BIGARREN PARTEA: ZUZENKETA ORTOGRAFIKOA

V. Erroreen zuzenketa.

Euskararako zuzentzaile ortografiko bat burutzeko ideia taldearen helburu nagusien artean zegoen hasiera hasieratik. Horri ekiteko arrazoi nagusia, premia zen, ez baitzegoen halako produkturik euskararako, baina, baita ere, martxan dagoen batasun-prozesuak are interesgarriago bihurtzen zuelako halako tresna bat.

Berrerabilgarritasun eta zehaztasun irizpideak hobetsiz, horrek eraginkortasunaren aldetiko galera eragin bazezakeen ere, bi ideia nagusitu ziren laster: egiaztatzaile/zuzentzailea morfologian oinarritu behar zela batetik, eta martxan den batasun-prozesu irekiak eragiten dituen akatsei aurre egiteak lehentasuna zuela bestetik.

Izan ere, ideia horiek ondo finkatzeko eta arazoei aurre egiteko, gai honen inguruko kontzeptu eta ideia nagusiak eztabaidatzen eta argitzen joan ginen, bibliografia lagun genuela —azpimarratzekoa da Kukich-ek *ACM Computing Surveys*-erako (1992) idatzitako bilketa, gaur egun arlo honetan sartzeko oso gomendagarria dena.

Ideia horiek kapitulu honetan biltzen dira, euskararako egin dugun gauzatze konkretua hurrengorako utziz. Zuzenketaren arlo honetan aurrerapen eta proposamen asko dagoenez, lehenengoz mugatu dugu gure lanaren helburua, hitz isolatua aztergai hartuz, eta testuingurua kontuan hartzen duen zuzenketa ondorengo proiektuetarako utziz. Muga hori ezarri ondoren, hartutako esparruan kokatzen diren teknikak gainbegiratzen dira: hitzen ezagutza batetik, eta ez-onartuei dagozkien ordezkapen/proposamenei buruzkoak bestetik.

Zuzenketari buruz asko ikertu arren, helburu-hizkuntza eranskaria edo flexio-aukera handikoa denean, problematika konplexuagoa izanda ere, erreferentzia bakan batzuk baino ez dira aurkitzen.

Kapitulu hau lau ataletan banatu da: aplikazioak, hitzen egiaztapena, hitzen zuzenketa eta flexio handiko zein hizkuntza eranskarietan aurkezten diren arazo bereziak.

V.1. Aplikazioak, sailkapena eta irizpideak.

Testuen zuzenketa aplikazio desberdinetarako garatzen ari den ikerkuntza-arlo irekia da. Erabilpen ezagunenak testuen edizioaren eta karaktereen ezagutza optikoaren (OCR) esparruetan kokatzen badira ere, pertsona-ordenadore interfaze sistema orok, komando-lengoaia erabiltzen dutenak barne, hobetzen dira halako teknikak erabiliz. Horien artean honako hauek azpimarra daitezke: datu-baseen gaineko biltegitze/berreskuratze interfazeak, lengoaia naturalezko interfazeak, OLI sistemen interfazeak —eta baita beraien ezagutza-basea ere OLIren helburua idazketa denean—, testu-hizketa hizketa-testu bihurketak, eta elbarritu edo behar berezietako pertsonentzako komunikazio-sistemak.

Aplikazioaren arabera betebehar eta ezaugarri desberdinak egon arren, aplikaturiko estrategiaren arabera bi multzo handi bereizten dira:

- Elkarrekintzazko zuzenketa: erabiltzailearen esku utzi ohi da azken erabakia erroreak ordezteko orduan —gutxienez programak zalantza-tarte handia duenean—. Aplikazio tipikoa zuzentzaile ortografikoa da. Zuzenketarako proposamen bat baino gehiago lor daiteke emaitza gisa, eta hauen artean probabilitate-sailkapen bat egiten da. Proposamenik ez eskaintzea ez da oso egokia baina onar daiteke salbuespen gisa. Testuingurua kontuan hartzea ez da ezinbestekoa baina bai lagungarria, beste moduz detektaezinak diren akatsak aurki baitaitezke eta proposamenak zehatzago ordena baitaitezke. Testuingurua kontuan hartzen bada, zuzentzaile hauen zehaztasuna igotzeaz gain, sintaxia eta estiloa kontuan hartzen duten zuzentzaile aurreratuak egin daitezke.
- Zuzenketa automatikoa: zuzenketarako proposamen bakar bat lortu behar da, giza-laguntzarik ez dago eta. Testuingurua kontutan hartzea ezinbestekoa da emaitza dotoreak lortzeko. Denbora errealeko hizketa-ezagutza bezalako aplikazioetan behar da.

Ohizkoa da, Kukich-ek aipaturiko artikuluan (Kukich, 92) horrela egiten duen bezala, gai honi buruzko ikerkuntza hiru ataletan banatzea[1]:

- hitz ezezagunen detekzioa edo testu-egiaztatzea
- testuingururik gabeko hitz-zuzenketa
- testuinguruaren araberako hitz-zuzenketa

Lehen atalean ez dira sartzen Mitton-ek (1987) *real-word errors* deitutakoak —*benetako hitzaren erroreak* deituko ditugunak—, hitz ezaguna sortzen duten errore hauek detektatzeko testuingurua aztertu behar delako, hori dela eta hirugarren atalari dagozkion tekniken bidez tratatu behar direlarik.

Hirugarren atalari dagozkion teknikak txosten honek jasotzen duen proiektutik at daudenez, taldearentzat irekitako ikerlerro bat bada ere (Agirre *et al.*, 94) (Gojenola & Sarasola, 94), teknika-multzo horretaz ez gara ariko lan honetan zehar. Bibliografia gisa, Kukich-en aipatutako artikuluaren hirugarren kapitulua eta Vosse-rena (1992) aipa daitezke.

V.2.Egiaztatzea.

Zuzenketa bideratzeko ezinbesteko lehen urratsa hitzen arteko bereizketa edo egiaztatzea da, hau da, testuan zehar agertzen diren hitzen artean aukeratzea zeintzuk diren zilegiak edo ezagutuak eta zeintzuk ez.

Prozesu honetan egiaztatzailearen helburua ahalik eta zehaztasun handienaz lan egitea da. Zehaztasun-tasa jaisten duten bi akats-mota gertatu ohi dira egiaztatze-prozesuan (Peterson, 80):

- Erroretzat hartzen diren hitz zilegiak. Muga batetik behera mantentzen diren bitartean behintzat, elkarrekintzazko aplikazioetan oso kezkagarriak ez diren bitartean, erabiltzaileak ontzat emango baititu, zuzenketa automatikoan erabat ekiditera jo behar da, erroreen kopurua handitzen dute eta.
- Zilegitzat hartzen diren erroreak. Bi multzotan bereizten dira, hizkuntzan existitzen ez diren hitzak direnak batetik, eta, bestetik, *benetako hitzaren erroreak* deitu ditugunak, hau da, errorearen ondorioz testuinguru horretan ez baina hizkuntzan zilegia den hitza sortzen dutenak. Esan den bezala azken hauek lan honetatik kanpo utziko ditugu.

[1] Kukich-en terminologiaren arabera *nonword error detection*, *isolated-word error correction* eta *context-dependent word correction*.

Egiaztapena burutzeko metodoak sailkatzeko orduan irizpide desberdinak erabiltzen dira. Askotan metodo heterogenoak erabiltzen diren arren, bilketa honetan hiru multzotan banatu ditugu metodo hauek, erabiltzen dituzten datuen arabera: hitz-zerrenden bidezkoak, hitz-zatien bidezkoak eta morfologian oinarritutakoak.

Bakoitza bere aldetik aztertuko dugu, dagozkien abantailak eta eragozpenak azpimarratuz eta erreferentziaren bat aipatuz[1].

V.2.1 Hitz-zerrendetan oinarritutako metodoak

Sistema hauetan hitza onartu egiten da baldin eta hitz-zerrendan agertzen bada. Zehaztasun minimo bat lortzeko behinik behin, hitz-zerrenda oso luzea izaten da: hizkuntzaren arabera aldatzen da, baina 50.000 edo 100.000 hitzetik gorakoa izan ohi da. Datu-kopuru horiekin lan egiteko datuak gordetzeko eta atzitzeko bideak sakonean aztertu behar direlarik.

Gainera, halako hiztegi erraldoi bat eraikitzeko hitz askoko corpus zuzen bat (errorerik gabekoa) behar da iturburu gisa, miloi bat hitzetatik gorakoa gomendatzen da. Horrela egiten ez denean zehaztasunaren kalterako izaten da eta honako ondorio hauek izaten ditu: corpus txikien bidez hitz zilegi batzuk erroretzat hartuko dira, eta zuzendu gabeko corpusen bidez, berriz, errore batzuk ontzat hartzeko arrisku dago. Azken hau oso kaltegarria litzateke euskara bezala ondo batu gabeko hizkuntzetan.

Metodo horien ezaugarriak hauek izaten dira orokorrean:

- Flexio aberatsa duten hizkuntzetarako desegokia, gordetzeko zerrenda izugarri luzatzen delako, eta ondorioz “ikasteko” corpusak askoz handiagoa izan behar duelako. Gainera, sistema ez da malgua izango jakintza-arlo berri edo termino berrientzat, erro berri bakoitzeko bere deklinabide guztia sartu beharko litzatekeelako.
- Flexio txikiko hizkuntzetarako aurreko arazoak saihets badaitezke ere, beti iraungo du batek, koherentziarenak alegia. Erabiltzaileari oso ulergaitz egiten zaio erro baten flexio batzuk onartzen dituen bitartean beste batzuk errefusatzen direla ikustea, azken hauek probabilitate txikiagokoak izan arren.
- Garatzeko azkarrak eta merkeak.

[1] Kasu praktikoen adibide asko eman litezke, oso bibliografia oparoa dago eta. Hala ere, nahiago izan dugu ezaugarrien arabera gauzatze garrantzitsuenak azpimarratzea erreferentzia gehiegi ematea baino. Kukich-en aipaturiko artikuluan bibliografia osoa aurki daiteke.

- Ondoren aipatuko diren teknikak erabiliz, abiaduraren aldetik azkarrak izaten dira eta arazorik ez dute biltegi-tokiaren aldetik.

- Zehaztasuna, iturburu-corpusaren eta egiaztatzeko testuaren araberakoa izango da, baina aipatutako corpusa zuzena bazen ziurta daiteke ez dela ontzat emango hizkuntzan existitzen ez den hitzik. Dena dela testua erreferentzia-corpusarekin bat ez badator —jakintza-arlo bereziak, mailegu berriak, etab. direla eta— hitz zilegi batzuk ez dira ezagutuko.

Metodo honen bidez hitz-zerrenda edo hiztegia oso luze izan daitekeenez, hiztegia gordetzeko teknika desberdinak ikertu dira zehaztasunaren, toki-hartzearen eta atzipen-denboraren artean ahalik eta orekarik handiena lortzeko.

Teknika horietako batzuk aztertuko ditugu ondoren:

- Mailakako biltegitzea (Peterson, 80) memoria-hierarkien ideiari jarraituz hiru maila bereizten dira: lehena txiki samarra eta azkarra, maiztasun handieneko hitzak azkar atzitzeko —testuetako hitzen %50a hartzen duten hitzak kokatu ohi dira bertan—; bigarrena, dokumentuan bertan lehenago azaldu diren hitzak tarteko abiaduraz atzitzeko, eta hirugarrena, masa-biltegia, atzipen-abiadura motelenarekin. Hizkuntzaren arabera alda badaiteke ere, lor daiteke azken maila hori atzitu behar izatea %10ean baino gutxiagotan.

- *Hashing*-teknikak erabiltzea, taula batzuk eraikiz hiztegiaren atzipena azkartzeko. Gakoa kolisio gutxi sortzen duen *hash*-formula egokia aurkitzea da. Sistema batzuetan, Unix-eko *spell*-en (McIlroy, 82) adibidez, hitza gorde beharrean bere agerpena adierazten duen bit bat baino ez dute gordetzen. Horrek trinkotzen du hiztegia, baina, horren truke, hitz zilegiekin kolisioa sortzen duten akatsak onartzera darama.

- Bilaketa azkarra bideratzen duten zuhaitz-egiturak eta bigarren kapituluan aztertu den *trie egitura*.

Hitz-zerrendetan oinarritutako metodo hau izan da erabiliena, orain dela gutxi arte eta flexio-sistema sinplea duten hizkuntzetarako behintzat, zuzentzaile ortografikoen arloan; joera aldatzen ari da, ordea, eta morfologiaren bidezko tratamenduak —sarritan sinplifikatuak— ugalduz doaz.

V.2.2 Hitz-zatietan oinarritutako metodoak

Hitzak ezagutu gabe akatsak detektatu nahi direnean edo hitz-zerrenda luzeegiak gertatzen direnean aplikatzen dira aurrekoen aldaeratzat har daitezkeen teknika hauek.

Corpus batean oinarriturik ere, hitz-zati horiei buruzko informazioa lortu eta gordetzen da. Corpusak, normalean, ez du aurreko teknika-multzokoa bezain handia izan behar, baina zuzen-zuzena izatea horietan bezain komenigarri edo gehiago da.

Zatiak bi motakoak izan daitezke:

- *n-grama*k: *n* karaktereko luzera duten hitz-zatiak. *n* handitu ahala zehaztasun handiagoa lortzen da, baina toki-hartzea ere handiagoa da. n-gramen posizioa kontuan har daiteke zehaztasuna hobetzeko toki-hartze handiagoaren truke (Hull & Srihari, 82). Trigramak dira n-grama erabilienak, zehaztasun eta toki-hartzearen artean oreka onena lortzen delakoan (Zamora *et al*., 81).

- Luzera aldakorreko zatiak. Trataera eta burutzapen konplexuagoa eskatzen du, sasi-morfemak bilatzea baita helburua. Horretarako aipatutako corpusa konpilatu behar da, hitz-zatien artean loturak inferitzeko, adibidez egoera finituko automata bat sortuz (Aho, 90) (Meddeb, 94).

Hitz-zatietan oinarritutako metodo horien ezaugarriak hauek izaten dira orokorrean:

- Zehaztasunaren aldetik du arazo nagusia, bi motako akatsak egiten direlako: existitzen ez den hitzak zilegitzat hartu —hau da arazo nagusia— eta zilegi diren hitzak erroretzat. Hala ere, azken multzo honi dagokionez, corpusetan agertzen ez ziren forma zilegiak onartzeko gaitasuna du, hitz-zatien arabera zilegiak badira.

- Malgutasun falta, ia ezinezkoa baita sistema aberastea zehaztasuna hobetzeko.

- Garatzeko azkarrak eta merkeak, n-grametan oinarrituak behintzat.

- Azkarrak izaten dira eta, biltegi-tokiaren aldetik, oso trinkoak.

- Koherentziaren arazoa. Aurreko teknika-multzoan bezala, baina probabilitate txikiagoarekin, flexio batzuk onartzen diren bitartean beste batzuk errefusaturik izan daitezke.

n-grametan oinarrituriko sistemak OCR motako aplikazioetan erabili ohi dira, honako arrazoi hauengatik: sortzen diren erroreak nahikoa espezifikoak dira n-grama arraro samarrak sortuz, eta hizkuntzaren ohizko n-gramekin bat datozen hitz “berriak” onartzen direlako ezer eguneratzen ibili gabe.

V.2.3 Morfologian oinarritutako metodoak

Hitz bat zilegia den ala ez erabakitzea, hitzak deskonposaketa morfologiko zilegia duenentz aztertzea da; hau da, hitz bat ontzat ematen da deskonposaketa morfologikorik baldin badu. Beraz, analizatzaile morfologiko baten bidez bidera badaiteke ere, lan honetarako ez da analizatzaile osoa behar, deskonposaketa morfologiko posiblerik duenentz zehaztea nahikoa delako.

Metodo hauek hitz-zerrendetan oinarritutakoen alternatiba dira, hizkuntzaren flexioa aberatsa denean eta baita hitz-zerrenda laburtu nahi denean ere. Hajic-ek eta Droza-k (1990) sarreran diotena aldatuko dugu hona:

> " ... From different reasons, among which the speed of processing prevails, they are usually based on dictionaries of word forms instead of words. This approach is sufficient for languages with little inflection such as English, but fails for highly inflective languages such as Czech, Russian, Slovak or other Slavonic languages. ..."

Aurreko sistemen bilakaera "logikotzat" har daiteke teknika-multzo hau. Morfemak lirateke luzera aldakorreko hitz-zatiak, eta corpusetik lortzen den informazioa morfologiari buruzko ezagutza.

Morfologian oinarritutako metodoen ezaugarriak hauexek dira:

- Garatzeko garestiak dira denbora eta kostuaren aldetik. Horren truke berrerabilgarritasuna dugu, egindako lana helburu anitzekoa baita.
- Koherentziaren eta malgutasunaren aldetik sistema onenak dira. Erro berri bat behin sartuz gero bere forma flexionatu guztiak, eta kasu batzuetan forma eratorriak eta elkartuak ere, ezagutzen dira; ondorioz, sistemaren aberasketa erraztu egiten da.
- Biltegi-tokiaren aldetik aurreko bi teknika-multzoen artean dago. Abiaduraren aldetik, berriz, formalismo morfologikoaren menpe izanda ere, besteak baino motelago izan ohi da. Azkartzeko maiztasun handieneko hitzen zerrenda batekin konbinatu ohi da.
- Zehaztasunaren aldetik inoiz ez da ontzat emango hizkuntzan existitzen ez den hitzik, morfologiaren bidez gainsorkuntza onartzen ez bada behintzat. Ezagutzen ez diren hitz zilegien kopurua lexikoaren araberakoa izango da, baina, esan den bezala, aberasketa erraza eta koherentea bideratzen duenez, erabiltzailearen esku utz daiteke aberasketa hori. Hori dela eta, idazlearen estiloari eta jakintza-arloari ondo egokitutako erabiltzailearen lexiko baten bidez oso zehaztasun handia lor daiteke.

- Lortutako informazio morfologiko partziala interesgarria izan daiteke zuzenketa garaian. Inguruko hitzen informazio morfologikoak testuingurua kontuan hartzen duen zuzenketa bidera dezake.

Flexio aberatsa duten hizkuntzetarako ezinbestekotzat jo daitekeen bitartean, gainontzeko hizkuntzetarako gero eta erabiliagoak dira, elkarrekintzazko aplikazioetan batez ere: (Means, 88), (Hajic & Droza, 90), (Solack & Oflazer, 93), (Aduriz *et al*., 93), (Oflazer & Guzey, 94), (Vagelatos *et al*. 95).

V.3.Zuzenketa.

Ezagutzen ez diren hitzak zein diren jakin eta gero —aplikazioaren arabera hitz susmagarriak edo erroreak deitzen zaie— forma horien kudeaketa dator. Kudeaketa hori automatikoa izan daiteke, eta, orduan, zuzenketa automatikoa deitzen zaio, edo erabiltzailearen laguntzaren bidezkoa, kasu honetan proposamenen sorkuntza eta sailkapena izena egokiago delarik. Gure helburua bigarren kudeaketa-mota hori bada ere, bi multzoetan erabiltzen diren teknikak nagusiki amankomunak dira: proposamenak egiteko erroreen tipologia eta ezaugarriak aztertu behar direlako, eta, hautapen bakarra edo sailkapena egiteko, hurbilpen- edo antzekotasun-neurriak aztertu behar dira irizpideak finkatzeko.

V.3.1 Errore-motak eta ezaugarriak.

Erroreak zuzentzeko, edo dagozkien proposamenak lortzeko, erroreen ezaugarriak aztertzea interesgarria da oso. Erroreei buruz sailkapen desberdinak egin badaitezke ere, honako hiru multzotan sailkatzea proposatzen dugu argigarria delakoan:

- Oinarrizko erroreak: aplikazio guztietan agertzen dira mota honetako erroreak, jatorriak desberdinak izan arren. Askotan errore tipografikoak deitzen dira.
- Aplikazioarekin lotutako berezitasunak: testua lortzeko bidearekin lotura zuzena duten erroreen ezaugarriak kokatzen dira honetan.
- Hizkuntzaren ezaugarriekin lotutako erroreak, hizkuntzaren zenbait ezaugarri ez ezagutzeak, nahasteak edo aldaketa dialektalak eraginda, gizakiek modu kontzientean egindako erroreak dira beti. Aldaerak edo gaitasun-erroreak deituko ditugu.

V.3.1.1 Oinarrizko erroreen sailkapena.

Damerau-ren (1964) tipifikazioa hartzen da oinarritzat garatutako zuzenketa-aplikazio ia guztietan. Sailkapen honen arabera errore gehienak —testu-edizioan %80a da Damerauk ematen duen neurria— hauetako lau gertakizunetako bakar baten eraginez sortzen dira:

- Karaktere baten **aldaketa**. Karaktereetako bat ez da jatorrizkoa, bere ordez beste bat kokatu delako. Adib. *kame kale*-ren ordez
- Karaktere baten **sorrera**. Karaktere bat gehiago dago jatorrizko forman baino, bertako biren artean edo mutur batean txertatu da eta. Adib. *ssistema sistema*-ren ordez
- Karaktere baten **desagerpena**. Karaktereetako bat desagertu da, jatorrizko hitza karaktere batean laburtuz.Adib. *ed edo*-ren ordez
- Bi karaktere jarrairen arteko **trukea**. Bi karaktere jarrairen artean ordena aldatu egin da, hitzaren luzera mantenduz baina bi posizioetako karaktereak aldatuz. Adib. *bania baina*-ren ordez. Hauek ez dira orokorrak teklatua erabiltzen den aplikazioetan bakarrik agertzen baitira; beraz, gainontzeko aplikazioetan ez da kontuan hartuko.

Hitz batetik sor daitezkeen errore bakunak —hitz osoan aipatutako erroreetako bakar bat duten formak— kuantifikatu egin dira, eta *n* luzera duen hitz baterako hauexek dira kalkuluak (hizkuntzaren alfabetoko karaktere-kopurua *k* izanik): *n(k-1)* aldaketa, *(n-1)k* sorrera, *n* desagerpen eta *(n-1)* truke. Beraz, konbinazio kopurua *2nk*-tik oso gertu dago. Hala ere, hauetako konbinazio asko eta asko bazter daitezke, hasieratik metodo estatistikoak (n-gramen analisien bidez adibidez) erabiliz (Pollock & Zamora, 84).

Bakunak ez diren erroreak gertakizun hauen konbinazioaren bidez adieraz daitezke, eta **errore anitzeko akatsak** —*multi-error misspelling*— deitu ohi zaie. Hauen kopuruaz oso datu kontrajarriak daude: Pollock-ek eta Zamorak %6a aipatzen duten bitartean, Mitton-en (1987) ustez %31raino iristen dira. Badirudi datuen eta aplikazioaren arabera oso aldakor izan daitekeela.

Hitzen luzera eta erroreen arteko eraginaz zenbait zehaztasunen berri ematen da. Errore gehienak bakunak direnez, zera inferi daiteke: erroredun hitzaren eta jatorrizkoaren arteko luzera-diferentzia bat edo gutxiago dela. Hori dela eta, hitz-zerrendetan oinarritutako zuzentzaileak luzeraren arabera antolatzen dute hitz-zerrenda askotan. Hitz luzeetan motzetan baino errore gehiago egiten ote den ez dago argi, baina gaizki zuzendutako hitzak motzak izaten dira askotan.

Gertakizun bakunen bidez gerta daitezke aipatu behar diren bi fenomeno: benetako hitzaren erroreak eta hitz-mugaren gaineko erroreak.

Benetako hitzaren erroreak, aurretik esan den bezala gure lan-eremutik at daude, baina beren kuantifikazioa interesgarria da —ideia ematen baitigu testuingururik gabeko tratamenduen mugaz—. Honetaz ere neurriak ez datoz bat; horrela eta beti ingeleserako hartutako neurriez, Peterson-en (1986) ustez behe-muga %16a den bitartean Mitton-ek (1987) %40a aipatzen du. Kontuan hartzekoa da bi esperimentuen arteko desberdintasuna zuzentzaile arrunten erabileraren eragina izan daitekeela, hein batean behintzat. Alegia, zuzentzaile arrunten erabilerak errore-kopurua laburtzen du benetako hitzarenak kenduta, ondorioz azken hauen portzentaia igoz.

Aipatutako datuak ingeleserako datuak dira, eta beste hizkuntzetarako ez da halako daturik aurkitzen. Euskara bezalako hizkuntza eranskarietan portzentaia hori txikiagoa izatea espero daiteke hitzak luzeagoak izan ohi direlako eta zera baitago frogatuta: benetako hitzaren errore bat sortzeko probabilitatea txikiagoa dela hitz luzeetan, motzetan baino. Gainera, zuzentzaile ortografikoen erabilera murritzagoa dela eta bestelako erroreak ez dira gutxiagotzen.

Hitz-mugaren gaineko erroreak askotan ez dira kontuan hartzen, beren tratamendu konplexuagoa dela eta, *token* bat baino gehiago kontuan hartu behar baita. Hala ere, oinarrizko errore bezala ikus daiteke zuriunea[1], karaktere arruntzat jotzen baldin bada. Bi multzotan bana daitezke errore hauek: zuriunearen galerarengatik bi hitz bakar batean biltzekoak (*run-on words*), eta zuriuneren baten agerpenarengatik hitz bat bitan zatitzekoak (*split words*). Kukich-en ustez (1992), detektaturiko erroreen %15a mota honetakoak dira (%13 eta %2 hurrenez hurren). Tratamendu egokirik gabe hauei dagozkien zuzenketak edo proposamenak desegokiak izango dira. Mota honetako erroreek hitz ezagun bat noiz sortzen duen ere azterturik dauka Mitton-ek.

Errore hauen tratamenduari dagokionez, berriz, tratatzen dituztenen artean bi multzo bereiz daitezke: tratamendu berezitu partikularra ematen dutenak (Pollock &Zamora, 84) (Kernighan, 91) batetik, eta *lattice*[2] izeneko sare batez teilakatze-aukera guztiak aztertzen dituztenak (Carter, 92).

---

1 Zuriunea hitz-mugaren sinonimotzat hartuko dugu.

2 Vosse-k (1992) *lattice* egitura bera proposatzen du lokuzioen eta hitz anitzeko terminoen tratamendurako.

V.3.1.2 Aplikazioarekin lotutako erroreak.

Aurreko atalean azaldutako baieztapen edo neurri batzuk berraztertu egin behar dira aplikazioaren eta testu-iturriaren arabera, zeren desberdinak baitira pertsona batek teklak sakatzean sortzen dituen erroreak, OCR unitate batek sortzen dituenak edo mikrofono eta hizketa-testu sistema batean sortzen direnak. Kasuaren arabera, aurretik ikusitako zenbait errore maiztasun handiagoz edo gutxiagoz gertatuko dira, edo kasuistika bereziak sortuko dira. Horren aurrean teknika berriak edo teknika orokorren egokitzapenak izango dira gomendagarri. Ongien aztertutako aplikazioak OCR motakoak eta testu-edizioa direnez bi horietan zentratuko gara.

OCR bidezko irakurketetan honako ezaugarri hauek detektatu dira:

- Gertatzen diren erroreen ondorioz n-grama arraroak sortzen dira askotan, beste aplikazioetan baino gehiagotan. Horregatik erabiltzen da, besteak beste, n-grametan oinarritutako metodoa erroreen detekziorako.
- Erroreen iturburu nagusia karaktereen arteko antzekotasuna izaten denez, errore gehienak aldaketa baten bidez gertatzen direla suposa daiteke, aldaketen probabilitatea antzekotasunaren arabera defini daitekeela. Sistema batzuetan antzekotasun hori definitzeko irakurritako testuaren letra-mota hartzen da kontuan.
- Aurreko puntuan esandakoaren arabera, erroredun hitzetan jatorrizko luzera mantentzen dela esan badaiteke ere, aldaketa batzuek ondorioak dituzte luzeraren gainean. Horrela *ri =n* edo *m=iii* aldaketak sarri gertatzen dira. Kasu berezi horiek behintzat aztertu ohi dira, beti luzera mantentzen delako erregela gaindituz.
- Hitz-mugaren gaineko erroreei dagokionez, berriz, aipatutako bi motetakoen arteko banaketa alderantzizkoa da testu-edizioarekin konparatuz, hau da, hitz-zatiketa gehiago gertatzen da biren bilketa baino.

Testu-edizioari dagozkion ezaugarriak hauek dira:

- Teklatuan karaktereek duten posizioaren arabera karaktereen arteko distantzia fisikoa eta antzekotasun-neurria lot daitezke. Irizpide hau, interesgarria badirudi ere, sistema gutxitan erabiltzen da.
- Yannakoudakis-en (1983) iritziz errore gehiago gertatzen dira hitzaren azken karaktereetan hasierakoetan baino. Lehen karakterean errore gutxi egon ohi delakoan, zerrendetan oinarritutako zuzentzaile ortografiko askok hiztegia lehen

karakterearen arabera antolatzen dute; honek, kasu batzuetan, zuzenketa zilegia ez aurkitzera eramaten ditu.

V.3.1.3 Gaitasun-erroreak.

Idazlearen arabera izen desberdinak esleitzen zaizkie hizkuntzaren ezaugarriekin lotutako erroreei: *orthographical errors* (van Berkel & de Smedt, 88); *competence errors* (Veronis, 88), *cognitive errors* eta *phonetic errors* (Kukich, 92). Gaitasun-erroreak eta, morfologian (ikus IV kapitulua) ikusitakoarekin bat etorriz, aldaerak deituko ditugu.

Hizkuntzaren ezaugarriekin lotutako errore hauek bereziki tratatzea ez da ohizkoa, baina, egiten denean, emaitza onak lortzen dira. Errore fonetikoak izaten dira tratamendu berezia merezi ohi dutenak. Mitton-en arabera aztertutako corpusean aurkitzen diren erroreen artean %44ak homofonoekin du zerikusirik. Hala ere, bibliografian agertzen diren aplikazioak, leku-izenekin eta pertsona-izenekin lotuak dira. Mota honetako bi adibide ditugu: orri horiak Minitelaren bidez frantsesez kontsultatzeko Veronis-ek (1988) garatutako zuzentzailea, batetik, eta bestetik Van Berkel eta De Smedt-ek holanderazko pertsona-izenen gainean egiten dituzten neurriak, beren Triphone sistemarako. Argi dirudi halako eremuetan handiago dela fonologiaren eragina erroreetan.

Sistema batzuetan bestelako erroreekin batera tratatzen dira errore hauek, maiztasun handia duten erroreak eta dagozkien zuzenketak buffer batean gordez.

Normalean silaba/fonemen bidez eta ezagumendu linguistikoa erabiliz tratatzen dira, eta errore tipografikoen trataerarekin konbinatzen diren azpisistemak izan ohi dira. Honetaz V.4 atalean sakonduko badugu ere, zenbait sistemaren oinarriak zerrendatuko ditugu hemen:

- Aipatutako Triphone-n homofonoak bilatzeko fonemen araberako lexiko bat erabiltzen da.
- TWBn (Kese *et al*., 92) alemanerarako erabilitako zuzenketan lexiko berezi bat eratzen da errorea, testuingurua, dagokion zuzenketa eta arrazoiaren azalpena edukitzeko.
- Fonemen arteko baliokidetasun-taulen eraketa erabili da frantsesezko interfaze-sisteman (Veronis, 88), eta greziera modernorako zuzentzaile batean (Vagelatos *et al*., 95).

Gure proiektuan, eta hurrengo kapituluan sakonduko dena laburbilduz, laugarren kapituluan aipatu den aldaeren tratamendurako informazioa erabiltzen da hizkuntzaren ezaugarriekin lotutako erroreak zuzentzeko. Alegia, azpilexiko-multzo bat morfemetan eta

morfotaktikan gertatu ohi diren akats edo gaitasun-erroreetarako definitzen dira, eta bi mailatako morfologiaren araberako erregela-multzo bat aldaketa morfofonologiko erregularrak adierazteko (Aduriz *et al*., 93). Honekin bibliografian agertzen den gaitasun-erroreen tratamendurik osoena egiten da.

V.3.1.4 Tratamenduaren garrantzia errore-mota eta aplikazioaren arabera.

Aplikazioaren arabera errore-mota batzuk ager daitezke eta beste batzuk ez. Erroreen tratamendua erabakitzeko orduan, irizpide nagusia maiztasuna izan ohi da, baina honek ez du zertan beti horrela izan behar.

Elkarrekintzazko zuzenketan askoz inportanteagoa da gaitasun-erroreen tratamendua errore tipografikoena baino, hauek maiztasun handiagokoak izan badaitezke ere. Azken hauetan erabiltzaileak hitz egokiaren idazkera gehienetan dakien bitartean, aldaeretan askoz arazo gehiago dauka berak bakarrik zuzentzeko. Aldaeren tratamenduak are garrantzi handiagoa hartzen du OLIren arloko aplikazioetan, edo hizkuntzaren ezagumendua baldintzaturik dagoenean —euskararen kasuan aipaturiko batasun-prozesuarengatik dago baldintzatua, adibidez—.

Baieztapen honekin bat datoz ikerlari bat baino gehiago, ondoren ikus daitekeenez:

> "... In man-machine communication, the correction of competence errors is far more important than the correction of performance ones. ..." (Veronis, 88:708).

> "... Most of the correction methods currently in use in spelling checkers are biased toward the correction of typographical errors. We argue that this is not the right thing to do. Even if orthographical errors are not as frequent as typographical errors, they are not to be neglected for a number of good reasons. First, orthographical errors are *cognitive* errors, so they are more persistent than typographical errors: proof-reading by the author himself will often fail to lead to correction. Second, orthographical errors leave a worse impression on the reader than typographical errors. Third, the use of orthographical correction for standardization purposes (e.g. consistent use of either British or American spelling) is an important application appreciated by editors. ..." (van Berkel & de Smedt, 88:77).

Zaila egiten zaigu beste zerbait eranstea; bakarrik gehitzea ideia horiek buruan geneuzkala proiektuari ekin genionean. Estandarizazioari egiten zaion aipamenarekin lotuz, euskararen estandarizazio-prozesuan lagungarri den zuzentzaile bat diseinatzea izan da gure lanaren helburu nagusietako bat.

V.3.2 Antzekotasun-neurriak.

Aurreko ezaugarriak oinarritzat hartuz, zuzentzeko metodo/algoritmoak diseinatzen dira. Metodo hauetan, askotan, zuzenketa automatikoa edo errore tipografikoak zuzentzeko

egindakoetan batez ere, agertzen den arazo nagusia zera da: zenbait hitzen artean nola aukeratu erroredun hitzarekiko “antz” handiena duena. Honetarako neurri desberdinak erabiltzen dira ondoan ikusiko ditugunak dira erabilienak.

Gai hau oso zabala eta korapilatsua da, beraz, sarrera bat besterik ez da egingo. Honetaz sakontzeko Kukich-en artikuluko bigarren kapitulua gomendagarria da oso.

Edizio-distantzia

Damerau-ren tipifikazioaren arabera, bi formaren artean dauden bihurketa bakunen kopuru minimoa da distantzia hau. Horrela diren *baina* eta *bania* testu-hitzen artean dagoen edizio-distantzia batekoa da, bi karaktere jarrairen truke bakar batez (*in* - *ni*) batetik bestera igaro baitaiteke.

Neurri horren arabera, bateko distantzia minimoan hitz asko egon daitezkeenez, haien artean aukeratzeko bestelako irizpideak ere erabil daitezke: hurbilpen fonologikoa edo teklatuaren araberakoa testu-ediziorako, hurbilpen grafikoa OCR aplikazioetarako, maiztasun handieneko hitzak, etab.

Neurri horrek bi eragozpen aurkezten du: (1) edozein bi karaktere-kateren arteko edizio-distantzia kalkulatzea ez da berehalakoa; eta inportanteena (2), hitz bat ondo zuzentzeko hitz posible guztiekin alderatu behar da hitz akasduna, dagokion zuzenketa zehatzena lortzeko, eta hau oso garestia da konplexutasunaren aldetik.

Bigarren puntua izan da oso ikerlerro garrantzitsua, batez ere OCR aplikazioetan. Aldaera kopuru izugarria laburtzeko, besteak beste —programazio dinamikoa, luzeraren araberako bilaketa, etab.—, distantzia-neurri berriak proposatu dira. Horien artean inportanteenak diren honako bi hauek aipa daitezke: kodeen arteko distantzia eta n-gramen arteko distantzia.

Kodeen arteko distantzia

Hashing tekniketan oinarriturik hiztegiko forma guztiei kode bat esleitzen zaie; ondorioz, hiztegia kodeen arabera antolatzen da eta distantzia hitzen artean kalkulatu beharrean kodeen artean kalkulatzen da.

Normalean kontsonanteei, batez ere hasierakoei, balio handiagoa ematen zaie eta karaktere errepikatuei ez zaie jaramon handirik egiten. Pollock eta Zamora-k (1988) teknika hau erabiltzen dute SPEEDCOP sisteman, baina kode bakarraren ordez bikoitza, *skeleton key* eta *omission key*, erabiltzen dute, bilaketa zehatzago eta azkarrago burutzearren.

Multzo honetako emaitzak oso onak izan daitezke, baina horretarako kodeketa eta informazioaren antolaketa konplexu samarrak behar dira.

n-gramen arteko distantzia

Kodeak erabili beharrean n-gramak (trigramak normalean) erabili ohi dira karaktere-kateen arteko distantziak kalkulatzeko eta konplexutasuna txikitzeko. Hitzen arteko distantzia haien arteko amankomuneko trigramen araberakoa izango da. Erabiltzen diren trigrama-egiturak (edo orokorrean n-gramenak) bitarrak izaten dira —trigrama onartzen den ala ez esanez— eta trigramaren posizioa kontuan har daiteke edo ez. Hiztegia dagoenean, hitzen trigramen arabera indexatu ohi da bilaketa errazteko.

Honen adibidea ACUTE sistema (Angell *et al*., 83) dugu. Trigrametan eta hitzen luzeran oinarritutako sistema honetan distantzia neurtzeko formula honako hau da:

$$d = c / \max(n, n')$$

non *c* amankomuneko trigrama kopurua den eta *n* eta *n'* hitzen luzerak.

Oso emaitza onak azaltzen dituzte, karaktere jarraien arteko trukearen kasuaren salbuespenaz.

Saio hauez gain, hitzak trigrama-bektoreen bidez adierazten dituzten teknika sofistikatuak ere proposatzen dira; horien gainean Hamming-en distantziak bezalako neurriak aplikatzen direlarik.

V.3.3 Zuzenketa-metodoak

Berriro azpimarratu behar da hitz isolatuak zuzentzeko teknikak beti oso mugatuak direla, ondo zuzentzeko testuingurua kontuan hartzea ezinbestekoa da eta. Honen adierazgarri pertsonekin egindako testak ditugu: zenbait laguni emandako testuingururik gabeko akatsen aurrean eskatzen zitzaien hiru edo lau hitzen artean aukera zezaten, batez-besteko asmatze-tasa %75ean ezarriz (Kukich, 92:411). Sistema automatiko onenetan antzeko zenbakiak lortzen dira.

Aurretik esandakotik ondoriozta daitekeenez hitz baten zuzenketa dirudiena baino eginkizun konplexuagoa da. Horren lekuko da PF-474 txipa (Yianilos, 83), helburu berezitu honetarako diseinatu dena.

Ondoren zuzenketa-metodo adierazgarrienak azalduko dira; oinarrizkoak aurretik eta konbinatuak ondoren. Azpimarratzekoa da zenbait sistemaren inguruan dagoen informazio-falta, gaiak duen interes komertziala dela eta.

V.3.3.1 Oinarrizko metodoak

Zuzenketara aplikatzen diren funtsezko metodoak azaltzen dira honako lau multzo hauetan bereiziak: (1) alderantzizko edizio-distantziaren bidezkoa; (2) hitz guztiekiko distantziaren bidezkoa; (3) erregelen bidezkoa eta (4) metodo estatistikoak.

Alderantzizko edizio-distantziaren bidezko metodoak.

Metodoaren funtsa hauxe da:

- Akasdun formatik abiatuta Damerau-ren legeak aplikatzen dira alderantziz. Horrela, eta bateko distantziara mugatuz, V.3.1.1 atalean aipatutako zenbakiak erabiliz *2nk* hipotesi sortzen dira, *k* alfabetoaren karaktere-kopurua eta *n* hitzaren luzera izanik.
- Hipotesiak egiaztatzen dira hizkuntzaren hitzak diren ala ez jakiteko, horretarako V.2 atalean aipatzen diren metodoak erabil daitezkeela.
- Ontzat hartutako hipotesiak sailkatzen dira; lehena aukeratzeko zuzenketa automatikoan edo lehenengo batzuk elkarrekintzazko zuzenketan.

Sistema askotan erabiltzen da metodo hau: (Peterson, 80), (Kernighan *et al*., 90), (Church & Gale, 91) (Vagelatos *et al*., 95). Gure proiektuan erabilitako metodoaren barruan ere, teknika hau errore tipografikoei aurre egiteko erabiltzen da.

Metodo honen ezaugarriak hauek dira: hiztegi osoa ez duten sistemetan aplikagarria, programatzeko eta memoria-hartzearen aldetik sinplea, baina bakun ez diren erroreentzat ez du proposamen egokirik eskaintzen. Azken eragozpen honen aurrean, metodo bera biko distantziarekin aplika daiteke, baina horren ondorioz, aukera-kopurua izugarri haziko litzateke eraginkortasunaren kalterako.

Hitz guztiekiko distantziaren bidezko metodoak.

Helburua zera da: errorea eman duen testu-hitza hizkuntzaren hitz posible guztiekin —edo askorekin, optimizazio-teknikak erabiltzen badira— alderatzen da, berarekiko distantzia txikiena duena aukeratuz zuzenketa automatikoan eta hurbilen dauden lehenak elkarrekintzazko zuzenketan.

Hau da metodo erabiliena eta gehien ikertu dena, testu-edizioan batez ere. Sistemen artean hauek daude: (De Heer, 82), (Angell, 83), (Pollock & Zamora, 84), (Hull & Srihari, 82), (Tanaka, 87). Aipatutako PF-474 txipa eragiketa hau arintzeko diseinatua da.

Metodo honen ezaugarriak hauek dira: emaitza onak lortzen dira, eta gainera, beti aurkitzen da zuzenketaren bat; baina horretarako hiztegi osoa biltegiraturik eduki behar

da. Aurreko atalean aipatu den bezala, distantzia- edo hurbiltasun-neurri desberdinak erabil daitezke, teknika horien artean zehaztasun, memoria-hartze eta eraginkortasun neurri desberdinak lortuz; hiru irizpideak batera optimizatzeko metodorik ez dago ordea.

Erregelen bidezko metodoak.

Ezagumendu linguistikoa erabiliz erregelak sortzen dira akasdun formatik dagokion forma zilegia lortzeko. Erregela hauek gehienetan morfofonologikoak izaten dira, baina aztertutako corpusetan gertatzen diren errore tipografikoetatik inferitutako erregelen bidezko sistemak ere sartzen dira multzo honetan.

Sailkapena zehaztearren bi multzotan bana daitezke metodo hauek:

- Zuzenean hitz zilegiak lortzen dituzten metodoak; hauetan, erregelak zeharo linguistikoak direnez, lortzen diren zuzenketak hizkuntzaren formak izanik. Hauek dira benetako “metodo linguistikoak”. Multzo honetan kokatzen da gure proiektuaren barruan Xuxen zuzentzaile ortografikorako egindako aldaeren tratamendua (ikus §VI.4 atala).

- Proposamen hipotetikoak lortzen dituztenak, ondoren hipotesi hauek egiaztatu behar direla. Multzo hau lehen multzoaren aldaketa bezala ere ikus daiteke, Damerau-ren erregelak aplikatu beharrean beste batzuk aplikatzen direlarik. Mota honetako metodoa dugu Yannakoudakis eta Fawthrop-ek (1983) proposatutakoa, non erregelak baino heuristikoak erabiltzen diren.

Ezaugarrien aldetik, berriz, metodo hauek oso emaitza onak ematen dituzte erroreak aurrikusitako parametroen barruan gertatzen badira, baina oso txarrak gainontzekoetan; eta horrexegatik osatu ohi dira beste metodo batzuekin sistema konbinatuak eginez.

Metodo estokastikoak.

Aurreko erroreetan oinarriturik, automatikoki inferitutako informazioa erabiliz zuzentzen dira akats berriak. Normalean, ikasketa-prozesu bat behar dute aurretik; prozesu horretan, eskuz prestatutako edo zuzendutako corpus batez, akatsak eta hitz zilegien artean erlazioak bilatzen dira. Ezagumendua inferitzeko teknika nagusiak hiru dira: taula estatistikoak, eredu markoviarrak, eta sare neuronalen eredua. Teknika hauek etiketatze-lanetan (*tagging*) erabiltzen diren berberak dira, eta lexikorik erabili gabe lan egiten duten zuzentzaileetan erabiltzen dira batez ere, egiaztapena hitz-zatietan oinarrituz.

OCR aplikazioak izan ohi dira teknika hauen helburua, OCR dispositiboek akatsak modu erregularrean egiten baitituzte, pertsona desberdinen akatsak askoz irregularragoak izanik —pertsona bakoitzeko ikasketa-prozesu berezia beharko litzateke—. Gainera, lexikorik edo hiztegirik gabe lan egiten den aplikazioetan —OCR eta hizketaren

tratamendua normalean— eta lehen bi teknika-multzoak ezin direla erabili kontuan hartuz, lortzen diren emaitzak aipagarriak dira.

Adibide gisa *correct* programa (Kernighan *et al*., 90) dugu. Corpusen gainean lortutako probabilitateetan oinarritzen dira bere oinarrizko sistema osatzeko —oinarria alderantzizko edizio-distantziaren bidez eraikitzen da— eta oso emaitza onak azaltzen dituzte. Damerau-ren lau errore motetako bakoitzerako "nahasketa-matrize" bat osatu dute datuen arabera, eta hitz akasduna ordezkatzeko gai direnen artean sailkapen bat egiten dute, hitzaren probabilitate absolutua eta akatsekiko desberdintasunaren probabilitatea biderkatuz. Metodo estokastiko honekin oso emaitza onak lortzen dira errore bakunetarako.

V.3.3.2 Metodo konbinatuak.

Testu-edizioan aplikatutako zuzenketarako metodo konbinatuak ari dira proposatzen azken urteotan, eta hauetan sakonduko dugu.

Berriro De Smedt eta Van Berkel-en hitzak aldatuko ditugu hona:

> "Of the method described in the previous chapter, no single method sufficiently covers the whole spectrum of errors. Because each method has its strengths and weaknesses, it is advantageous to combine two methods which supplement each other." (van Berkel & de Smedt, 88:80)

Lehentxeago aipatutako *correct* da hauetako bat, alderantzizko edizio-distantzia eta metodo estokastikoak konbinatzen dituena. Ematen duten asmatze-tasa %87a da, baina neurria ez da estandarra. Normalean lehen edo lehen hiru proposamenekin zenbatetan asmatzen den izaten bada neurria, beraiek testuingurua kontuan hartu gabe hiru pertsonek emandako epaien artean gutxienez birekin bat etortzea hartzen dute neurri-unitatetzat.

Ondoren beste bi metodo konbinatu azaltzen dira gainbegirada osoa lortzearren.

Triphone (van Berkel & de Smedt, 88)**.**

Entziklopedia bat kontsultatzeko diseinaturiko sistema honek bi metodo konbinatzen ditu: errore fonetikoak tratatzeko fonemen gaineko erregelak erabiltzen zituen Spell Therapist batetik, eta forma guztiekiko distantzian oinarritutako trigramen bidezko FUZZIE[1] (De Heer, 82) metodoa.

[1] Metodo hau egokiagoa da aipatutako ACUTE (Angell, 83) baino, azken honetan luzerak funtsezko papera duelako eta fonema batek karaktere kopuru aldakorra duelako.

Proposatzen duen irtenbidea hau da: distantzia kalkulatzeko trifonemen arteko distantzia erabiltzea trigramena erabili beharrean, ondorioz hiztegia trifonemen arabera antolatuz. Jarraitzen den algoritmoa honako hau da:

- bere fonemen arabera hitza trifonemetan banatzen da (banaketa-aukerak optimizatuz)
- trifonema bakoitzari dagokion maiztasuna lortzen da
- zenbait trifonema aukeratzen dira, maiztasun-muga batetik behera dauden hautapen-trifonemak deitutakoak, eta horien arabera antolaturiko fitxategian bilatzen da.
- bide honetatik aurkitutako hautagai guztiekin amankomuneko trifonemen arabera antz handiena dutenak aukeratzen dira.

Azaltzen dituzten emaitzak oso onak dira, %92 lehen proposamenean, baina eremua oso mugatua da pertsona-izenekin bakarrik probatzen delako.

Forma guztiekiko distantzia + morfofonologia (Veronis, 88)**.**

Minitel kontsulta-sistemarako garatutako sistema honetan, hitz isolatuak zuzentzeko metodoa —komunztadura-akatsak zuzentzeko beste metodo bat ere badago— ezaguna den beste batean (Durham *et al*., 83) oinarritzen da eta hiru multzotan banatzen da:

- Forma guztiekiko distantzia da jarraitzen den oinarrizko metodoa. Errore tipografiko bakar bat hartzen du kontuan.
- Fonologiaren menpe dagoen karaktere-multzoen artean antzekotasun-taula bitarra (antzekoak ala ez esaten baitu) erabiltzen du, antzeko multzoak berdintzat joz distantziak neurtzerakoan.
- Aurreko guztia erroekin egiten da, ondoren atzizkien tratamendu morfologiko ad-hoc sinple bat eginez.

V.4.Hizkuntza flexionatuen eta eranskarien zuzenketa.

Flexio handiko hizkuntzetan zein hizkuntza eranskarietan erroreen zuzenketa korapilatsuagoa da beste hizkuntzetan baino. Hala ere, eta ingelesaren flexio-sistema sinplea dela eta, hizkuntza hauen gaineko zuzenketa ez da sakonean aztertu.

Lexikoa erabiltzen duten aplikazioetara murriztuz —lexikorik gabekoetan, hitz-zatietan oinarritutako egiaztapenean, eta erregelen zein metodo estokastikoen bidezko zuzenketan ez baitago alde nabarmenik hizkuntza-motaren arabera—, honela labur daiteke hizkuntza hauetarako zuzenketa-mekanismoa:

- **Egiaztapena** analisi morfologikoaren bidez burutzen da, hitz-zerrenda oso bat desegokia eta ezinezkoa edo memoria-hartzearen aldetik garestiegia baita. Mota horretako hizkuntzetarako lortutako erreferentzia guztietan horrela egiten da. Izan ere, lexikoa aberasteko orduan, geroxeago aztertuko denez, informazio linguistikoa jaso behar da erabiltzailearengandik.

- **Zuzenketa** egiteko arazoak handiak dira, eta proposamen desberdin egin dira. Forma posible guztiekiko distantzia kalkulatzea ezinezkoa da —forma guztiak ez baitira inon gordetzen—, eta, ondorioz, morfologia eta zuzenketa-metodoak konbinatu behar dira.

V.4.1 Lexikoaren aberasketa.

Esan bezala, euskara bezalako flexio handiko hizkuntzetan errorea dagoenentz jakiteko testu-hitzen analisi morfologikoa egin ohi da. Horretarako sisteman, lexikoan normalean, morfologiari buruzko informazioa metatzen da. Lexikoa itxia denean ez dago arazorik, baina lexikoaren aberasketa erabiltzaile arruntaren esku uzten denean ondoko arazoak sortzen dira:

- sarrera berriei buruzko informazio morfologikoa beharrezkoa da berauen flexioa ondo era dadin.

- informazio linguistiko hori modu automatikoan sortzea ezinezkoa izaten denez, elkarrizketa-protokolo bat diseinatu eta erabili behar da erabiltzaileari informazio hori eskatzeko.

- informazio linguistikoa linguistikan aditua ez den erabiltzaile bati eskatzen zaionez gero, elkarrizketa sinpleaz baina ahalik eta informaziorik zehatzena lortu behar da, eta hau, askotan, ez da erraza.

Gai honetaz bibliografian dagoen informazioa hutsetik hurrena da. Gure zuzentzailearen barruan elkarrizketa-modulu hau diseinatu dugu helburu bikoitz horrekin: sinplea izatea baina zehaztasunik galdu gabe (ikus §VI.6 irudia).

V.4.2 Zuzenketa. Zenbait sistema.

Flexio aberatseko hizkuntzetan, V.3 atalean ikusitakoaren arabera, eta morfologian oinarrituriko sistema baterako aritzen garela suposatuz, akats baten gaineko zuzenketarako ondoko algoritmoa har daiteke oinarritzat:

- *alderantzizko edizio-distantzia*ren metodoa erabiliz, hitz-forma honetarako proposamen hipotetiko guztiak sortu.

- Hitz-forma hipotetiko guzti hauek benetako hitzak diren ala ez jakiteko berauen analisi morfologikoa burutu, arrakastatsuak aukeratuz.

- Benetako hitzen artean sailkapena egin.

*Forma hipotetikoen analisi*an oinarritutako algoritmo horren eragozpen nagusiak bi dira:

1) Oso motela gerta daiteke, batez ere analisi morfologikoa konputazio-konplexutasun handi samarrekoa denean. Azkartzeko metodoak —maiztasun handieneko hitzak, trigrama okerren bidezko bazterketa, etab.—bidera badaitezke ere, sakoneko arazoa izan daiteke.

2) Akatsa bakuna ez denean ez da proposamenik (edo proposamen egokirik) sortuko, alderantzizko edizio-distantziaren metodoa bateko edizio-distantziarekin lan egiten du eta. Biko distantziara heda liteke metodoa baina horrekin izugarri areagotuko litzateke lehen eragozpena.

Oinarrizko algoritmo hori izan da, funtsean, guk errore tipografikoen tratamendurako erabili duguna. Ikus dezagun bibliografian agertzen diren zuzenketa-tratamendu garrantzitsuenak morfologian oinarrituriko sistemetarako:

- **Tratamendurik ez**. Zenbait sistematan ez da zuzentzeko laguntzarik ematen: (Hajic & Droza, 90), (Solack & Oflazer, 93).

- **Forma hipotetikoen analisia**: Aurretik ikusitako tratamendua, alderantzizko edizio-distantziaren metodoan oinarritzen dena, edo horren deribazioren bat da hedatuena, errore tipografikoetarako behinik behin: (Means, 88), (Vagelatos *et al*. 95).

- **Erro-hizkiak**. Hitza ezagutzen ez den analisian erro eta hizkien banaketa burutzen da, eta bakoitzaren zuzenketari ekiten zaio V.3 atalean aipatutako metodoren batez. Bukaeran sorkuntza morfologikoaren bidez erroa eta hizki zilegien bidez lortzen dira zuzenketarako hitzak. Metodo honen eragozpena lehen banaketan datza, baina trukean hipotesien analisia saihesten da. Horren adibidetzat har daiteke gure zuzentzailean egiten den aldaeren zuzenketa. Veronis-ek (1988) frantseserako sistema batean ideia honi jarraitzen dio. Autore horrek burutzen duen tratamendu morfologikoa, hala ere, ez da osoa, atzizkien tratamendu partikular bat besterik ez baitu egiten; beraz, ez da egokia izango hizkuntza eranskarietarako.

Oflazer eta Guzey-k ondoren azalduko dugun tarteko tratamendua proposatzen dute, erro guztiekiko distantzian eta sorkuntza morfologikoan oinarritzen dena.

Hizkuntza eranskarietarako proposatutako metodoa (Oflazer & Guzey, 94)**.**

Turkiera hizkuntza eranskaria denez erroreen zuzenketa korapilatsua da. Aipatu den bezala, lehen zuzentzaile batean ez da zuzenketarako mekanismorik eskaintzen (Solack & Oflazer, 93). Azken urteetan, eta bi mailatako ereduan oinarritutako prozesadore morfologiko bat burutu ondoren (Oflazer, 93), zuzenketaren arazoari ekin diote.

Algoritmoaren funtsa bi urratsetan banatzen da:

1) Akasdun hitzaren erro posible guztiak lortzea erro-hiztegitik.

2) Lortutako erroetatik abiaturik akasdun formarekin antza duten formak sortzea.

Lehen urratsa burutzeko erroen azalak eta akasdun hitzaren hasierako azpikateak[1] alderatzen dira edizio-distantziaren metodoa erabiliz, distantzia honi muga bat jarriz. Bilaketa hau laburtzearren, bigramen araberako bektoreen indizeak eratzen dira lexikoa atzitzeko —bakarrik akasdun formarekin amankomuneko *k* bigramak dituzten erroak hartzen dira kontuan.

Bigarren urratserako bi hipotesi egiten du: balizko erroari dagokion eskuineko azpikatea —atzizki-multzoa osatuko lukeena— zuzena dela suposatuz batetik eta ez dela suposatuz bestetik.

Lehen kasuari "*on the left edge of the word*" izena eman diote eta tratamendu erraza du: erroari dagokion bukaerako azpikatea erantsi eta analizatu, analizatuz gero proposamen bat lortzen delarik. Erroa eta hasierako azpikatearen arteko distantzia mugan bada, hori da erro horretarako saio bakarra.

Bukaerako azpikatean akatsik egon daitekeela suposatzen bada, erroaren araberako sorkuntzari ekiten zaio, eta hasierako formarekiko distantzia osoa —erroari eta atzizkiei dagozkienak batuz— mugatik behera duten kasuak aukeratzen dira. Aztertzeko kasuak laburtzearren, "*Cut-Off Paths*" izeneko teknika (Du & Chang, 92) erabiltzen dute.

Erroaren eta atzizkiaren artean gerta daitezkeen karaktere-aldaketak eta galerak kontuan hartzeko, distantzia-mugaren gainean "ukituak" egiten ditu.

Azkenik, distantzia bera dutenen proposamenen artean sailkapena egiteko aztertutako datuen gainean egindako estatistikak darabiltzate.

Azaltzen dituzten emaitzak onak dira zehaztasunaren aldetik, baina ez hainbeste eraginkortasunaren aldetik. Adibidez, bateko distantzia-muga ezarriz —errore bakunak

[1] Aurrizkiak ez ditu kontuan hartzen.

bakarrik zuzenduko dira—, zehaztasun handia lortzeko aipaturiko *k* faktoreak hiru izan behar du eta, ondorioz, batez-besteko analisiak 31 dira, sorkuntzak 311 eta distantzia-eragiketak 2500. Biko distantziarekin, zehaztasuna mantentzeko, neurri horiek bost aldiz handiagoak dira.

Neurri hauekin eta datorren kapituluan ikusiko dugun gure sistemarekin konparatuz, zera ondoriozta daiteke: metodoaren sofistikazioak ez du ebazten *forma hipotetikoen analisi*an oinarritutako metodo klasikoaren eragozpen nagusia, eraginkortasunarena hain zuzen.

V.4.3 Ondorioak.

Zuzenketa-metodoen gainean egindako azterketa honetatik ondorio hauek lortzen dira:

- Bi metodo multzo bereizten dira: informazio lexikoa erabiltzen dutenak eta hitz-zatietan oinarritzen direnak. Lehenak zehatzagoak dira, baina ez dute malgutasun handirik, lema edo hitza lexikoan ez dagoenean hitz zilegi bat akastzat hartzen baita. Bigarrenak OCR motako aplikazioetan erabiltzen dira nagusiki, zuzenketa automatiko, azkarra eta malgua behar izaten delako aplikazio hauetan.

- Lexikoa erabiltzen duten metodoek eraginkortasunaren eta memoria-hartzearen artean oreka lortu behar dute, eta horretan hizkuntzaren ezaugarri morfologikoek zerikusia handia dute. Forma-zerrendak oso azkarrak dira baina memoria hartze handia dagokie, eta alderantziz gertatzen da morfemetan oinarritutako metodoetan. Mailatan eratutako sistemak interesgarriak dira.

- Testuingurua kontuan hartzen ez bada zuzenketa, automatikoa batez ere, ez da zehaztasun handikoa izango.

- Elkarrekintzazko zuzenketan errore tipografikoen zuzenketak interes txikiagoa du fonetikak edo ez-jakiteak eragindakoenak baino; haiek nola zuzendu jakin ohi den bitartean besteak ez.

- Hizkuntza eranskarietan hitzen egiaztapena morfologian oinarritu ohi da, eta zuzenketa konplexu samarra gertatzen da.