de paragraf) - „Limbajul folosit în paragraf este colocvial. Se recomandă reducerea verbelor la diateza pasivă pentru a păstra acest stil."
- Diateza reflexivă (nivel de paragraf) - „Paragraful conține prea multe verbe la diateza reflexivă. Recomandăm diminuarea acestora."

- Numărul de paragrafe (nivel de document) – „Documentul ar trebui să fie structurat adecvat și să conțină cel puțin 3 paragrafe (introducere, cuprins și încheiere)."
- Diversitatea conceptelor (nivel de document) - „La nivel global aveți o diversitate prea mică a lexicului utilizat. Vă recomandăm introducerea de noi concepte pentru varietate și pentru a evita monotonia."
- Lungimea medie a tuturor cuvintelor (nivel de document) - „Textul pare mult prea complicat din prisma lungimii medii a cuvintelor. Recomandăm utilizarea unui limbaj simplificat pentru a facilita lectura textului."
- Coeziunea inter-paragraf (nivel de document) - „Textul are o coeziune globală scăzută între paragrafe. Vă recomandăm construirea unei structurii în cadrul căreia să maximizați coeziunea între paragrafe adiacente."
- Coeziunea intra-paragraf (nivel de document) - „Textul are o coeziune locală scăzută în paragraful selectat. Vă recomandăm construirea unei structurii în cadrul căreia să maximizați coeziunea între propozițiile constituente."

Revendicări

1. Metodă şi sistem de îmbunătăţire a stilului de scriere, cuprinzând o fază de antrenare şi o fază de test, **caracterizată prin aceea că**, în faza de *antrenare*, se folosesc două tipuri de corpusuri pentru crearea de modele specifice, respectiv colecţii de documente de referinţă pentru un anumit domeniu, în vederea stabilirii de valori admisibile pentru diverşi indecşi de complexitate textuală, respectiv o colecţie de fraze greşite cu sugestiile aferente de corecţie, peste care se aplică într-o primă etapă un algoritm bazat pe reţele neuronale în vederea augmentării corpusului, urmat în a doua etapă de antrenarea unui model bazat pe reţele neuronale axat pe corecţia automată la nivel morfo-sintactic a frazelor, şi se trece la faza de *test* în care, într-o primă etapă, se introduce un text pe dispozitivul utilizator; într-o a doua etapă se pre-procesează textul folosind tehnici de procesare a limbajului natural; într-o a treia etapă se calculează indecşi de complexitate textuală care pot fi utilizaţi în caracterizarea stilului de scris şi care pot fi grupaţi prin intermediul unei descompuneri PCA (Principal Component Analysis) în componente care caracterizează diverse dimensiuni de scriere; într-o a patra etapă se aplică reguli ajustate fiecărui domeniu din cadrul unui sistem bazat pe reguli care verifică dacă valorile aferente indicilor / componentelor sunt admisibile, raportat la colecţiile de documente de referinţă utilizate în etapa de antrenare; într-o a cincea a etapă se aplică modelul de corecţie antrenat pe colecţia augmentată de fraze greşite; într-o a şasea etapă se generează feedback comprehensiv pe multiple niveluri de granularitate (cuvânt, propoziţie/frază, paragraf, întregul document) luând în considerare recomandări rezultate în etapele patru şi cinci.

2. Sistem pentru îmbunătăţirea stilului de scris **caracterizat prin aceea că** integrează metoda conform revendicării 1 care preia textul de la utilizator, fie sub formă

cursivă convertită automat în format digital, fie text direct în format digital completat în cadrul unui formular.

3. Metodă conform revendicării 1, **caracterizată prin aceea că** oferă sugestii de corecție și de îmbunătățire a stilului de scris considerând într-o primă etapă elemente de suprafață (spre exemplu, abundență sau utilizare incorectă de semne de punctuație, fraze excesiv de scurte sau lungi ca număr de cuvinte), într-o a doua etapă erori gramaticale, într-o a treia etapă greșeli sintactice (spre exemplu, densitate prea mare de anumite părți de vorbire), într-o a patra etapă probleme semantice (spre exemplu, propoziții cu coeziunea locală mică raportat la contextul semantic, repetiții), într-o a cincea etapă elemente de discurs (spre exemplu, conectori de discurs utilizați excesiv, probleme de coeziunea globală care reflectă o incoerență a ideilor), și într-o a șasea etapă corecții specifice unei anumite limbi (spre exemplu, pentru limba română, lipsa diacriticelor, disonanțe, probleme de acord între diverse părți de vorbire, utilizarea excesivă a diatezei pasive sau a modului gerunziu).

4. Metodă conform revendicării 1, **caracterizată prin aceea că** efectuează o secvență de pre-procesare a textului care într-o primă etapă realizează prelucrări preliminare, respectiv: tokenizare, eliminare cuvinte de tip stop-words, adnotare cu părți de vorbire, lematizare și analiză sintactică bazată pe dependențe, într-o a două etapă efectuează prelucrări independente, și anume: recunoaștere entități cu nume, dezambiguizare cuvinte și identificare sensuri și coreferințe, și într-o a treia etapă vizează corecții suplimentare care adresează marcarea repetițiilor (inclusiv sinonime), identificarea disonanțelor și verificarea acordului.

5. Metodă conform revendicării 1, **caracterizată prin aceea că** feedbackul consideră atât regulile specifice domeniului, implementate la nivelul sistemului bazat pe reguli, cât și corecțiile sugerate de rețeaua neuronală, iar rezultatul analizei

automate include atât sugestii potențiale de corecție, cât și o reprezentare vizuală sub forma unei hărți de culori în cadrul căreia fiecare element de text (spre exemplu, cuvânt, propoziție/frază, paragraf, document) este identificat printr-un dreptunghi colorat conform nivelului de severitate al greșelilor identificate.

6. Metodă conform revendicării 1, **caracterizată prin aceea că** rețelele neuronale utilizate, atât de augmentare a corpusului de greșeli, cât și de corecție, sunt de tipul codificator-decodificator, iar modelul generativ rezultat asigură o libertate extinsă în schimbări, întrucât corecția frazelor poate include sugestii de corectare complexe, implicând schimbarea, adăugarea sau ștergerea unui număr extins de cuvinte.

7. Metodă conform revendicării 1, **caracterizată prin aceea că** rețelele neuronale utilizate includ mecanisme de atenție pentru construirea la fiecare pas a vectorului contextual.

8. Metodă conform revendicării 1, **caracterizată prin aceea că** modifică algoritmul de beam-search din cadrul decodificatorului utilizat în cadrul etapei de augmentare a corpusului cu fraze greșite, prin adăugarea la fiecare pas a unei valori k * **β** a distribuțiilor de probabilitate ale frazelor candidat, pentru a spori zgomotul (diversitatea) frazelor generate de către rețea.

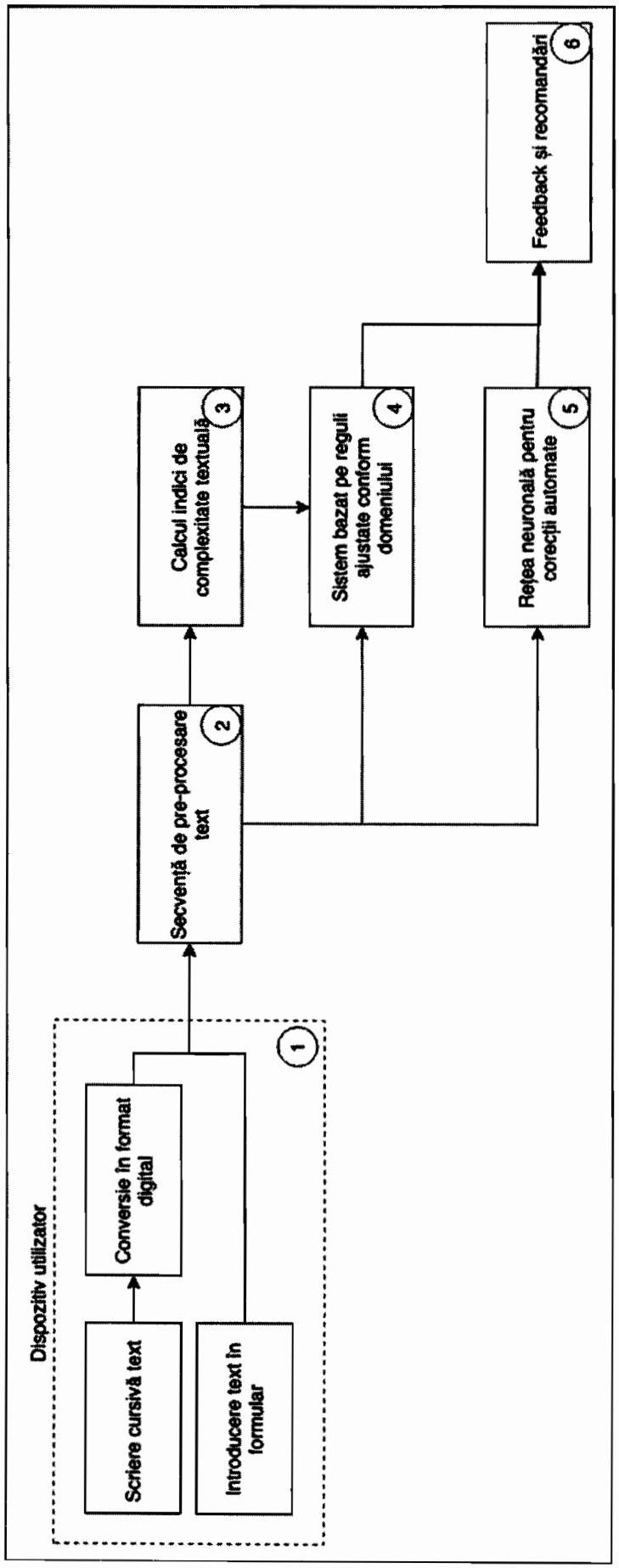

Fig. 1 – Schema bloc a metodei și sistemului de îmbunătățire a stilului de scriere

30

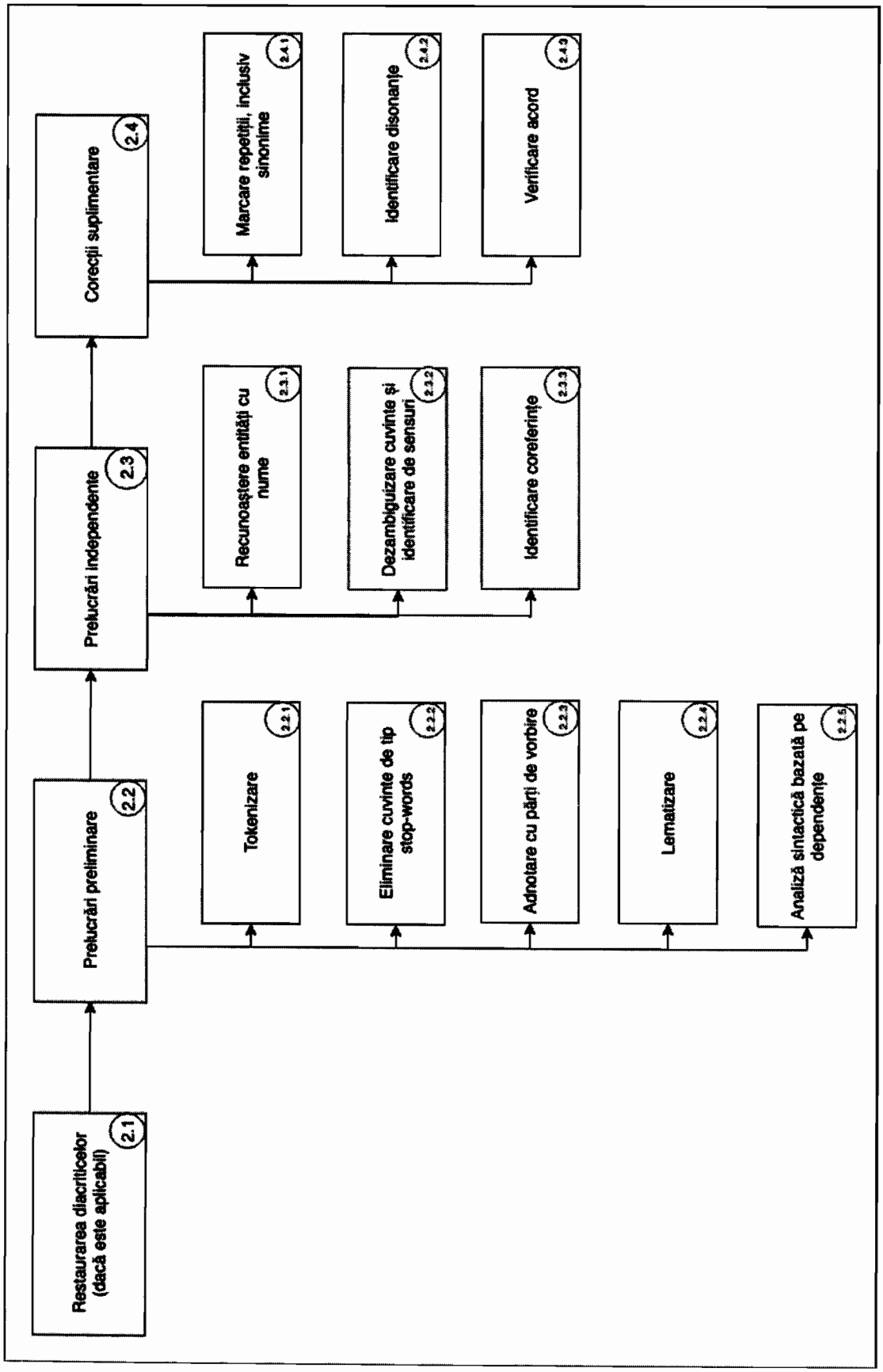

Fig. 2 – Secvența de pre-procesare a textului

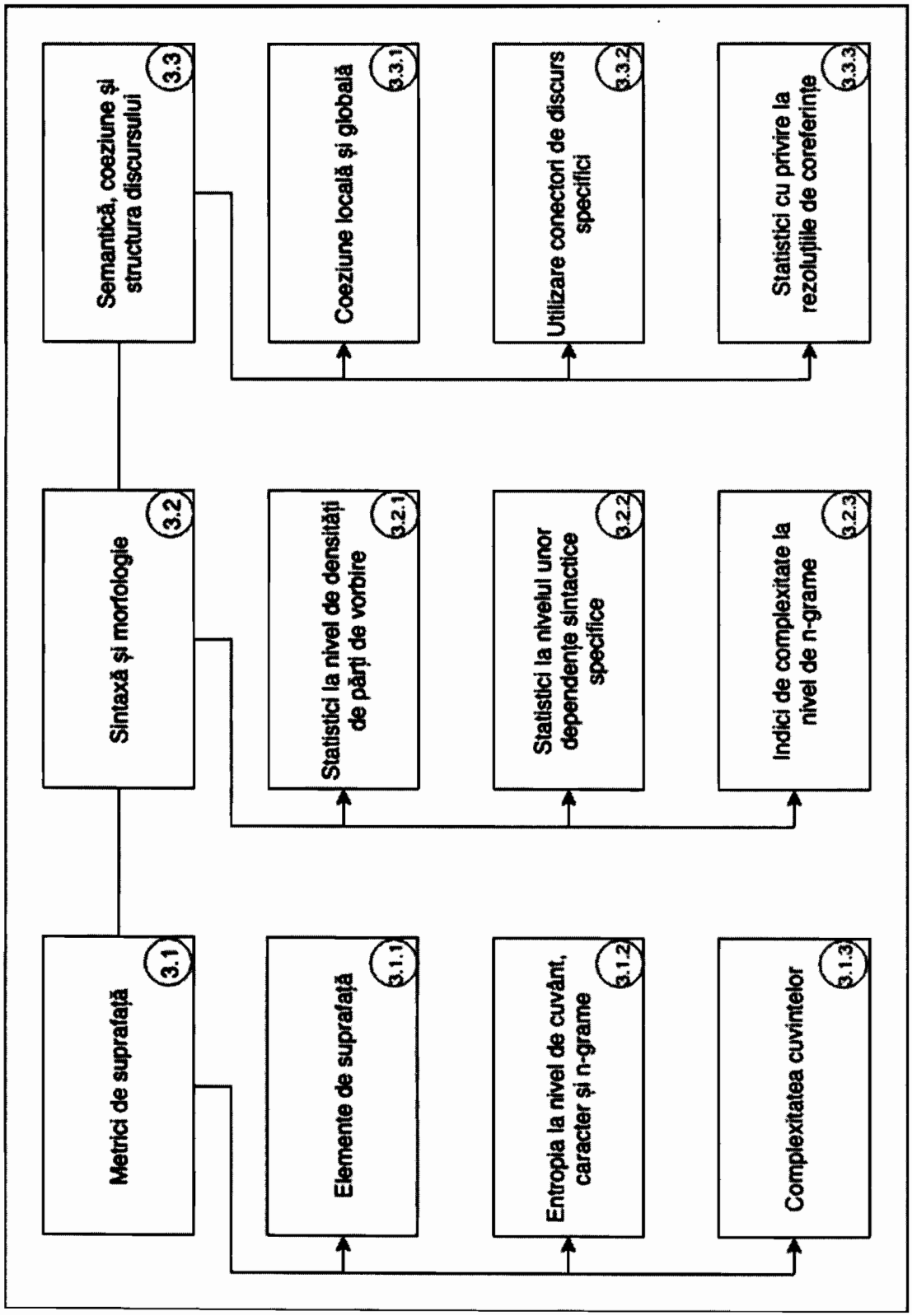

Fig. 3 – Exemple de indici de complexitate textuală utilizați în generarea feedback-ului

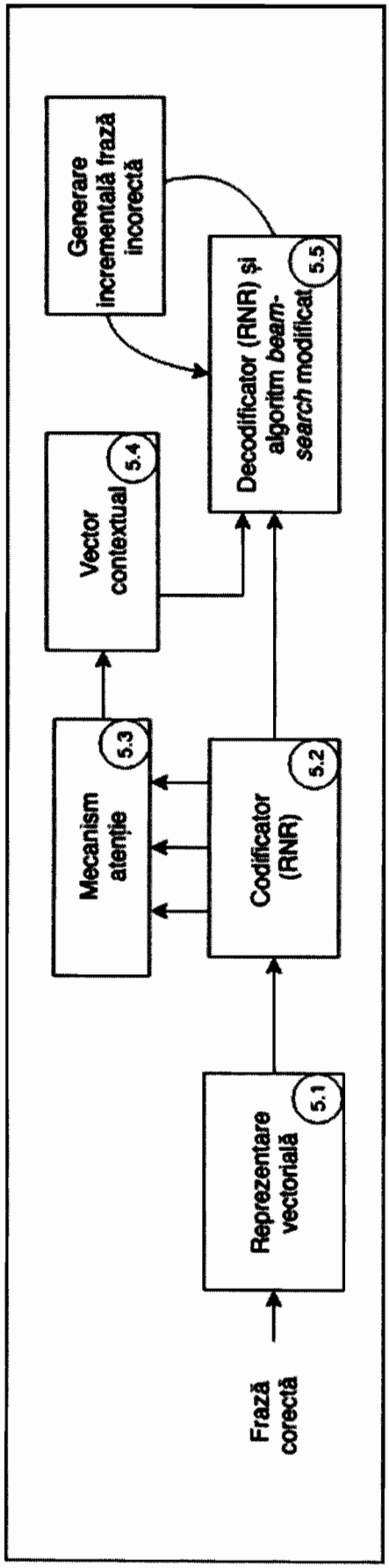

Fig. 4 – Rețea neuronală pentru augmentarea corpusului inițial cu greșeli gramaticale

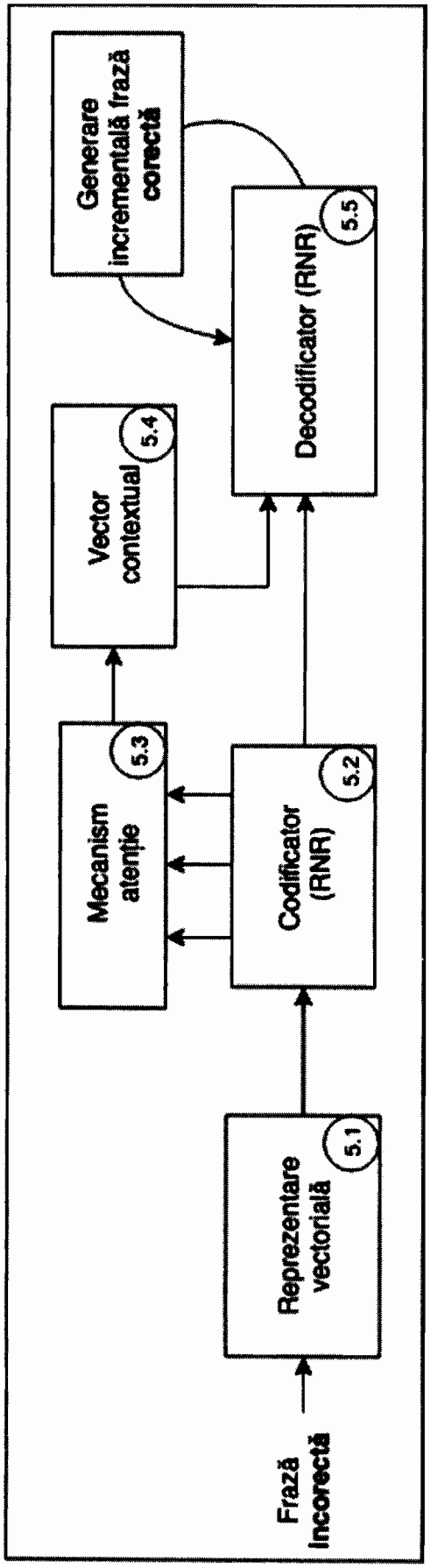

Fig. 5 – Rețea neuronală pentru corectarea frazelor

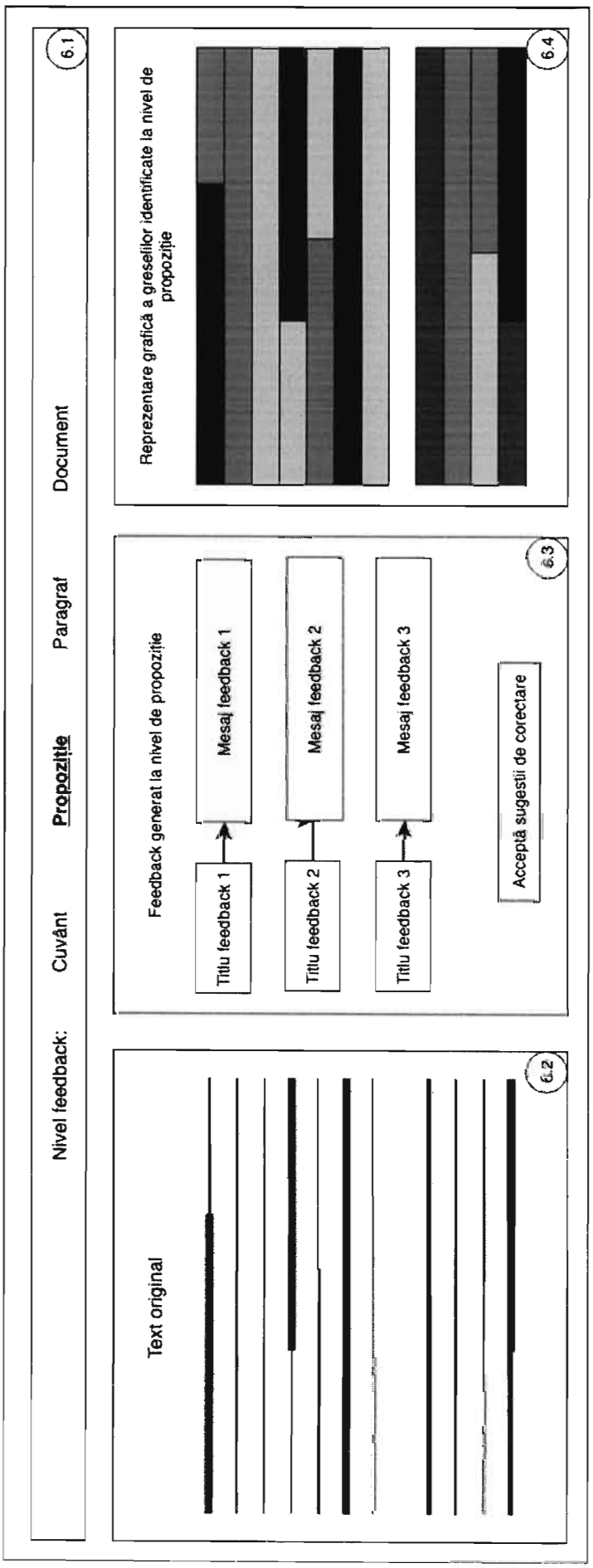

Fig. 6 – Reprezentare vizuală a feedback-ului primit de utilizator la nivel de propoziție