VI. Xuxen: bi mailatako morfologian oinarritutako zuzentzaile ortografikoa.

Euskararako zuzentzaile ortografiko bat burutzea zen gure helburua hasiera-hasieratik; euskara bizi den batasun-prozesurako oso garrantzi handikoa baita halako tresna bat.

Hasieratik argi zegoen zuzentzailea analisi morfologikoan oinarritu behar zela; ondorioz, eraikitzen ari ginen analizatzaile morfologikoa berrerabiltzea zen irtenbiderik, logikoena ez ezik, merkeena eta interesgarriena.

Euskararen batasuna erabat finkatu gabe egoteak problematika berria eta aberatsa dakar, ikergaia interesgarriago bihurtuz. Gainera, bibliografian agertzen ziren erreferentzia gehienak ez ziren oso baliagarriak izaten, euskara bezalako hizkuntza eranskarietan zuzenketa-prozesua korapilatsuagoa da eta.

Arazo hauen aurrean, eta egiaztatzea analisi morfologikoan oinarrituz burutzeaz gain, problema ebazteko sinpleago diren azpiproblematan banatzea izan zen erabili den taktika: errore tipografikoak alde batetik eta ez-jakiteak eragindako erroreak edo gaitasun-erroreak —analisi morfologikoaren aldaeren tratamendu bera erabiliko denez aldaerak ere deituko ditugunak— bestetik. Tratamenduaren diseinua egiterakoan eraginkortasuna eta zehaztasunaren arteko oreka abiapuntua izan da, erabilpen komertziala bilatzen ari ginen eta.

Analisi tipografikoen tratamendurako, aurreko kapituluan aztertutako alderantzizko edizio-distantziaren ohizko metodoa erabili da, flexio konplexuko hizkuntzetarako baliagarria dena. Metodo honi jarraituz, akasdun formaren gainean proposamen hipotetikoak eratu eta egiaztatzen dira, ondoren benetako hitzak direnak sailkatuz. Proposamen hipotetiko guztien analisia ekiditeko, egiaztatzea analisi morfologikoaren bidez egiten baita, hipotesiak murrizteko eta sailkatzeko zenbait metodo erabiltzen dira.

Gaitasun-erroreen zuzenketarako metodo berritzailea erabili dugu. Hitz bat erroretzat hartzen da analisi morfologiko estandarrik ez dagokionean. Horren aurrean, eta laugarren kapituluan azaldutako aldaeren tratamendu morfologikoa berrerabiliz, lexikoa eta erregela morfofonologikoak hedatzen dira aldaeren tratamendurako informazioarekin, eta analisi-saio berri bat burutzen da. Saio berri horretan analisirik lortzen bada, erroredun hitza gaitasun-erroretzat har daiteke eta baita sorkuntza morfologikoari esker dagokion forma estandarra sortu ere.

Bi tratamenduetan proposatzeko benetako hitzak sor daitezke. Proposamen hauek nola ordenatu behar diren erabakitzea ez da erraza, bibliografian dauden proposamenak oso heterogenoak izanik. Elkarrekintzazko testu-edizioan erabiliko den honetan funtsezkoena ez bada ere, estatistikan oinarritutako sailkapen bat erabaki da.

Proposamenen sorkuntza eta sailkapena egiten duten moduluez gain beste bi modulu garrantzitsu azpimarratu behar dira: iragazlea eta erabiltzailearen hiztegia eguneratzekoa. Biak morfologian erabilitako token-ezagutzailea eta erabiltzailearen lexikoak egokituz burutu dira.

Azkenik modulu guztiak integratzeko eta ingurune atsegina lortzeko elkarrekintza ere diseinatu da.

Inplementaturiko zuzenketa-sistemak zehaztasuna/eraginkortasuna oreka mantentzen duelakoan gaude. Izan ere, aukera berriak ari gara aztertzen bi bide nagusitatik: batetik, lexikorik gabeko analisia berrerabiliz erro-atzizki banaketan oinarritutako metodo bat diseinatzea errore tipografikoetarako; eta bestetik, eraginkortasuna hobetzearren, lexiko-itzultzaileen erabilera testu-zuzenketan.

VI.1. Sarrera.

Euskararako zuzentzaile ortografiko baten diseinuaren aurrean, aurreko kapituluan zehaztutako kontzeptuak gogoan hartuz, hauek izan ziren programaren funtzionalitateak mugatzen eta zehazten dituzten oinarrizko irizpideak:

- Eskala errealeko zuzentzaile erabilgarria burutzea, produktu komertzial baten oinarria izanik, euskararako horrelakorik ez baitzegoen. Gainera, indarrean den batasun-prozesuan horrelako tresna bat oso lagungarria da.
- Lehen belaunaldiko zuzentzaile ortografikoa zen helburua, hau da testuingurua kontuan ez duen zuzentzaile arrunta, testu-edizioan erabili ohi direnak bezalakoa. Irekitako ikerlerro bezala gelditu dira testuingurua kontuan hartuz zuzentzaile hau hobetzea eta estilo-zuzentzailea edo zuzentzaile gramatikala egitea.
- Morfologian hitz anitzeko terminoen tratamendua baztertzeko erabilitako arrazoiek hitz-mugaren gaineko erroreak lan honen eremutik kanpo uzteko ere balio dute, tratamendu partzial bat egin bada ere aldaeren kasuan.

- Egiaztatzea analisi morfologikoan oinarrituz egitea, beste metodoak, n-grametan oinarritutako metodo estatistikoak zein forma-zerrendan oinarritutakoak, baztertuz.

- Euskararen batasunerako arauak ondo ez ezagutzeak zein erabilpen dialektalak eragindako erroreak zuzentzea lehenestea gainontzeko erroreen aurrean, haiek direlakoan erabiltzaileei zuzentzeko buruhauste handiena ematen dietenak eta batasun-prozesuan lagungarrien gertatzen direnak[1].

- Euskararen egoera kontuan hartuz erabiltzailearen hiztegiak sortu eta eguneratzeko aukera ezinbestekoa da, lexiko orokorra zeharo finkatu gabe baitago, pertsona- zein leku-izenen eta termino teknikoen aldetik batez ere. Gainera, hiztegi hauetan termino bat sartuz gero bere flexio guztia ere ezagutu beharko du zuzentzaileak.

VI.2. Egiaztatzea.

Esan den bezala hitzen ezagutza tratamendu morfologikoaren bidez burutzen da; hau da, hitz bat onartzen da analisi edo deskonposaketa morfologikorik baldin badu, bestela erroretzat hartzen da.

Beste metodoak, n-grametan oinarritutako metodo estatistikoak, zein forma-zerrendan oinarritutakoak, baztertu egin ziren honako hiru arrazoi hauengatik:

1) Berrerabilgarritasuna. Irtenbide *ad-hoc*etik ihes egitea eta berrerabilgarritasuna bultzatzea, bide batez helburu orokorreko prozesadore morfologikoa burutzeko asmoa indartuz.

2) Ortogonaltasuna. Erabiltzaileari oso ulergaitza gertatzen zaio lemaren flexio batzuk onartzea eta beste batzuk ez. Hori gerta daiteke beste metodoekin baina ez ondo eratutako analizatzaile baten bidez.

3) Segurtasuna. Testu-edizioan funtsezkoa da, eta euskararen egoera kontuan hartuz are gehiago, ez ematea ontzat hizkuntzan existitzen ez diren hitzak. n-grametan oinarritutako metodoetan hau gertatu ohi da, eta horixe da metodo hauek baztertzeko arrazoi nagusia aplikazio-mota honetan.

[1] Lan honetan zehar euskara batuarekin edo hobetsitako lexikoarekin bat ez datozen testu-hitzei erroreak edo akatsak deituko diegu, batzuentzat termino hauek gogor samarrak izan arren.

Behin analisi morfologikoaren oinarria aintzat harturik, hirugarren kapituluan azaldu den bi mailatako morfologian oinarritutako analizatzaile morfologiko estandarra egokitu zen, honako ukitu hauek eginez:

- Informazio morfologikoaren bazterketa memoria-hartzea laburtzearren, aplikazio honetan ez baita beharrezkoa. Beraz analisi morfologikoa baino, burutzen dena segmentazio morfologikoa da.

- Hitz zilegien segmentazio morfologiko guztiak ez dira beharrezkoak, analisi zuzen bat duela jakitea nahikoa baita. Hori dela eta, hobekuntza hau burutu da: lehen segmentazioa posiblea lortu bezain laster prozesua eten eta hitza zilegitzat ematen da. Gainera, lehen segmentazioa lehenbailehen lortzeko backtracking-aren bidezko analisi-aukerak sakonean[1] jorratuko dira (ikus §II.5 irudia backtracking-prozedura gogoratzeko). Honen ondorioz hitz zilegien egiaztatze-prozesua azkarragoa izango da erroreena baino, haietan egiaztatzea lehen segmentazioan bukatzen den bitartean, bigarrenetan bide guztiak jorratu behar baitira segmentazio-aukerarik ez dagoela egiaztatu arte.

Dena den, eta aipatutako hobekuntza kontuan hartuz ere, hitz bakoitzaren segmentazio morfologikoak eraginkortasun-galera dakar beste egiaztatze-metodoekin alderatzen badugu. Galera hori ahalik eta gehien murriztearren Peterson-ek (1980) aipatzen zituen buffer-en ildotik egin dugu lan. Buffer horiek lehen kapituluan aipatutako corpusaren bidez osatu dira eta berauetan beti testu-hitzak daude, bilaketa bitarra da. Ondoko hiru mailatan banatzen dira:

- Maiztasun handieneko hitz zilegien bufferra. 8000 bat sarreraz osaturik, horrekin testuetako hitzen %75-80 bat ezagutuz.

- Maiztasun handieneko erroreen bufferra. Hitza buffer honetan aurkituz gero egiaztatzea eten egiten da, zuzenean hitza erroretzat hartuz. Buffer honetan hitz hauei dagozkien proposamenak ere gordetzen dira, zuzenketa azkartzeko asmoz. 800 bat hitzez osaturik dago, eta testuaren arabera estaldura desberdina bada ere, %5-10 tartean dago. Beste hizkuntzetan ez da ohizkoa halako bufferrik erabiltzea, baina, aipatutako batasun-prozesua dela eta, errore “tipikoen” kopuru handiak bultzatzen du buffer honen erabilera.

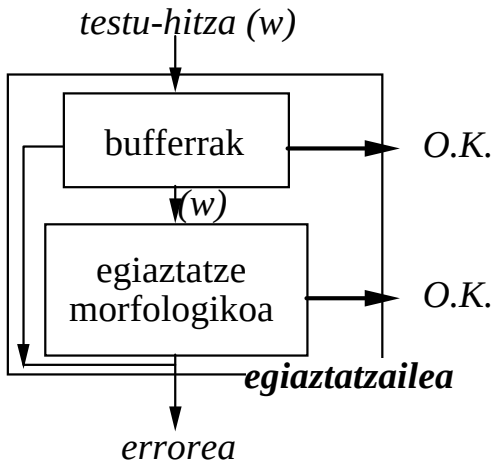
- Testuan aurretik agertutako hitz analizatuak, zilegiak diren ala ez zehaztuz. Hauek dokumentuko bufferra osatzen dute, eta dokumentu-motaren arabera emaitza desberdinak eman ditzake.

---

[1] Analisi morfologikoan sakonean edo zabalean aritzea ez da axola, analisi posible guztiak lortu behar direlako.

VI.1 irudian egiaztatzaile ortografikoaren eskema sinplifikatua ikus daiteke.

Hitz bat ezagutzen ez bada erroretzat hartzen da. Errore baten aurrean zuzentzaile ortografikoak bi bide jorratzen du: errore tipografikoei dagokiena, eta aldaerak edo gaitasun-erroreak deitu ditugun ezjakintasunak bultzatutako erroreei dagokiena. Bi bide hauek paraleloz jorra litezke ordenadorearen ezaugarriek horrela gomendatuko balute; ondoren proposamenak ordenatu egingo baitira.



VI.1 irudia.- Egiaztatzaile ortografikoaren eskema orokorra.

VI.3. Errore tipografikoen tratamendua.

Aurreko kapituluan aztertutako alderantzizko edizio-distantziaren metodoa erabili da zuzenketak edo, aplikazio honetarako egokiago den izenaz, zuzenketa-proposamenak sortzeko.

Metodo honi jarraituz (aipatutako metodoa §V.3.3.1 atalean azaldu zen), akasdun forma batetik abiatuta honako urrats hauek jarraitzen dira:

- Bateko edizio-distantzia duten proposamen hipotetiko guztiak sortu.
- Lortutako proposamen hipotetiko guztiak egiaztatu, VI.2 atalean azaldutako egiaztatzailea erabiliz. Arrakastaz egiaztatzen direnak dira benetako hitzak, gainontzekoak zuzenketa-proposamen gisa baztertuz.

Metodo hau sinplea da, baina bi eragozpen inportante du zehaztasun aldetik, emaitza onak eman baditzake ere:

1) Akatsa eta dagokion zuzenketaren artean edizio-distantzia bat baino gehiago denean, ez da zuzenketa egokirik eskainiko.

2) Forma hipotetiko guztiak egiaztatu behar izatea ez da eraginkorra, morfologian oinarritutako egiaztatze-prozesu bat burutzen denean batez ere.

Lehen eragozpenaren aurrean edizio-distantzia bira zabal liteke baina horrekin zuzenketak lortzeko denbora izugarri haziko litzateke, zuzenketa-prozesua ia ezinezkoa bihurtuz —gogoratu behazehaztasuna/eraginkortasuna oreka bermatu behar dela. Horren arrazoia zeran datza: proposamen hipotetikoen kopuru izugarria eta hauetako bakoitza morfologikoki egiaztatzeko beharra.

Hala ere, eta lehen eragozpen hori erlatibizatuz, bi irizpide hartu behar dira kontuan:

- Zuzentzaile ortografikoa biko edo edizio-distantzia handiagoko erroreak zuzentzeko gai izango dela, erroreek aldaeren ezaugarriak dituztenean.
- Aurretik esan dugun bezala, errore tipografikoen zuzenketa ez da diseinu-helburu nagusia.

Dena dela kapitulu honen bukaeran (ikusi §VI.8) eragozpen hauek berraztertzeari eta aurreko kapituluan hizkuntza eranskarietarako egindako proposamenarekin alderatzeari ekingo diogu.

VI.3.1. Azkartzeko bideak.

Aipatutako bigarren eragozpenaren aurrean, proposamen hipotetiko guztien egiaztatzearen beharraz hain zuzen, zenbait heuristiko erabili dugu proposamenen sorkuntza ahalik eta azkarren buru dadin. Bestela, aurreko kapituluan esan zen bezala, bateko edizio-distantzian egon daitezkeen forma hipotetikoak *2nk* inguru dira *n* hitzaren luzera eta *k* alfabetoaren karaktere-kopurua izanik (ikus §V.3.1.1), eta denak egiaztatu beharko lirateke akats bakoitzeko. Gainera, hauetako proposamen hipotetiko gehienak benetako hitzak ez direnez, haien egiaztatzea motelagoa izango da benetako hitzena baino.

Azkartzeko teknika horiek bi motakoak dira: batetik, zuzenean aplikatzen direnak, zehaztasunean eragin adierazgarririk ez dutelako; eta aukeran jartzen direnak bestetik, zehaztasunean eragina izanik zehaztasuna eta eraginkortasunaren artean hautatzen den orekaren arabera.

Lehen multzoan daude proposamenen bufferra eta trigramen bidezko proposamenen sorrera. Bigarrenean aldiz, morfemen selekzioa eta proposamenen zein analisien kopurua murriztea koka daitezke.

VI.2 irudian prozesu osoaren eskema azaltzen da.

Proposamenen bufferra.

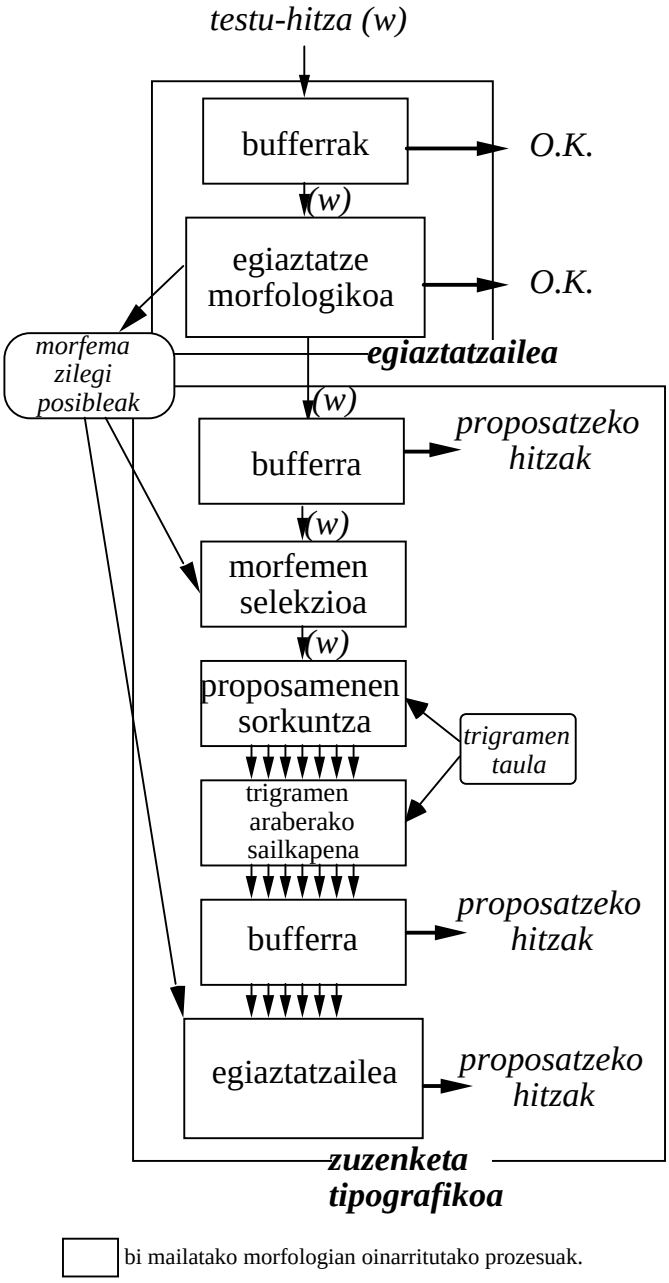
Egiaztatze-prozesua azaldu denean, esan den bezala maiztasun handieneko hitz zilegiak aparte, maiztasun handiko erroreak ere gordetzen dira, azken hauek dagozkien proposamenekin batera.

Azken buffer honen bidez maiztasun handieneko akatsak ezagutu eta zuzen daitezke urrats batean, hipotesien sorrera eta egiaztatzea saihestuz. Bufferrean gordetzen diren proposamenak errore tipografikoei zein aldaerei dagozkienak dira, eta aurreprozesu batean lortzen dira.

Prozedura azkartzeko helburuari jarraituz eta VI.2 irudian ikusten denez, proposamen hipotetikoak bilatzen dira hitz zilegien bufferrean, egiaztatzera joan baino lehen.

Trigramen bidezko proposamenen sorrera.

Aurreko kapituluan aipatu den bezala n-gramen erabilerak arrakasta handia izan du zuzenketaren alorrean, Trigramak dira n-grama erabilienak memori hartzearen eta esanguratasunaren artean oreka handiena eskaintzen dutelako. Xuxen zuzentzailerako trigramen taula bat osatu da, trigrama posible bakoitzari dagokion maiztasuna esleituz, corpusetatik lortutako datuen arabera.



VI.2 irudia.- Errore tipografikoen tratamendu eraginkorra.

Proposamenen sorkuntza azkartzeko trigramek tratamendu bikoitza bideratzen dute:

- Erroredun formaren trigramak aztertzen dira eta zilegi ez diren trigramak —taulan agertzen ez direnak hain zuzen— bakarrik hartzen dira kontuan Damerau-ren oinarrizko tipifikazioa aplikatzean. Trigrama guztiak zilegiak badira, aldiz, hitz osoaren gainean aplikatzen dira.

- Damerau-ren oinarrizko tipifikazioa aplikatutakoan trigrama ez-zilegirik duten proposamen hipotetikoak egiaztatu gabe baztertzen dira. Horrez gain,

proposamen hipotetikoak haien barneko trigramen pisuaren arabera sailkatzen dira egiaztatzeari begira.

Morfemen selekzioa.

Esan den bezala, aukeran den tratamendu hau hiru urratsetan burutzen da:

1) Hitzaren egiaztatze morfologikoa burutzen den bitartean, aurkitutako morfemak edo morfema-multzoak eta dagozkien lexikoko posizioa zein automaten egoerak gordetzen dira datu-egitura batean. Prozesua eskerretatik eskuinetara burutzen denez aurkitutako morfemak edo morfema-multzoak beti dira hasierakoak.

2) Hitza zilegia bada aurretik gorde dena baztertu egiten da baina ezagutzen ez bada proposamenen tratamendua metatutako morfemetatik abiatzen da. Horrela hasierako morfemak ontzat emango dira eta Damerau-ren araberako aldaketak ezagutu ez den partean baino ez dira aplikatuko. Oinarrizko proposamen kopurua *2nk* inguru izan beharrean, *2(n-l)k* inguru izango da, *l* ezagututako morfemaren luzera izanik.

3) Proposamen hipotetikoak egiaztatzeko analisi morfologikora jo behar baldin bada, gordetako egoeratik hasiko da analisia, eta ondorioz askoz azkarragoa izango da.

Aurkitutako morfemak anitzak badira, egoera bakoitzetik abiatzen da analisia; hipotesiak sortzeko, ordea, morfema edo morfema-multzo luzeena hartzen da erreferentziatzat.

Esan den bezala zehaztasunaren kaltean izan daiteke tratamendu hau, zeren akats baten bidez hitzaren ezkerreko partea morfema desberdin bat bilakatzen bada, morfema horretatik abiatuta ezinezkoa izango baita akatsa zuzentzea. Izan ere, §V.3.1.2 atalean azaldu den bezala, edizioan probabilitate handiagoz egon daiteke errorea hitzaren bukaeran hasieran baino (Yannakoudakis, 83).

Proposamenen kopurua murriztea.

Proposamenak sortzeko prozesua azkartzeko funtsezko parametroa analisi morfologikoen kopurua da, hauxe baita atalik konplexuena konputazioaren ikuspuntutik. Horren ondorioz hipotesi batzuk baztertu egingo dira egiaztatu gabe; baina litekeena da horietako batzuk benetako hitzak izatea, eta are gehiago hitzari zegokion zuzenketa. Beraz, ondorio kaltegarria ekar diezaioke zehaztasunari eraginkortasun hobetze honek, analisi-kopurua murriztea proposamen kopurua murrizteak ekar baitezake. Ondorio kaltegarri hori dela eta, tratamendu hau aukerazkoa eta parametrizagarria izango da.

Parametrizatze hori bi aldagairen arabera egin daiteke —argi egon arren haien artean erlazio zuzena dagoela—:

- proposamen kopuruari muga jartzea analisiak burutzen hasten denetik.
- analisi kopurua mugatzea proposamen kopuruari jaramonik egin gabe.

Bi irizpideak konbina daitezke eta horrela egin dugu gure produktu komertzialean (ikus §VI.6 atala).

Beste aldetik, eta VI.2 irudia aztertuz ikus daitekeenez, proposamen hipotetikoak banan-banan egiaztatu beharrean bi urratsetan burutzen da: lehenean hipotesi guztiak bilatzen dira hitz zilegien bufferrean; horrela ez da analisi morfologikorik egin behar, han aurkitutakoekin proposamen kopurua osatzen baldin bada.

VI.4. Gaitasun-erroreen zuzenketa.

Errore mota hau da diseinu irizpideen arabera zuzentzeko lehentasuna duena. V.3.1.4 atalean aipatzen zen bezala, errore hauek tratatzeko arrazoi nagusia zera da: beste motako erroreen zuzenketa nola egin erabiltzaileak jakin ohi duen bitartean, hauena normalean ez du jakiten. Hainbestetan aipaturiko euskararen egoera dela eta, hauen portzentaia beste hizkuntzetakoa baino handiagoa denez, are interesgarriago bihurtzen da prozesaketa hau.

Analisi estandarraren bidez ezagutu ez den hitz bat aldaeratzat hartzeko, laugarren kapituluko bigarren atalean azaldutako aldaeren analisi morfologikoaz ezagutu behar da. Han azaldutakoa puntu hauetan labur daiteke:

- Sistema estandarreko lexikoa eta erregela-sistema osatzen dira, azpilexiko multzo berri eta erregelen azpisistema berri bana erantsiz.

- Azpisistema osagarri horren bidez horrelako kasuak aurrikusten dira:
  1) Morfemen aldaera: morfema baten ordez beste bat erabiltzetik datozen akatsak. Jatorrizko morfemaren eta aldaerari dagokionaren arteko desberdintasuna diakritikoren bat bada, morfema konkretu hau lotzean egiten diren akatsak ere ezagutzen dira. Morfemaren aldaera eta dagokion zilegia lotzen dira lexikoan.
  2) Morfotaktikaren aldaera: morfema baten ondoren etor daitezkeenak aldatzetik datozen erroreak.
  3) Aldaera erregularrak: errore fonologiko, morfologiko eta ortografiko erregularrak bi mailatako erregela osagarrien bidez ezagutzen dira.

- Analisia burutzean azpisistema osagarria estandarrarekin batera ibiltzen da, hitzetan bi motako morfema zein aldaketa morfofonologikoak gerta daitezke eta.

Analisirik aurkitzen baldin bada, morfemen aldaerei dagozkien morfema estandarrak bilatu behar dira lexiko-loturaren bidez.

Orain arte ikusitakoa laugarren kapituluan sakonean azaltzen da, baina zuzentzaile ortografikoan gaitasun-erroreetarako aldaeren analisi morfologikoan egiten ez zen prozedura bat burutu behar da: aldaerari dagokion zuzenketa edo forma estandarra lortu behar da. Zuzenketa hau burutzeko sorkuntza morfologikoa erabiliko da: analisian lortutako morfema estandarrak lotzen dira erregela estandarrak erabiliz.

Hala ere **arazo** bat dago, morfotaktikaren aldaerak sortutako akatsak ezagutu arren ezin baitira zuzendu. Arrazoia sinplea da, aldaera honen bidez erabiltzaileak eraikitako hitzaren osagaiak zilegiak dira baina ez kateatze konkretu horretan. Lexikoa diseinatu den bezala behintzat, ezinezkoa da zuzentzea, aldaeraren deskribapenean dagoen jarraitze-klasea estandarrarekin loturik ez dagoenez gero, ez baitago jakiterik bi jarraitze-klase horietan dauden morfemen artean zer erlazio dagoen. Aldaera hauen deskribapen konplexuago baten bidez zuzentzeko aukera egon liteke, baina ez dirudi halakorik egiteak pena merezi duenik, aldaera mota hau bereziena dela kontuan hartzen badugu.

Beste irtenbidea hauxe izan daiteke: morfotaktikaren aldaera morfemen aldaera multzo bezala adieraztea, hau neketsua izan badaiteke ere.

Adibide gisa *batzu* morfemaren jarraitze-klasea dugu. IV.3 irudian azaltzen zen bezala, jarraitze klase estandarra *PLU* (plurala) da, baina aldaera gisa *MG* (mugagabea) erabili ohi da[1]. Honen zuzenketa burutzeko zeharkako aukera hau legoke:

- *PLU* azpilexiko bakarra denez, berau osatzen duten morfemak azpilexiko berri batean bikoiztu, alomorfoak sortuz. Horrez gain, *batzu*-ren jarraitze-klasea aldatu beharko litzateke.
- Azpilexiko hori bikoiztu osagarri ez-estandar batean, morfema bakoitzari dagokion *MG*-ko morfema zehaztuz morfemaren aldaera gisa.

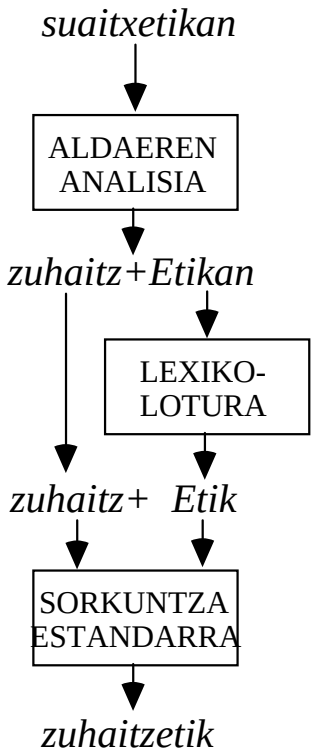
Lehen urratsa aurrez daiteke baina horrela izugarrizko gainsorrera bultzatzen da, *PLU* jarraitze-klasea erabiltzen duten lema guztietarako suposatzen ari baikara aldaera lokal bat.

Maiztasun handia duten mota honetako aldaeren zuzenketa *ad-hoc* edo partikularra bidera daiteke maiztasun handieneko hitzen bufferraren bidez. Morfotaktikaren aldaerei dagozkien formatarako, eta aipatu den aurreprozesaketaren bidez proposamen egokirik lortu ez bada, salbuespen gisa ukituak egin daitezke bufferrean.

---

1 Orain dela gutxi arte bien erabilera onartuta zegoen, baina orain pluralarena bakarrik onartzen da.

Gaitasun-erroreen zuzenketa bere osotasunean azaltzeko har dezagun *suaitxetikan* hitzaren **adibidea**. Forma honi dagokion zuzenketa-prozesua VI.3 irudian agertzen da.



VI.3 irudia.- *suaitxetikan* hitzaren zuzenketa aldaeren analisi eta sorkuntza morfologiko estandarraren bidez.

suaitxetikan zuhaitzetik forma estandarraren aldaera bezala ikus daiteke —adibidea muturrera eraman da, baina zuzentzeko aukerez jabetzeko oso egokia—. Azpimarratzekoa da zuzentzea badagoela edizio-distantzia bostekoa izan arren. Dagokion analisia IV.2.3.1 atalean azaldu zen, bertan zera ikus daitekeela:

1) *zuhaitz* lema lortzen da *suaitx* hitz-zatitik bi erregela osagarriri esker: arrakasta bi aldiz duen txistukarien arteko aldaketarena (z-s, z-x) eta h-ren galerarena.

2) *Etikan* atzizkia lortzen da azpilexiko osagarriei esker.

3) Aipaturiko azpilexikoetan *Etikan Etik* forma estandarrekin agertzen da lotuta.

Behin analisia burutu ondoren, eta zuzentzeko arazorik ez dagoela ikusita sorkuntzara pasatzen da, *zuhaitz+Etik* lexiko-karaktereetatik erregela estandarrei dagozkien itzultzaileen bidez *zuhaitzetik* lortuz.

Bibliografian aipatutako errore fonologikoen zuzenketarekin alderatuz gero, gaitasun-erroreak tratatzeko sistema orokorra, dotorea zein beste moduluekiko homogenoa bihurtzen da gure burutzapenean.

Gure sistemarako 30 aldaketa baino gehiago deskribatzen duten 18 erregela osagarri (ikus §IV.2.2.2 atala), eta ia mila morfema ez-estandar landu dira, oso azpisistema ahaltsua osatuz.

Azkenik aipatu behar da metodo honen bidez “hitz-mugaren gaineko errore” batzuk zuzen daitezkeela. Banaturik idazten diren zenbait forma, *hitz egin* adibidez, batera idaztea aldaera bezala jaso daiteke morfema ez estandarretan, *hitzegiN hitz_egiN* forma estandarrarekin lotuz; eta horrela, morfema horien flexioa zuzen daiteke.

VI.5. Sistemaren arkitektura eta ezaugarriak.

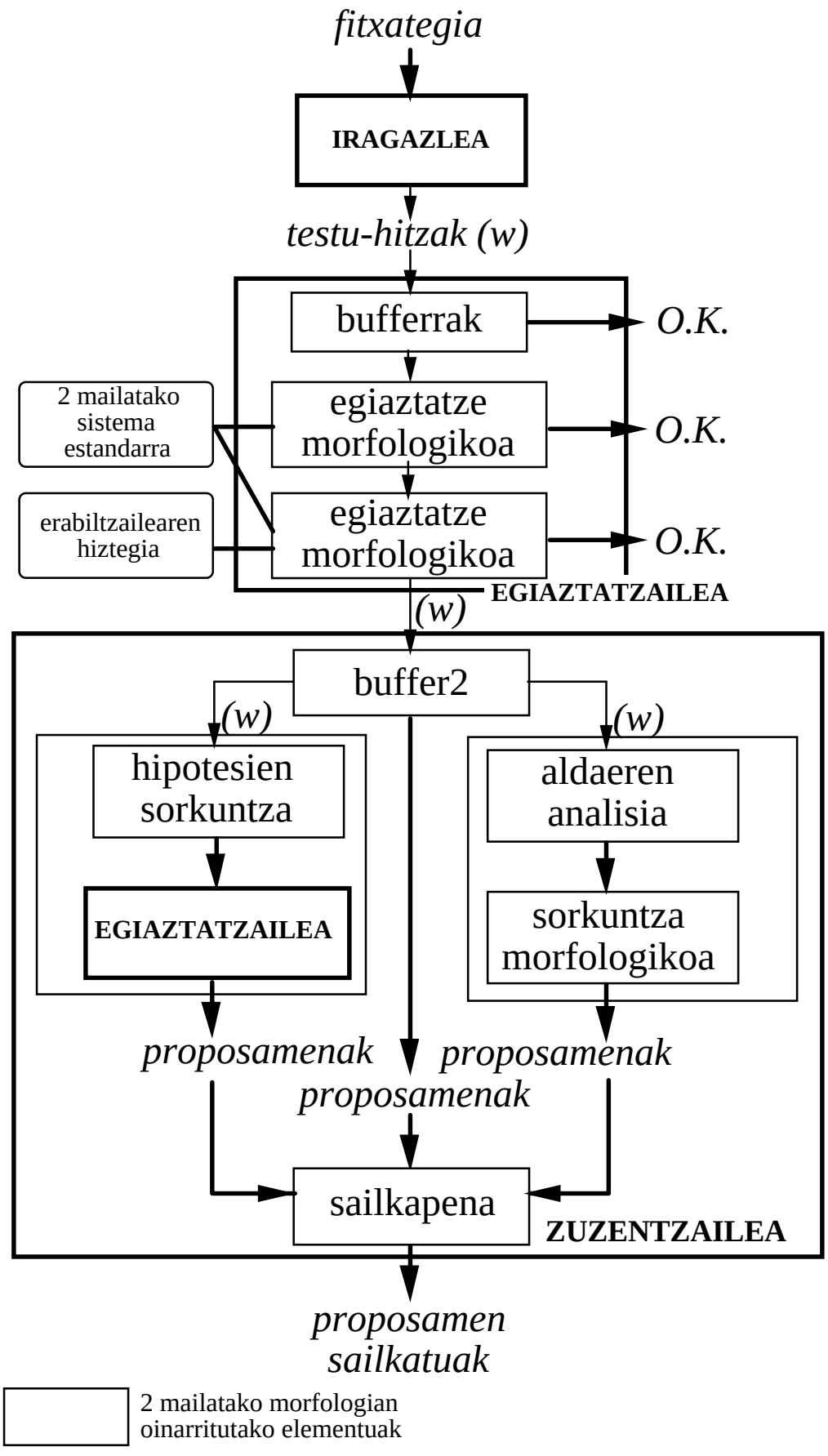
Aztertutako modulu egiaztatzailea eta zuzentzailea zuzentzaile ortografikoaren funtsa dira baina ez osagai bakarrak. Horiekin batera ondoko modulu hauek gehitu behar dira, kalitateko zuzentzaile bat burutu nahi bada:

- Proposamenak sailkatzeko modulua. Zuzentzaileak sortutako proposamenak ordenatzea da horren helburua. Zuzenketa automatikoa behar den sistemetan funtsezko modulua da.
- Erabiltzailearen hiztegiaren kudeatzailea. Egiaztatzailearekin lotuta dago, hiztegi hau kontuan hartu behar delako egiaztatzean, baina eguneratzeko aukera ere eskaini behar zaio erabiltzaileari.
- Iragazlea edo token-ezagutzailea. Fitxategi bat irakurriz hitzen eta testu-unitate berezien ezagutzaz arduratzen da, eta egiaztatu aurretik burutzen da.

Aipatutako osagai hauek banan-banan aztertuko dira ondoren, eta aurretik deskribatutako funtsezko moduluekin VI.4 irudian azaltzen den arkitektura osatzen dute.

VI.5.1. Proposamenen sailkapena.

Proposamenak sailkatzerakoan komenigarria litzateke testuingurua kontuan hartzea baina, esan den bezala, hau gure lanaren esparrutik kanpo dago. Testuingurua erabiltzen ez bada proposamenak sailkatzeko honako teknika hauek erabil daitezke (ikus §V.3.3.1):



VI.4 irudia.- Xuxen zuzentzaile ortografikoaren arkitektura.

- Edizio-distantzia edo aipatutako bestelako distantziaren neurriak. Azken hauek dira zuzenketa automatikoan erabiltzen direnak. Dena den edizio-distantzia erabiltzen bada metodoren batez bereizi beharko dira distantzia bera dutenak.

- Errore-corpusen gainean aplikatutako metodo estokastikoak, taula estatistikoak, eredu markoviarrak edo sare neuronalen bidezkoak erabiliz.

- Proposamenen maiztasuna: estatistikak, errorearen eta dagokion zuzenketaren arteko erlazioan oinarritu beharrean, hitz zilegien datu sinpleetan oinarri daitezke. Metodo hau aurrekoa baino askoz sinpleagoa da, baina errore-corpusik ez da behar.

Gure kasuan sailkapena proposamen hipotetikoen gainean egin zitekeen, baina kontuan hartu behar da kasu horretan aldaeren tratamenduaz sortutako proposamenak sailkapenetik at geldituko liratekeela. Beste aldetik errore-corpus fidagarririk ez dagoenez bigarren puntuko irizpideak ezin izan dira aplikatu, eta beraz, hirugarrenekoak aplikatu dira.

Proposamenak egiteko honako algoritmoari jarraitzen zaio, bai maiztasuneko handieneko akatsei dagozkien zuzenketak sortzean, bai zuzenketa-prozesuan zehar:

1) Bateko edizio-distantzian eta maiztasun handieneko hitz zilegien bufferrean dauden proposamenak. Hauek maiztasunaren arabera sailkatu beharko lirateke, baina *maiztegi* izeneko buffer horretan maiztasunaren balioa ez da jasotzen memoria-hartze arrazoiak direla eta. Horren ordez barneko trigramen pisuaren arabera sailkatzen dira.

2) Aurreko puntuan sartzen ez diren aldaeren tratamenduz lortutako proposamenak, hurrenez hurren edizio-distantziaren arabera eta barneko trigramen eraketaren arabera sailkatuak.

3) Gainontzeko proposamenak morfologikoki egiaztatu ahala, aurretik barneko trigramen eraketaren arabera sailkaturik baitaude.

Algoritmo honekin lortutako emaitzak VI.7 atalean azaltzen dira.

VI.5.2. Erabiltzailearen hiztegia.

Hizkuntza eranskarien zuzenketan dauden arazo berezien artean, hiztegiaren aberasketa aipatu da aurreko kapituluan (ikusi V.4.1 atala).

Sistema komertzial batzuetan egiten den hitz-zerrendaren bidezko erabiltzailearen hiztegia ez da, inola ere, egokia hizkuntza eranskarietarako. Hitzen ordez morfemak —gehienetan lemak— metatu eta erabili behar dira.

Xuxenerako IV.1 atalean azaldutako erabiltzailearen lexikorako burutzapena berrerabili da, horrekin helburu bikoitza lortuz:

- Azpilexiko **irekiak** definitzeko aukeraren bidez lema berriak erabiltzailearen hiztegietan gordeko dira. Lemari dagokion azpilexikoa, jarraitze-klasea eta bi mailatako morfologiaren araberako lexiko-maila (diakritikoak eta guzti beharrezkoak bada) lortu behar dira linguistikan aditua ez den erabiltzailearengandik. Horretarako IV.1.3 atalean zehazten den informazio morfosintaktikoa —kategoria eta, kasuaren arabera, azpikategoria eta ezaugarriren bat— eskatzen zaio erabiltzaileari interfaze atsegin eta sinple batez (ikus VI.6 irudia).
- Egiaztatze morfologikoa bi urrats edo gehiagotan burutzen da: lehenean lexiko orokorra kontsultatzen den bitartean, ondorengoetan erabiltzailearen hiztegiak kontsultatzen dira lexiko orokorreko azpilexiko orokorrak ere erabiliz, VI.4 irudian ikus daitekeenez. Aplikatzen diren erregela morfofonologiko estandarrak ez dira aldatzen urrats desberdinetan.

Horrela hiztegi desberdinak erabil daitezke jakintza-arloaren arabera eta “benetako hitzaren errore” batzuk saihestu egingo dira, hitz orokor bat, akatsen baten eraginez, beste jakintza-arloko hitz berezitu bat bihurtzen denean.

VI.5.3. Iragazlea edo token-ezagutzailea.

Hitzak eta beste testu-unitateak bereiztea da modulu honen helburua. Analizatzaile estandarrerako egindakoa berrerabili da zenbait aldaketa eginez. Bi automatatan oinarritzen da eta bere zeregina ondoko puntuetan bana daiteke:

- Zenbakiak, arruntak edo erromatarrak, bereiztea dagokien deklinabidearekin batera. Deklinabidea ondo dagoen ala ez ziurtatzeko egiaztatzera bidaltzen da.
- Laburdurak eta siglak identifikatzea dagokien deklinabidearekin. Egiaztatzera bidaltzen dira.
- Lerro-bukaeran hitza banatzen duen marratxoa (*hyphenation*) ezagutu eta kontuan ez hartzea. Marratxoaren tratamendu korapilatsua azpimarra daiteke: aipatutako funtzioaz gain elkarketarena, erdal hitzen atzizkiekiko lotura eta bereizgarri-funtzioa ere izan baititzake.
- Zuriuneak, puntuazio-zeinuak eta gainontzeko hitzen arteko bereizgarriak ezagutzea.

- Maiuskulaz osoki idatzitako hitzekin tratamendu berezia egitea, maiuskulaz ezagutzen ez badira minuskulaz ere egiaztatuko direlarik.
- Testuetan karaktere arraroak, bereizgarriak edo hizkuntzarenak ez direnak agertzen badira, horren berri ematea erabiltzaileari.

Zenbait informazio metatzen da zuzenketak eskaintzeko orduan kontuan har dadin. Horrela, hitza osoki edo hasieran maiuskula duen ala ez, zenbaki eta laburduren kasua, etab.

Zuzentzaile komertzial batzuetan aurreko eragiketa batzuk aukeran ematen zaizkie erabiltzaileei, horrela hauek maiuskulaz idatzitakoa, zenbakiak, siglak, laburdurak, karaktere arraroak etab. egiaztatu nahi dituzten ala zuzenean ontzat eman nahi dituzten hauta dezaten.

Proiektuan gure diseinu-filosofiarekin bat etorriz —zalantza kasuan nahiago izan dugu abisua ematea ontzat ematea baino, eta horrexegatik ekidin dugu gainsorrera— nahiago izan dugu dena egiaztatzea. Hala ere, etorkizunean halako parametrizazioak egitea litzateke egokiena.

VI.6. Produktu komertzialaren diseinua.

Aurreko ataletan azaldutako osagaiekin zuzenketarako prototipo parametrizagarria osatu genuen VI.4 irudian agertzen den arkitekturari jarraituz. Prototipo hori produktu komertzial bihurtzeko eman behar izan diren urrats garrantzitsuenak honako hauek izan dira:

- Aukerazko azkartze-mekanismoak erabakitzea, makinaren arabera zehaztasuna/ eraginkortasuna oreka mantenduz.
- Interfaze atsegina diseinatzea, erabiltzailearekiko hartu-emanak ahalik eta modu sinple bezain esanguratsuenean buru daitezen.
- Makina zein testu-editore zehatz bakoitzerako egokitzapena. Macintosh eta PC izan ziren aukeratutako oinarri-ordenadoreak eta Word eta WordPerfect testu-editoreak.

Lehen bi puntuak interesgarriak izan daitezkeelakoan zabaldu egingo ditugu ondoren.

VI.6.1. Zehaztasuna/eraginkortasuna oreka.

Zehaztasuna/eraginkortasuna oreka mantentzea da LNPko aplikazio guztien ardatzetako bat, eta produktu komertzial baten arrakastarako aldagai nagusietako bat.

Xuxen zuzentzaile ortografikoa merkaturatzeko orduan prototipoa erabiliz neurriak hartu genituen, ondoko ondorio kualitatiboak lortuz:

- Zehaztasun aldetik kalitate handiko egiaztatze/zuzenketa egiten zuela, bateko distantziatik gorako errore tipografikoen kasuaren salbuespenaz.
- Abiaduraren aldetik motela zela konputagailu pertsonaletan korritzeko beste produktu komertzialekin alderatuz gero, hizkuntza eranskarietarako emandako datuekin parekagarria bada ere.

Horren aurrean abiaduraren aldera orekatzea erabaki genuen aukerazko azkartze-metodo guztiak erabiliz, zehaztasunean ahalik eta eragin txikiena eraginez noski. Ondorioz, horrela moldatu genituen azkartze-metodoak:

- Hitzaren hasieran aurkitutako morfemei buruzko informazioa erabiltzea hipotesi kopurua laburtzearren eta hipotesien analisi morfologikoa azkartzearren.
- Proposamenen kopurua mugatzea analisi morfologikoen kopurua laburtzeko. Bide horretan gaitasun-erroreen tratamendutik eta bufferrean bilatzetik sor daitezkeen proposamen guztiak lortzen dira, baina hipotesien analisi morfologikoari aurreko bideetatik proposamenik sortu ez bada soilik ekiten zaio. Beraz, proposamen bat lortuz gero ez da analisi morfologiko gehiagorik egiten. Gainera, analisi kopuru mugatu batera iritsiz gero, proposamenik lortu ez bada ere, zuzenketa-prozesua eten egiten da, eta horrexegatik da funtsezkoa hipotesien sailkapena barneko trigramen arabera. Analisi kopuruaren muga hitzaren luzeraren menpe dago, hitza zenbat eta luzeago analisi morfologikoa hainbat eta motelago baita.

Bide honetatik lortzen da zuzenketarako abiadura onargarria konputagailu pertsonaletan.

VI.6.2. Erabiltzailearekiko interfazea.

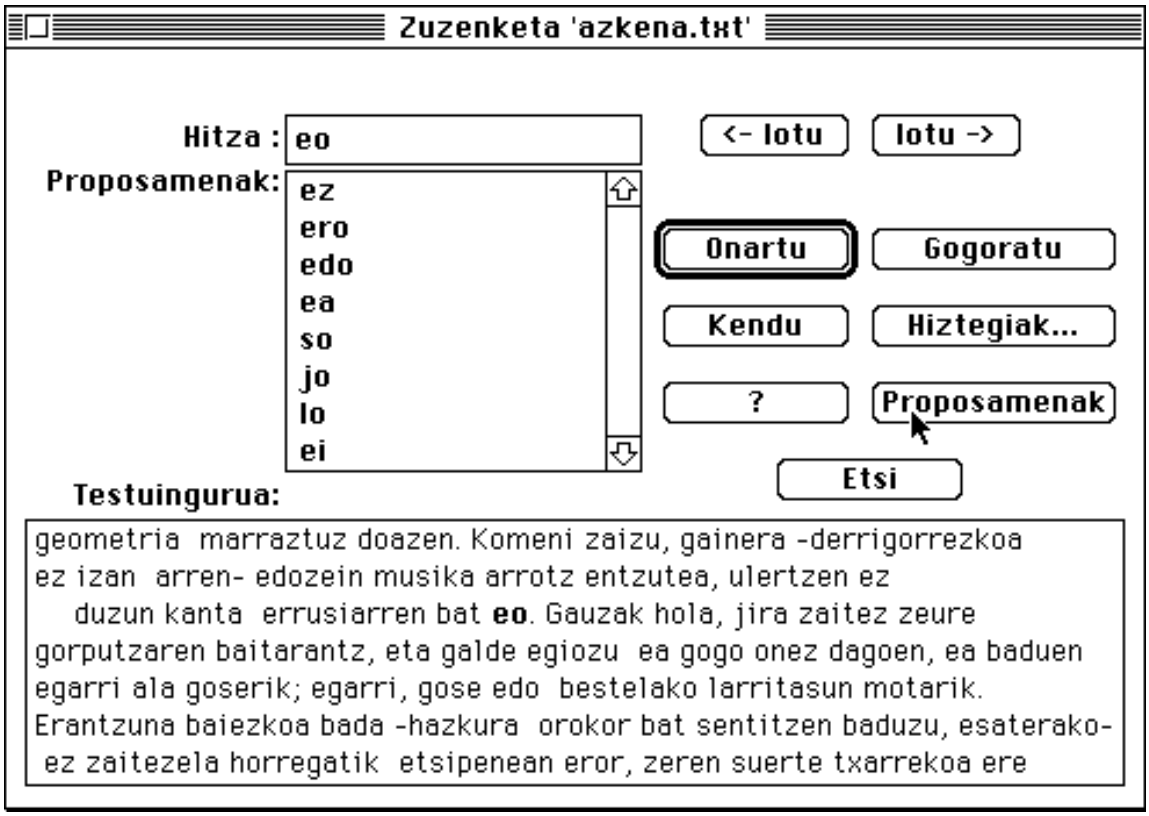
Zuzentzailea merkaturatzeko beste funtsezko ekimen bat interfazea egoki eta atsegina izaten da.[1]- GUIen garaian, leihoetan oinarritutako interfaze-sistema bat, objektuei

[1] GUI: Graphical User-Interface (Erabiltzaile-Interfaze Grafikoa)

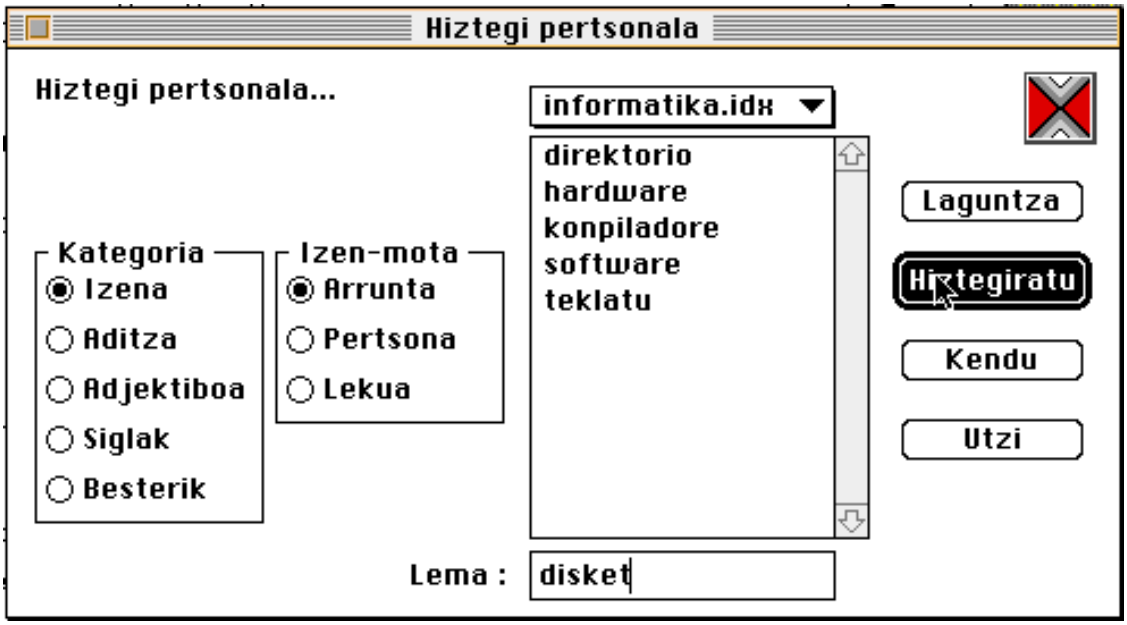
zuzendua eta atzean gertaerei zuzendutako programazio-eredua duena, ia-ia ezinbestekoa da produktua arrakastatsua gertatzeko.

Ildo honetatik garraigarritasuna ziurtatzen duen ingurune batean programatzea ezinbestekotzat jo daiteke zuzentzailea plataforma desberdinetarako eskaini nahi bada. XVT izan zen ingurune garraiagarria garatzeko aukeratu genuen software-paketea.

Interfazeari buruzko zehaztasun gehiago ematearren VI.5 eta VI.6 irudietan bi leiho nagusiak azaltzen dira: zuzenketa-prozesua kudeatzekoa eta erabiltzailearen hiztegia aberastekoa.



VI.5 irudia.- Xuxen zuzentzaile ortografikoaren leiho nagusia



VI.6 irudia.- Xuxen zuzentzailean erabiltzailearen lexikoa aberasteko leihoa.

VI.7. Zehaztasuna eta eraginkortasuna.

Aurreko ataletan deskribatutako egiaztatze- eta zuzenketa-prozesuen gainean hartutako datuak azaltzen dira pasarte honetan.

VI.7.1. Egiaztatzea.

Zehaztasunaren aldetik III.9 eta III.10 irudietan azaldutako emaitzak aipa daitezke hastapen gisa. Horren arabera testu-hitzetako %91-96a ezagutzen da, beren analisia lortzen da eta. Erabiltzailearen lexikoa erabiliko balitz portzentaia hori igo egingo litzateke, analizatzen ez diren hitzen erdia baino gehiago ez baitira erroreak; baina ez dira ezagutzen dagokien lema lexikoan ez dagoelako. Hala ere, kontuan hartu behar da aipatutako testuetan akats tipografiko gutxi dagoela, argitaraturiko testuak dira eta.

Egiaztatu gabeko testuak izaten dira zuzentzaile ortografikoen helburua eta horrelako testu batzuk aztertuz lortu dira VI.7 irudian azaltzen diren emaitzak. Bertan hiru iturburu desberdinetatik eskuratutako testuak agertzen dira, antzeko tamainakoak direnak:

- Ilazki euskaltegian euskara ikasten duten azken urratseko ikasleek egindako testuak[1], ikasle batek sakaturik —ikasleen testuak deitutakoak. Hauetan aldaera gehiago dago besteetan baino, aldaeretan gaitasun-erroreak ere sartzen baitira.

Testuak	**hitzak**	**ezagutu gabe**	**ondo[2] daudenak**	**ezagutut. aldaerak**	**errore tipograf.**
1.-Ikasleen testuak	3.959	326 **%8,2**	63 [3]%19,3	171 **%52,5**	92 **%28,2**
2.-Testu teknikoa	3.891	220 **%5,6**	32 %14,5	32 **%14,5**	156 **%71**
3.-Prentsako testuak	4.319	205 **%4,7**	42 %20,5	79 **%38,5**	84 **%41**
GUZTIRA	12.097	751 %6,3	137 %18,2	282 %37,6	332 %44,2

VI.7 irudia.- Egiaztatzeari buruz hartutako estatistikak.

- UZEIn sortu eta sakatutako testu tekniko zuzendu gabeak. Terminologia teknikoa kenduta testu estandarra da eta horregatik aldaera gutxi detektatzen dira. Terminologiaren arazoak ebazteko hiztegi berezitu bat aberastu da aurretik.

- *Egunkaria*tik lortutako prentsako testu zuzendu gabeak. Estandarrak dira hein batean, hala ere izen nagusien eragina saihesteko maiuskulaz hasitako izenak ez dira kontuan hartu, eta horregatik ezagutu gabeko hitzak gutxiago dira beste testu-zatietan baino.

Beraz, irudian azaltzen diren datuetatik atera daitezkeen ondorioak espero zitezkeenak dira.

Jakintza-arloarekin lotutako lexikoa hiztegi berezituetan antolatzeagatik, eta egindako deskribapen morfologikoari dagokion gainsorrera-ezarengatik, **benetako hitzaren erroreak** ekiditen dira ahal den neurrian. Horien zenbatekoa corpusetan oinarriturik kalkulatzea zaila da, automatikoki egitea ezinezkoa da eta, ondorioz lagin adierazgarria aztertzea oso neketsua izango litzateke. Horren ordez hurbilpen estatistikoa egin dugu.

---

1 Testu hauek M. Maritxalarrek bildu ditu bere ikerketarako OLIren esparruan (Maritxalar & Diaz de Illarraza, 93).

2 Lexikoan ez egoteagatik edo beste hizkuntzetako hitzak izateagatik ondo dauden baina ezagutu ez diren hitzak.

3 Ondoko hiru portzentaiak egiaztatu gabeko hitzen gainean kalkulatzen dira.

Luzera	Hitz Kopurua.	Benetako hitzak(%)
2	18	9,7
3	74	6,8
4	81	6,0
5	56	5,5
6	67	3,0
7	61	2,6
8	39	1,7
9	41	1,5
10	21	0,9
11	20	1,0
12	13	1.0
13	3	0.5
14	3	0.4
best.	3	0.0
GUZT.	500	4.0

VI.8 irudia.- Benetako hitzaren erroreen probalititatea (hurbilpena).

Hurbilpen horretan testu bateko 500 hitz[1] zilegi jarrai hartu dira, eta bateko edizio-distantzian dauden forma guztien artean benetako hitzen portzentaia kalkulatu da, %4 ingurukoa izanik errore hauen probabilitatea. VI.8 irudian luzeraren araberako datuak aurkezten dira.

Datu hauek minimotzat jo behar dira; hurbilpenean egin diren bi sinplifikazioen artean, erroreak beti bateko distantzian daudela eta bateko distantzian daudenek probabilitate bera dutela alegia. Bigarrenak batez ere eragin nabarmena eduki dezake emaitza hori txikiagotuz, frogatutzat har baitaiteke erroreak sortzean benetako hitzak idazteko joera handiago dagoela arrazoi sikolinguistikoak direla eta.

Abiaduraren aldetik analisi morfologikoarena 2 hitz segundoko da—III.5.4n esaten zen bezala Sun-Sparc IPX baterako—, eta egiaztatzearena 25-30 hitz segundoko izatera iristen da bufferren erabilerarengatik eta lehen analisiarekin analisi-prozesua bukarazteagatik. Izan ere, zuzenketa-fasean egiten diren egiaztatzeak, existitzen ez diren hitzei dagozkienez, analisi morfologiko osoen denboratik gertuago daude hitz arrunten egiaztatzeenetik baino.

1 Neurri honekin lortzen diren emaitzak egonkorrak direla egiaztatu da.

VI.7.2. Zuzenketa.

Zuzenketaren zehaztasunari eta abiadurari buruzko datuak lortzea oso inportantea da bi arrazoirengatik: bi aldagai horien artean bilatu beharreko oreka-puntua bilatzeko eta bibliografian dauden beste sistemekin alderatzeko.

Egiaztatzeari buruzko estatistikak lortzeko erabilitako corpus bakoitzetik 100 errore hartu eta horren gainean VI.9 irudian azaltzen diren estatistikak lortu ditugu. Datu kopurua dela eta fidagarritasuna mugaturik egon arren, estatistiken emaitzak interesgarriak dira. Estatistika horietan erabilitako aldagaiak hauek dira:

A) Zutabeetan lau aukera agertzen dira: 1) Xuxen zuzentzaile komertzialean erabili dugun zuzenketa azkarra, proposamen hipotetikoen analisia aurkitutako morfemetatik abiatzen duena eta hauen kopurua mugatzen duena. 2) Aurreko berbera baina analisi kopurua murriztu gabe. 3) Proposamenen analisi aukera guztiak kontuan hartzen direnekoa baina analisi kopuru mugatuarekin. 4) Proposatutako metodoa batere murriztapenik gabe zuzentzen duena. Lehena azkarrena den bitartean, laugarrenean ziurtatzen da bateko edizio-distantzian dauden errore guztien zuzenketa agertuko dela[1] proposamenen artean. Bigarrena eta hirugarrena tarteko ebazpideak dira. Kontuan hartu behar da guztietan amankomunean dagoena: akats ohizkoenen bilaketa, proposamen hipotetiko guztien bilaketa maiztasun handieneko hitz zilegien bufferrean, eta gaitasun-erroreen tratamendua.

B) Corpus bakoitzeko eta aurretik aipatutako zuzenketa-metodo bakoitzeko lau neurri ematen dira: azkena hitz bakoitzari dagozkion proposamenak sortzeko batez-besteko denbora den bitartean, lehenengo hirurak zehaztasunarenak dira, erroreari dagokion zuzenketa zenbatetan proposatzen den *(n)*, zenbatetan proposatzen den lehen hiruren artean *(3)* eta zenbatetan lehena *(1)*.

[1] Bi edo edizio-distantzia handiagoko erroreen zuzenketa proposamen gisa agertuko da, baldin eta aldaera bezala ezagutzen bada.

Testuak		**Xuxen** (mugatua)	**Xuxen** (muga gabe)	**morfema guztiak** (mugatua)	**guztiak**
1.-Ikasleen testuak	(n)	%82	%87	%81	%89
	(3)	%81	%85	%80	%86
	(1)	%74	%76	%73	%75
	denb.	0,3 s	6,2 s	0,6 s	15 s
2.-Testu teknikoa	(n)	%63	%73	%64	%88
	(3)	%62	%72	%63	%86
	(1)	%49	%56	%50	%68
	denb.	0,4 s	2,7 s	0,5 s	12,5 s
3.-Prentsako testuak	(n)	%70	%80	%71	%89
	(3)	%68	%78	%69	%85
	(1)	%59	%64	%60	%71
	denb.	0,35 s	4,5 s	0,6 s	16,7 s
GUZTIRA	(n)	%72	%80	%72	%89
(300 akats)	(3)	%**70**	%**78**	%**71**	%**86**
	(1)	%61	%65	%61	%71
	denb.	**0,35** s	**4,5** s	**0,6** s	**14,7** s

VI.9 irudia.- Zuzenketaren zehaztasuna eta abiadurari buruz hartutako estatistikak.

Irudian azaltzen diren datuak aztertuz ondoko ondorioak aipa daitezke:

- Denborak: Kontuan hartu behar da emandako denborak batez-bestekoak direla, baina proposamen kopurua mugatzen duten bi metodoetan desbiderapen handia dagoenez gero, kasu batzuetan proposamenak sortzeko denbora luzeagoa izan daiteke.

- Zehaztasuna: Egiaztapen oso guztiekin hiru corpusekin emaitza antzekoak lortzen badira ere, gainontzekoetan askoz emaitza hobeak lortzen dira aldaera edo gaitasun-errore asko duten testuetan (ikasleenak) besteetan baino. Horrekin baieztatzen da aurretik esan dugun zerbait: aldaeren tratamendua osoagoa da errore tipografikoena baino. Egiaztapen guztiekin zehaztasunak muga bat du %90ean, eta hori bateko edizio-distantzian dauden ordezkagaiak bakarrik aztertzetik dator nagusiki —distantzia handiagoan dauden batzuk zuzentzen dira aldaeren tratamenduari esker—.

- Metodoen arteko desberdintasunak: Xuxen zuzentzaile ortografiko komertzialerako erabilitako metodoak (lehen zutabekoa) nahikoa oreka ona lortzen du zehaztasuna eta eraginkortasunaren artean, batez ere gaitasun-errore

asko dagoenean. Hala ere, eta egiaztatze morfologikoa azkartu ahala, bigarren eta laugarren zutabeei dagozkien metodoetara jo beharko da, hirugarrenekoan zehaztasuna apenas hobetzen ez baita.

Flexio handiko hizkuntzetan aplikatzen diren beste sistemetarako ematen diren neurriekin konparatuz gero, nahikoa emaitza onak lortzen direla esan daiteke, gaitasun-erroreen tratamendua horretan garrantzi handia izanik.

VI.8. Proposatutako hobekuntzak.

Egindako sistemaren gainean aztertzen ari garen hobekuntzak azaltzen dira kapituluaren azken atal honetan. Hobekuntza hauek, normala denez, bi bidetatik joan behar dute, abiadura eta zehaztasunaren aldetik, alegia.

Abiadura da zuzentzailearen alde ahulena beste hizkuntzetarako merkatuan dauden zuzentzaileekin alderatzen badugu. Hobekuntza horretarako funtsezkoa da analisi morfologikoaren denborak laburtzea, horretan aztertutako lexiko-itzultzaileek markatutako bidea jorratu behar delarik. Abiadura hobetzea zehaztasunaren onerako izango da, oraingo metodoarekin egiten diren mozketak, VI.6.1 atalean azaldu direnak, bazter daitezkeelako.

Zehaztasunaren aldetik zuzentzaile zehatza bada ere, hauek dira puntu ahulenak eta hobetu behar direnak:

- Bateko distantzia baino gehiago duten errore tipografikoen tratamendua. Dagoen zuzentzailearekin eginez gero, abiaduran oso eragin handia jasango genuke.
- Testuingurua kontuan hartzea, proposamenak sailkatzea, eta benetako hitzaren erroreen detekzioa hobetu.

Hobekuntzarako bide horiek bi urratsetan laburtuko ditugu: proposamen-sistema erro-hizkiaren bidez burutzea batetik, eta lexiko-itzultzaileak erabiltzea bestetik.

VI.8.1. Lexiko-itzultzaileen erabilera.

Lexiko-itzultzaileen ezaugarri nagusietako bat analisi morfologikorako eskaintzen duten abiadura dugu. Beraz, III.6 atalean azaldutako euskara estandarrerako prozesadore morfologikoa egiaztapen morfologikorako erabiliz, eta IV.2.4 atalean azaldutakoa aldaeren tratamenduari aplikatuz emandako denborak ehun bat aldiz azkar litezke, ondorioz zehaztasuna laburtzen duten mugak kentzeko aukera emanez, eta morfologikoki sinpleago diren beste hizkuntzen zuzentzaileekin parekatuz.

Erabiltzailearen hiztegiarekin dira arazo bakarrak lexiko-itzultzaileak zuzenketan aplikatzeko garaian. Honen integrazioa Carter-ek (1995) adierazitako bidetik etor liteke, lexiko-itzultzaileetan egiten den lexiko osoaren konpilazioaren ordez lema irekiak ez direnak soilik konpilatuz.

VI.8.2. Erro-hizkiaren bidezko proposamen-sistema.

Erro-hizkian oinarritutako metodoa honetan datza:

- Hitz baten erro-hizkien banaketa posibleak lortu. Hau da egitekorik zailena.
- Alde bakoitzaren zuzenketa posibleak sortu, horretarako aurreko kapituluan azaldutako mekanismoak erabil daitezkeela.
- Sorkuntza morfologikoaren bidez proposamenak eskuratu. Lan honetan morfologia nahiz morfotaktika eduki behar da kontuan, morfotaktikaren aldetik jatorrizko zatien kateatzea zilegia bada ere, zati horietatik sortutako zuzenketa-zatiak elkartezinak izan baitaitezke morfotaktikaren aldetik.

Lehen urratsari dagokionean zerbait egin da aurretik. Batetik, lema hipotetikoak gordetzen dira analisia egin ahala, VI.3.1 atalean aipatutako morfemen selekzioaren bidez. Bestetik, laugarren kapituluan azaldutako lexikorik gabeko analisiari esker lema ezezagun bati dagozkion atzizki-multzo posibleak lor daitezke.

Arazo nagusia ondokoa da: akatsek erroari eta hizkiren bati batera eragiten badiete —biko edizio-distantziaz edo mugako bi karaktere jarrairen arteko trukeaz— lehen urratsa egitea oso konplexu bihurtzen da.

Aurreko eragozpenen aurrean, hobekuntza hau irekitako ikerlerro gisa utzi dugu.