**ROMÂNIA**

**MINISTERUL APĂRĂRII NAŢIONALE**

**ACADEMIA TEHNICĂ MILITARĂ „FERDINAND I”**

**FACULTATEA DE SISTEME INFORMATICE ȘI SECURITATE CIBERNETICĂ**

**Specializarea: Calculatoare și sisteme informatice pentru apărare și securitate națională**

**RAPORT PRACTICĂ:**

**DEZVOLTAREA UNEI PLATFORME EDUCATIVE CU AJUTORUL ÎNTREBĂRILOR GENERATE PROCEDURAL**

**Sd. Sg. Porfirie-Denissa PILIPĂUȚANU**

**Gr. C-113C**

**BUCUREŞTI**

**2020**

Cuprins

[**1. Introducere 1**](#_Toc76118567)

[**2. Generarea procedurală 1**](#_Toc76118568)

[**3. Generarea automată a întrebărilor 2**](#_Toc76118569)

[**3.1. Metode generative 2**](#_Toc76118570)

[**3.2. Procesul generator 3**](#_Toc76118571)

[**4. Evaluarea calității întrebărilor 6**](#_Toc76118572)

[**5. Concluzii 6**](#_Toc76118573)

[**Bibliografie 7**](#_Toc76118574)

[**Generarea de text cu ajutorul gramaticilor 8**](#_Toc76118575)

[**Definirea unei gramatici 8**](#_Toc76118576)

[**Utilizarea librăriilor pentru generarea de text 10**](#_Toc76118577)

# Introducere

În procesul educațional, testarea acumulării de cunoștințe prin întrebări este una dintre metodele fundamentale de evaluare. Acestea oferă beneficii precum stimularea memoriei pentru regăsirea și sedimentarea informației, focusarea pe conceptele importante din materialul studiat și în anumite circumstanțe obținerea de feedback atunci când există o nelămurire.

Provocarea întâlnită în momentul utilizării acestui instrument este validarea actului de învățare care scade proporțional cu numărul de folosiri, creându-se o familiaritate cu itemii respectivi.

Compunerea manuală a acestora poate deveni un proces complex ce necesită experiență, resurse și timp. Astfel, nevoia de a avea un set amplu de întrebări inhibă folosirea resurselor educaționale pentru antrenament.

# Generarea procedurală

Generarea procedurală este una din ramurile creării sintetice de conținut multimedia, domeniu cunoscut pentru producerea, manipularea și modificarea datelor prin mijloace automate, în special, prin utilizarea algoritmilor de inteligență artificială[[1]](#footnote-1). Aceasta se bazează pe fuziunea dintre conținutul generat de om și aleatorismul utilizat în cadrul algoritmilor.

În funcție de implementare, printre avantajele generării procedurale se numără dimensiunea redusă a fișierelor, creșterea cantității și a varietății de conținut[[2]](#footnote-2).

Aplicațiile acesteia se regăsesc cu precădere în industrii precum cea a jocurilor, cinematografică sau textilă[[3]](#footnote-3).

Pentru a vedea cum este utilizată această metodă în scopuri didactice, am efectuat căutări pe baza cuvintelor cheie “educational”, “procedural”, ”generation”, însă rezultatele direcționau tot către industria jocurilor. Printre acestea am observat o potențială alternativă, care ar fi însemnat schimbarea cuvântului cheie ”procedural” cu ”automatic”. De această dată, rezultatele se apropiau de tema lucrării de față, fiind legate de generarea chestionarelor cu ajutorul unor surse de informații structurate. Pentru a filtra mai bine aceste rezultate, am adăugat și cuvântul cheie ”question”.

În cele din urmă, am ajuns la rezultatul dorit, acesta fiind articolul ”A Systematic Review of Automatic Question Generation for Educational Purposes” (Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler & Salam Al-Emari) care include o analiză comparativă a 93 de lucrări științifice din domeniu. Imaginea de ansamblu asupra comunității AQG, direcțiile curente și progresele realizate furnizate de acest articol se dovedesc a fi un bun punct de pornire în a decide cum va fi abordată tema de față.

# Generarea automată a întrebărilor

Pentru a reduce costurile asociate procesului manual de generare al întrebărilor și pentru a rezolva problema unui flux constant de conținut, au fost dezvoltate tehnici de generare automată ale acestora[[4]](#footnote-4).

Domeniul AQG a cunoscut o creștere a numărului de publicații începând cu anul 2011, după desfășurarea unor workshop-uri în anii precedenți care au atras atenția cercetătorilor. Acest eveniment coincide cu lansarea platformelor de cursuri online precum Udemy, Coursera, Udacity sau edX, care cresc interesul pentru AQG. De asemenea, este important de menționat că atât procesarea limbajului natural (NLP) cât și web-ul semantic, arii strâns legate de generarea de întrebări, au devenit mult mai mature și au furnizat metode și unelte care au influențat cantitatea și calitatea cercetării.

## Metode generative

Metodele de generare ale întrebărilor pot fi clasificate din mai multe puncte de vedere:

* în funcție de nivelul de înțelegere al intrării necesar pentru crearea acestora:
  + bazate pe sintaxă (folosesc arborele sintactic al textului de intrare conform gramaticii independente de context[[5]](#footnote-5) și nu necesită înțelegerea semanticii intării, ex: selectarea distractorilor pe baza părții de vorbire);
  + bazate pe semantică (folosesc relații semantice de echivalență, ierarhice sau asociative[[6]](#footnote-6) și operează la un nivel mai profund, utilizând surse adiționale de informații precum taxonomii sau ontologii, acestea neregăsindu-se în mod explicit în intrare, ex: selectarea distractorilor pe baza similarității contextului sau caracteristicilor);
* în funcție de procedura utilizată pentru transformarea intrării într-o întrebare:
  + bazate pe șabloane (folosesc o structură de bază a întrebării, customizând golurile cu valori ce întrunesc caracteristicile sintactice sau semantice precizate);
  + bazate pe reguli (se anotează propozițiile textului de intrare cu informații sintactice sau semantice, iar apoi aceste anotații sunt folosite pentru a potrivi intrarea cu un tipar specificat în reguli; acestea descriu modul în care este selectat un anumit tip de întrebare și cum e manipulată intrarea pentru a o construi);
  + bazate pe statistică (abordează procesul generator ca pe o problemă de predicție a conținutului întrebării, pornind de la un segment de text dat; secvența următoare din cadrul acesteia este determinată cu ajutorul unui set de antrenament care furnizează probabilitățile de apariție a unor secvențe din textul de intrare în acel context);

Acestea nu sunt mutual exclusive.

Cele trei metode de generare a întrebărilor au avantaje și dezavantaje, însă ceea ce au toate în comun este faptul că sunt considerate scumpe din punct de vedere al costului. Șabloanele și regulile necesită intervenția umană prin construirea manuală a acestora, iar învățarea prin antrenament necesită cantități mari de date etichetate corespunzător, inexistente în multe domenii. Întrebările generate prin reguli și metode statistice sunt foarte similare cu textul de intrare, pe când șabloanele permit generarea unor întrebări cu o structură diferită de cea a intrării, însă limitată din punct de vedere al diveristății lingvistice.

## Procesul generator

Activitățile desfășurate în cadrul procesului generator pot fi grupate în trei categorii:

* preprocesarea;
* construcția întrebării;
* postprocesarea.

**Preprocesarea**

În funcție de metoda de generare se utilizează una sau două tipuri de preprocesare:

* preprocesarea standard;
* preprocesarea specifică generării întrebării.

Preprocesarea standard este utilizată în cadrul sarcinilor de procesare a limbajului natural (NLP), pregătind intrarea pentru sarcinile următoare. Aceasta presupune segmentarea intrării, împărțirea în propoziții, tokenizarea (împărțirea propoziției în cuvinte), etichetarea cuvintelor ca părți de vorbire (POS tagging), clusterizarea tuturor expresiilor aceleiași entități (identificarea relațiilor de coreferință). În unele cazuri, mai poate include extragerea de informații prin recunoașterea entităților numite (nume de persoane, organizații, locații, etc)[[7]](#footnote-7) sau extragerea de relații semantice dintre două sau mai multe entități.

Scopul preprocesării specifice generării de întrebări este de a crea sau de a selecta intrările potrivite pentru procesul generator. Acesta se regăsește în cele trei tipuri de preprocesare identificate:

* simplificarea frazelor (frazele complexe, ce conțin apoziții sau propoziții alăturate prin conjucții sunt convertite în propoziții simple pentru facilitarea taskurilor următoare, precum extragerea tripletului subiect-predicat-complement utilizat în generarea întrebării);
* clasificarea propozițiilor (propozițiile sunt repartizate unor categorii, pentru a determina ce fel de întrebări pot fi alcătuite pe baza acestora; aceasta se realizează cu ajutorul părților de vorbire etichetate și a relațiilor de dependență dintre acestea);
* selecția de conținut (cum numărul de întrebări în cadrul unei examinări e limitat, scopul acestui proces este de a determina conținutul important, precum propozițiile, părțile din propoziții sau conceptele din care pot fi generate întrebări, iar din varietatea produsă, examinatorii decid care dintre acestea vor fi alese; cu toate acestea, în mediile de învățare nesupervizată, acest lucru nu este fezabil, din cauza lipsei factorului decizional).

Selecția de conținut e specifică intrărilor bazate pe text, spre deosebire de cele bazate pe informații structurate.

**Construcția întrebării**

Principala sarcină a întregului proces de generare este alcătuirea întrebării (setul întrebare-răspuns corect-distractori). În funcție de tipul acesteia și de formatul răspunsului sunt implicate diferite abordări:

* pentru generarea de întrebări:
  + se transformă propoziția asertivă în una interogativă, atunci când intrarea este un text;
  + se determină tipul întrebării, selectând pronumele interogativ (care, cine, ce, cât, câți, câte) sau șablonul potrivit;
  + este selectată poziția locului liber în cazul propozițiilor stem (din care lipsesc anumite părți).
* pentru generarea opțiunilor incorecte (un task foarte important în generarea întrebărilor cu răspuns multiplu (MCQ) deoarece distractorii influențează calitatea întrebării):
  + se efectuează o selecție pe baza frecvenței cuvintelor alăturate unei anumite secvențe care a contribuit la generarea întrebării;
  + se aleg distractorii în funcție de similaritatea față de cheie din punct de vedere al sintaxei (parte de vorbire similară, litere similare), al caracteristicilor sau al contextului, în special când sunt utilizate baze de informații structurate (KB – knowledge bases);
  + se efectuează o interogare permisivă pentru a genera cheile întrebărilor, ce oferă ca rezultat și distractori care împărtășesc anumite caracteristici cu cheile respective;
* pentru generarea de feedback (o explicare a corectitudinii sau incorectitudinii unui răspuns la o anumită întrebare în funcție de alegerea celui evaluat):
  + fiind un aspect neglijat în literatura AQG, doar un studiu a tratat acest aspect prin generarea unei verbalizări a axiomei folosite pentru a selecta potențialele răspunsuri ale întrebării.
* pentru controlul dificultății întrebării (determinarea a cât de ușoră sau grea va fi o întrebare):
  + se determină numărul de persoane evaluate care au răspuns corect din totalul celor examinați;
  + se compară dificultatea prezisă cu cea precizată de un expert;
  + se compară dificultatea prezisă cu performanța obținută de subiecții evaluați;
  + se utilizează mecanisme de rezolvare automată.

**Postprocesarea**

Scopul postprocesării este de a îmbunătăți întrebările generate. Acest lucru se realizează de obicei prin două metode:

* verbalizare (procesul desfășurat pentru a retușa întrebările din punct de vedere gramatic și al fluenței sau pentru a furniza variații ale întrebărilor prin parafrazare);
* ierarhizarea întrebărilor produse prin modele statistice, pentru a facilita alegerea celor de o calitate mai bună.

Parafrazarea are rolul de a diversifica, de a-i provoca și de a se asigura că subiecții participanți în evaluare și-au însușit cu adevărat cunoștințele necesare.

De asemenea, conceptul de verbalizare este întâlnit și în lucrul cu baze de informații structurate care sunt reprezentate folosind diferite convenții precum litera mare la începutul fiecărui cuvânt (camel case, ex: exempluCamelCase) sau caracterul underscore (’’\_’’) între acestea, fiind nevoie de o reprezentare naturală. Prin urmare, se efectuează o procesare de bază ce include segmentarea în cuvinte și adaptarea la o formă morfologică adecvată.

# Evaluarea calității întrebărilor

În urma generării întrebărilor, acestea pot fi evaluate cu ajutorul unei referințe, în mod automat, de către experți sau de către studenți.

Seturile de date standard sunt una din practicile utilizate de anumiți cercetători. Un exemplu este SquAd (The Stanford Question Answering Dataset) ce conține un set larg de întrebări din articole de pe Wikipedia, cărora le sunt alăturate paragrafele de unde au fost extrase cu răspunsul corespunzător. Un astfel de set este folosit pentru a compara întrebările generate prin diferite metode cu cele generate manual.

O metodă de comparație utilizată într-o altă abordare este furnizarea întrebării originale alături de întrebarea generată și ierarhizarea calității acestora de către factorul uman.

Criteriile de evaluare se orientează asupra calității lingvistice și educaționale. Din punct de vedere lingvistic se urmărește corectitudinea gramaticală, fluența, abiguitatea semantică, lipsa erorilor și rezonabilitatea distractorilor. Din punct de vedere educațional se urmărește utilitatea întrebării, relevanța în cadrul domeniului și obiectivul didactic atins de subiectul evaluat prin răspunsul la acea întrebare.

De asemenea, mai sunt și alte metrici relevante pentru calitate precum dificultatea sau nivelul cognitiv țintă.

# Concluzii

Articolul studiat a analizat este 90 de lucrări legate de generarea întrebărilor în domeniul educațional, însă nu s-a putut identifica o performanță standard sau cele mai bune mijloace de generare din cauza eterogenității măsurătorilor ce privesc calitatea și a modului de raportare a rezultatelor (lipsa numărului de întrebări evaluate, a distribuției întrebărilor pe tipuri, nivele de dificultate, a numărului de participanți implicați în evaluări, a informațiilor despre strategia de eșantionare sau a mărimii eșantioanelor, a descrierii caracteristicilor participanților, etc). Astfel, se remarcă nevoia dezvoltării unui standard care să definească procedurile de evaluare și de raportare a rezultatelor.

Una din principalele limitări a lucrărilor analizate este simplitatea întrebărilor generate (bazate pe tripletul subiect-predicat-complement), majoritatea fiind constituite din puțini termeni și țintind nivele cognitive mai joase.

Cu toate acestea, am găsit acest studiu ca fiind foarte util prin prezentarea metodelor generative și înglobarea etapelor procesului generator pentru diferite tipuri de întrebări și răspunsuri. Un alt aspect foarte important este și referențierea lucrărilor din diferite domenii ale spectrului educațional, facilitând astfel, filtrarea celor care nu se potrivesc temei abordate.

De asemenea, țin să menționez că în documentarea acestei lucrări, un impact pozitiv l-a avut Kate Compton prin discursurile despre generarea procedurală din cadrul Game Developer Conference în care a prezentat și librăria javascript ”Tracery” de generare a textului cu ajutorul unei gramatici. Căutând detalii despre această librărie am descoperit cursul lui Daniel Shiffman care se focusează pe analiza și generarea procedurală a textului prin gramatici independente de context și metode statistice precum lanțurile Markov, pe care îl consider drept pas următor în documentarea temei.

# Bibliografie

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, Salam Al-Emari, *A Systematic Review of Automatic Question Generation for Educational Purposes*, International Journal of Artificial Intelligence in Education 30, 2020, p. 121-204

# Generarea de text cu ajutorul gramaticilor

## Definirea unei gramatici

Limbajul este un sistem de comunicare bazat pe cuvinte și pe combinarea cuvintelor pentru a forma propoziții și fraze. Acesta este descris de fonologie (regulile după care simbolurile sunt utilizate pentru a forma cuvinte – microsintaxă), sintaxă (regulile după care cuvintele se combină pentru a forma propoziții și fraze) și semiotică (procesul prin care semnele sunt legate de anumite înțelesuri – semantică).

Reprezentarea cuvintelor se realizează prin folosirea unor simboluri convenționale, care în cazul limbajelor de programare sunt șirurile de caractere. Pentru a defini un șir de caractere se utilizează o mulțime finită de simboluri, numită alfabet (ex: ∑ = {a, b, ..., z, #} – mulțimea literelor din alfabetul latin, plus simbolul special #).

Mulțimea de șiruri (finită sau infinită) de simboluri peste același alfabet (ex: ab#bc – șir de caractere peste ∑) constituie limbajul. Dacă acesta este finit, atunci el poate fi definit prin enumerare. Dacă este infinit, atunci există două mecanisme de definire:

* prin generare, cu ajutorul gramaticilor formale (știu să genereze toate propozițiile din limbaj (și numai pe acestea), astfel încât alegând o propoziție din limbaj, aceasta va fi generată într-un interval finit de timp);
* prin recunoaștere, cu ajutorul automatelor finite și expresiilor regulare (știu să recunoască (să accepte ca fiind corecte) propozițiile limbajului dat (și numai pe acestea)).

Limbajele naturale nu se bazează pe reguli stricte, ceea ce cauzează ambiguitatea semantică. Astfel, același cuvânt sau aceeași propoziție/construcție/frază poate avea mai multe înțelesuri, putând fi interpretat în mai multe moduri. Pentru oameni acest lucru nu constituie o problemă deoarece sensul este dedus imediat cu ajutorul contextului, însă mașinile de calcul nu au această capabilitate. Prin urmare, limbajele formale (cu reguli precise) constituie o soluție a acestei probleme.

O gramatică formală este un cvadruplu G = (VN, ∑, S, P), unde:

* VN se numește mulțimea neterminalelor;
* ∑ se numește mulțimea terminalelor;
* S este simbolul inițial (sau simbolul de start), S ∈ VN;
* P este mulțimea regulilor de producție.

V = VN ∪ ∑ reprezintă alfabetul gramaticii.

Observație: Acest cvadruplu definește numai lexicul (vocabularul) limbajului și sintaxa regulilor de producție, nu și semantica acestora.

Forma propozițională peste o gramatică G = (VN, ∑, S, P) este orice succesiune de simboluri terminale și/sau neterminale care poate fi obținută pornind de la simbolul de start prin 0 sau mai multe derivări. Atunci când este derivată o formă propozițională, se poate ajunge în situația în care aceasta să conțină mai multe simboluri neterminale. Prin urmare va trebui aleasă următoarea regulă de producție pentru unul din simbolurile neterminale, cu scopul de a face o derivare directă. Dacă se alege întotdeauna neterminalul cel mai din stânga, atunci derivarea se va numi derivare stânga. Dacă se alege întotdeauna neterminalul cel mai din dreapta, atunci derivarea se va numi derivare dreapta.

Exemplu:

Fie gramatica G = (VN, ∑, S, P) astfel încât:

* VN = {Subiect, Predicat, Propoziție};
* ∑ = {Ana, Ion, învață, doarme, ’ ’, .};
* S = Propoziție;
* P = {

Propoziție → Subiect ’ ’ Predicat .;

Subiect → Ana | Ion;

Predicat → învață | doarme;

}.

Propoziție → Subiect ’ ’ Predicat .

Derivare stânga:

Subiect ’ ’ Predicat . → Ana ’ ’ Predicat .

Ana ’ ’ Predicat . → Ana ’ ’ învață .

Derivare dreapta:

Subiect ’ ’ Predicat . → Subiect ’ ’ doarme .

Subiect ’ ’ doarme . → Ion ’ ’ doarme .

O formă propozițională peste o gramatică G, care conține numai simboluri terminale, se numește propoziție generată de G. Propoziția este deci orice succesiune de simboluri terminale care poate fi obținută pornind de la simbolul de start prin cel puțin o derivare.

Limbajul generat de gramatica G = (VN, ∑, S, P) se notează cu L(G) și reprezintă mulțimea tuturor propozițiilor care pot fi generate de către această gramatică (ex: Ana învață., Ion învață., Ana doarme., Ion doarme.). Prin urmare, gramatica este reprezentarea finită (toate elementele sale sunt finite) a unui limbaj care poate să fie infinit. Pentru a genera propoziții cu un număr oarecare de simboluri este utilizată recursivitatea (ex: A →\* Aα (stânga), A →\* αA (dreapta), unde A este un simbol neterminal, iar α este un șir de simboluri terminale și neterminale).[[8]](#footnote-8)

## Utilizarea librăriilor pentru generarea de text

**Tracery**

Librăria Tracery permite definirea gramaticii ce va genera limbajul dorit prin intermediul unui obiect javascript.

var poveste = {

"start": ['A fost odata ca niciodata un #adj## #erou#'.],

"erou": ['unicorn', 'dragon', 'urs'],

"adj": ['vrednic', 'dragut', 'suparat']

}

Elementele încapsulate de către simbolul # sunt marcate ca fiind neterminale, pentru a fi înlocuite în mod aleator cu un element din mulțime conform regulilor de producție aferente acestora. Cu cât regulile de producție sunt mai imbricate, cu atât rezultatul este mai complex.

Pentru a genera povestea se creează un obiect aferent gramaticii din datele furnizate.

var gramatica = tracery.createGrammar(poveste);

Pornind de la simbolul de start, orice rezultat poate fi obținut prin expandarea arborelui sintactic. Pentru a obține doar rezultatul final se apelează metoda flatten.

var rezultat = gramatica.flatten('#start#');

**RiTa.js**

Librăria RiTa.js definește o gramatică prin crearea obiectului RiGrammar și adăugarea de reguli acestuia. De asemenea, gramatica poate fi încărcată dintr-un fișier formatat JSON sau YAML, folosind metoda loadFrom aferentă obiectului.

var gramatica = new RiGrammar();

gramatica.addRule('<start>', 'A fost odata ca niciodata un <adj> <erou>.');

gramatica.addRule('<erou>', 'unicorn | dragon | urs');

gramatica.addRule('<adj>', 'vrednic | dragut | suparat');

RiTa identifică regula corespunzătoare simbolului inițial <start> pentru a genera un text prin apelul metodei expand.

var story = gramatica.expand();[[9]](#footnote-9)

1. https://en.wikipedia.org/wiki/Synthetic\_media, accesat la 18 iunie 2021 [↑](#footnote-ref-1)
2. https://en.wikipedia.org/wiki/Procedural\_generation, accesat la 18 iunie 2021 [↑](#footnote-ref-2)
3. https://youtu.be/WumyfLEa6bU, accesat la 20 iunie 2021 [↑](#footnote-ref-3)
4. https://link.springer.com/article/10.1007/s40593-019-00186-y, accesat la 22 iunie 2021 [↑](#footnote-ref-4)
5. https://en.wikipedia.org/wiki/Parse\_tree, accesat la 22 iunie 2021 [↑](#footnote-ref-5)
6. https://marciazeng.slis.kent.edu/Z3919/4relationship.htm, accesat la 22 iunie 2021 [↑](#footnote-ref-6)
7. https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da, accesat la 23 iunie 2021 [↑](#footnote-ref-7)
8. Mihai-Lica Pura, Curs Limbaje formale și translatoare, p. 1-55 [↑](#footnote-ref-8)
9. https://shiffman.net/a2z/cfg/ [↑](#footnote-ref-9)