Testuen sinplifikazio automatikoa: arloaren egungo egoera

Automatic Text Simplification: State of Art

Itziar Gonzalez-Dios
Ixa Taldea. Euskal Herriko Unibertsitatea (UPV/EHU)
itziar.gonzalezd@ehu.es

María Jesús Aranzabe
Ixa Taldea. Euskal Herriko Unibertsitatea (UPV/EHU)
maxux.aranzabe@ehu.es

Arantza Díaz de Ilarraza
Ixa Taldea. Euskal Herriko Unibertsitatea (UPV/EHU)
a.diazdeilarraza@ehu.es

Laburpena

Lan honen helburua testuen sinplifikazio automatikoaren arloaren egungo egoera aurkeztea da. Horretarako, ikerketa-lerro honetan egindako sistemak eta prototipoak jaso ditugu hizkuntzaren, helburu duen taldearen eta egiten duen sinplifikazio motaren (sintaktikoa, lexikala edo biak) arabera sailkatuz. Sistemak ebaluatzeko metodoak eta garatu diren baliabideak eta tresnak ere aurkeztuko ditugu.

Gako-hitzak

Testuen sinplifikazio automatikoa, esaldien sinplifikazioa, arloaren egungo egoera

Abstract

The aim of this paper is to give an overview of the state-of-art in automatic text simplification. To that end, we present the systems and prototypes according to the language the are built for, their target audience and the type of simplification (syntactic, lexical or both) they perform. Moreover, we expound the different evaluation methods that have been carried out with these systems and the resources and tools developed so far.

Keywords

Text simplification, sentence simplification, state-of-art survey

1 Sarrera

Testuen sinplifikazioak testu bat testu sinpleagoa lortzea du helburu jatorrizko testuaren esanahia mantenduz; egitura eta hitz konplexuak ordezkatuz sortzen den testua errazagoa izan behar da irakurle jakin batzuentzat. Ezaguna da testu errazek edo sinpleek abantaila ugari eskaintzen dizkietela, bai pertsonei, bai Hizkuntzaren Prozesamenduko tresnei (Chandrasekar, Doran, & Srinivas, 1996).

Testuen sinplifikazioa irakaskuntzan eta batez ere atzerriko hizkuntzen didaktikan landu izan da. Arlo horretan testu sinplifikatuak erabiltzearen helburua ulermena areagotzea eta karga kognitiboa leuntzea da (Crossley, Allen, & McNamara, 2012).

Irakaskuntzan, hain zuzen ere, Allen-ek (2009) eta Crossley-k, Allen-ek, & McNamara-k (2012) sinplifikatzeko bi aukera daudela azaltzen dute:

- Egituraren araberako sinplifikatzea: maila jakin bakoitzari dagozkion hitz zerrendak eta egitura sintaktikoen zerrendak erabilita gauzatzen da, eta kasu batzuetan, testuen konplexutasun maila neurtzen duten (*Readability*) formuletan oinarritzen da.
- Intuitiboki sinplifikatzea: maila jakin bakoitzari dagozkion hitz zerrendak eta egitura sintaktikoen zerrendak kontuan izan ditzaketen arren, oro har, intuizioari jarraiki sinplifikatzen da.

Young-ek (1999) hainbat metodo aipatzen ditu sinplifikatzeko atzerriko hizkuntzaren irakaskuntzan:

- Linguistikoki sinplifikatzea: testua berridaztea esaldiak laburragoak egiteko, esaera idiomatikoak ezabatzea edo parafraseatzea, hitz espezializatuak eta maiztasun gutxikoak ekiditea, eta sintaxi konplexua errebisatzea perpaus bakunak sortzeko.

- Materia sinplifikatzea: testua laburtu, paragrafoak edo atalak kenduz.
- Glosen bitartez sinplifikatzea: itzulpenak edo definizioak gehitzea.
- Prozesamendu kognitiboetan oinarritutako aldaketak eta elaborazioak eginez sinplifikatzea.

Simensen-ek (1987) atzerriko hizkuntzetako liburuak egokitzeko argitaletxeek dituzten gidalerroak aztertu zituen, eta gidalerro horietan oinarrituta, egokitzapenerako hiru printzipio (kontrolaren printzipioak) aurkeztu zituen: informazioaren, hizkuntzaren eta diskurtsoaren kontrola. Hizkuntzaren kontrolaren atalean, Crossley-k, Allen-ek, & McNamara-k (2012) ere azaldutako bi hurbilpenak aurkezten ditu: zerrendetan (egituraren araberako sinplifikazioa) eta intuizioan (intuitiboki sinplifikatzea) oinarritutakoak.

Atzerriko hizkuntzen didaktikaren arloan, testu sinplifikatuak edo originalak erabiltzeak dituen abantailak ere aztertu izan dira, emaitzak kontrajarriak diren arren. Esaterako, Youngen (1999) azterketaren arabera, testu sinplifikatuak ez dira baliagarriak ikasleak irakurketa globala egiten ari badira; are gehiago, kalterako izan daitezke. Oh-ren (2001) arabera, ordea, irakurmenaren ulermen globala erraztu egiten dute testu sinplifikatuek. Crossley-k, Allen-ek, & McNamara-k (2012), berriz, uste dute testu sinplifikatuek erreferentziakidetasun bidezko kohesio altuagoa eta lokailu eta hitz ezagun gehiago eskaintzen dizkietela ingelesa ikasten ari diren ikasleei. Dena dela, arlo honetako autore gehienek gaian sakondu behar dela adierazten dute (Young, 1999; Allen, 2009; Crossley, Allen, & McNamara, 2012).

Hizkuntzaren Prozesamenduari (HP) dagokionez, ikerketa-lerro hau azken urteetan garrantzitsua bihurtu da, ingeleserako ez ezik beste hainbat hizkuntzatarako ere proposatu direlako sistemak eta metodo eta teknika berri ugari argitaratzen ari direlako. Testuak eskuz sinplifikatzeak edo egokitzeak lan handia eta garestia eskatzen du; beraz, HPko tresnak erabiliz testuak sinplifikatzean lana erraztu eta azkartzen da. HPan hartu den sinplifikatzeko ildoa linguistikoki edo estrukturalki sinplifikatzearena izan da, zehazki bi sinplifikazio mota landu dira: sintaktikoa eta lexikoa.

Testuen Sinplifikazio (TS) automatikoa Siddharthan-ek (2002) berridazketa prozesu bat bezala definitzen du; ataza horren helburua konplexutasun lexikala eta sintaktikoa gutxitzea da. Aluísio-k & Gasperin-ek (2010), berriz, HPko ikerketa-lerro bezala definitzen dute PorSimples[1] proiektuan. Bertan Brasilgo portugesezko testuen ulermena areagotzeko, fenomeno lexikalak eta sintaktikoak sinplifikatzen dituzte; lehendabizi, pertsona gutxik ulertzen dituzten hitzak ezagunagoekin ordezkatzen dituzte eta bigarrenik, esaldiak banatu eta esaldien egitura aldatzen dute.

Era berean, TSak HPko beste ataza eta ikerketa-lerro batzuekin elkarrekintza zuzena du, bai atazaren antzekotasunagatik, bai antzeko metodoak erabiltzen dituztelako. Horien artean hurbilen laburpen automatikoa (*summarisation*) dago; bien arteko desberdintasun nagusia da laburpen automatikoak testua trinkotzea duela helburu eta TSak testuen konplexutasun linguistikoa gutxitzea, alegia, ez du testua kondentsatzea helburu. Bi ikerketa-lerro horien arteko muga ezartzea, gainera, kasu batzuetan zaila izaten da erabiltzen diren teknika eta metodo batzuk berdinak direlako. TSaren antzeko beste atazak dira batetik, parafrasiak ikastea eta sortzea (*paraphrase acquisition and generation*), testuak sinplifikatzeko maiz parafrasiak erabili eta berridazketak egiten direlako, eta bestetik, konplexutasuna ebaluatzea (*readability assesement*), TSren aurreprozesu bezala erabili ohi dena.

Lan honetan, HPan egin den testuen sinplifikazioan kontzentratuko gara ikerketa-lerro honen artearen egoera ezagutzera emateko. Sarrera honen ondoren, beraz, 2. atalean HPan zein hizkuntzatarako eta zein helburu taldetarako egin diren lanak aurkeztuko ditugu sortu diren baliabideekin batera. Sinplifikazio motak eta metodoak 3. atalean erakutsiko ditugu. Sistemak ebaluatzeko erabili diren metodoak 4. atalean aurkeztuko ditugu egin diren sistema eta prototipoen bitartez. Amaitzeko, ondorioak aurkeztuko ditugu 5. atalean.

2 Sinplifikazio automatikoa HPan

HPko testuen sinplifikazioan hasierako lanak ingeleserako egin ziren. Lehen lana Chandrasekaren, Doran-en, & Srinivas-ena (1996) izan zen eta TSrako motibazioak azaldu ziren. Hasierako beste lanen artean, PSET (*Practical Simplification of English Text*) proiektuan[2] (Carroll et al., 1998) egindakoa aurki daiteke. Proiektu horretan hizkuntzarekin arazoak zituztenei eta, batez ere, afasia zuten pertsonei zuzendutako sinplifikazioa egin zuten (Carroll et al., 1999). Siddharthan-ek

[1] `http://caravelas.icmc.usp.br/wiki/index.php/Principal` (2011ko irailean atzituta)

[2] `http://www.informatics.sussex.ac.uk/research/groups/nlp/projects/pset.php` (2013ko maiatzean atzituta)

(2002), berriz, testuen sinplifikazio automatikorako oinarrizko arkitektura finkatu zuen.

Ikerketa-lerro hau oso garrantzitsua bilakatu da 2009tik aurrera eta beste hizkuntzetara zabaltzeaz gain, metodo eta teknika berri ugari argitaratu dira, batez ere metodo estatistikoetan eta ikasketa automatikoan oinarrituz.

Esan bezala, TSrako sistema gehienak ingeleserako proposatu eta egin dira; horien artean ditugu Siddharthan-ena (2006) eta Zhu-ren, Bernhard-en, & Gurevych-ena (2010). Azken urteotan beste hizkuntzetarako ere egin dira: japoniera (Inui et al., 2003), Brasilgo portugesa (Candido et al., 2009; Gasperin et al., 2009; Aluísio & Gasperin, 2010), suediera (Rybing, Smith, & Silvervarg, 2010; Keskisärkkä, 2012), arabiera (Al-Subaihin & Al-Khalifa, 2011), gaztelania (Saggion et al., 2011; Saggion, Bott, & Rello, 2013), frantsesa (Seretan, 2012; Brouwers et al., 2012), euskara (Aranzabe, Díaz de Ilarraza, & Gonzalez-Dios, 2012), italiera (Barlacchi & Tonelli, 2013), daniera (Klerke & Søgaard, 2013), bulgariera (Lozanova et al., 2013) eta koreera (Chung et al., 2013).

Jarraian testuen sinplifikazioa zein helburu talderentzat den baliagarria (2.1 azpiatala) eta horientzat egin diren lanak aipatuko ditugu. Ondoren, TSrako sortu diren baliabideak zerrendatuko ditugu atal honen amaieran (2.2 azpiatala).

2.1 Helburu taldeak

Chandrasekar-en, Doran-en, & Srinivas-en (1996) lanean, TSa gizakientzat eta HPko tresnentzat erabilgarria eta onuragarria dela esaten da. Urteak pasa ahala, lan horretan proposatutako ideiak materializatu dira. Jarraian ikusiko ditugu gizakiei eta tresnei zuzenduta egin diren TS lanak:

- Gizakiak. Testu sinpleek informazioa eskuragarriago bihurtzen dute eta horrela testuak ulertzea errazagoa da. Ondoren, gizakientzat egin diren lanak azpitaldeen arabera zerrendatuko ditugu:
  - Urritasunak dituztenentzat (Carroll et al., 1999): afasikoak (Carroll et al., 1998; Max, 2005; Devlin & Unthank, 2006), jaiotzetiko entzumen arazoak dituztenentzat (Inui et al., 2003; Lozanova et al., 2013; Chung et al., 2013), irakurtzeko arazoak dituztenentzat (Bautista, Hervás, & Gervás, 2012), dislexikoak (Rello et al., 2013), adimen-urritasunak dituzten irakurle pobreentzat (Fajardo et al., 2013)
  - Atzerriko hizkuntzen ikasleentzat (Petersen, 2007; Burstein, 2009)
  - Gutxi alfabetatuentzat (Candido et al., 2009)
  - Haurretzat (De Belder & Moens, 2010; Brouwers et al., 2012; Barlacchi & Tonelli, 2013)
  - Orokorrean eta adimen urritasunak dituztenentzat teknologia munduan murgiltzeko (Saggion et al., 2011; Bott & Saggion, 2012)
- Oinarrizko tresnak edo HPko aplikazio aurreratuak. Esaldi luzeak eta konplexuak dituzten testuak zailagoak izaten dira automatikoki prozesatzeko; esaldi laburragoak eta sinpleak erabiltzen diren kasuetan, aldiz, tresnen eta aplikazio aurreratuen errendimendua hobea da. Hori dela eta, testuen sinplifikazioa aurreprozesu bezala erabil daiteke performantzia igotzeko. Hobekuntza hori bai oinarrizko tresnetan, bai aplikazio aurreratuetan gertatzen da. TSa edo esaldien sinplifikazioa erabili duten lanak dira:
  - Dependentzia-gramatikan oinarritutako analizatzaile sintaktiko edo *parser*ak (Chandrasekar, Doran, & Srinivas, 1996)
  - Laburpen automatikoak egiteko (Lal & Rüger, 2002; Siddharthan, Nenkova, & McKeown, 2004; Blake et al., 2007; Vanderwende et al., 2007; Silveira Botelho & Branco, 2012)
  - Informazioa bilatzeko aplikazioak (Beigman Klebanov, Knight, & Marcu, 2004)
  - Azpitituluak egiteko (Daelemans, Höthker, & Sang, 2004)
  - Itzulpen automatikoa (Doi & Sumita, 2004; Poornima et al., 2011)
  - Rol semantikoak etiketatzeko (Vickrey & Koller, 2008)
  - Arlo berezituetako analizatzailea, adibidez biomedikuntzako testuetan (Jonnalagadda et al., 2009)
  - Informazio erauzketa (Jonnalagadda & Gonzalez, 2010b; Evans, 2011)
  - Gertaeren erauzketa (Buyko et al., 2011)
  - Ahozko hizkuntza ulertzen duten sistemak (Tur et al., 2011)
  - Galdera-erantzun sistemak (Bernhard et al., 2012)

– Erlazioen erauzketa (Minard, Ligozat, & Grau, 2012)
– Corpus paraleloetan hitzak lerratzeko (Srivastava & Sanyal, 2012)

2.2 Baliabideak

TSren ikerketa-lerroan HPko beste atazetan bezala corpusak oso baliabide garrantzitsuak dira. Bi motakoak dira bereziki erabilienak: corpus paraleloak eta corpus ez-paraleloak.

Corpus paraleloetan jatorrizko testua eta testu sinplea lerratuta daude, esaldiz esaldi. Hau da, jatorrizko testuko esaldi bakoitzak bere baliokide sinplea du. Mota horretako corpusak Brasilgo portugeserako (Caseli et al., 2009), gaztelaniarako (Bott & Saggion, 2011) eta danierarako (Klerke & Søgaard, 2012) sortu dira.

Corpus ez-paraleloek, berriz, testu sinpleak eta jatorrizkoak gordetzen dituzte lerratu gabe. Modu horretan, bai Dell'Orletta-k, Montemagnik, & Venturi-k (2011) italierarako, bai Hanckek, Vajjala-k, & Meurers-ek (2012) alemanerako alde batetik, haurrei zuzendutako egunkarietako eta aldizkarietako testuak testu sinple bezala jaso dituzte eta bestetik, prentsa arrunteko testuak. Alemanerako ere, corpus ez-paraleloa eraiki dute Klaper-ek, Ebling-ek, & Volk-ek (2013) webeko testuak erabiliz.

Wikipedia entziklopedia ere corpus moduan erabili da. Wikipediaren ingelesezko jatorrizko bertsioaz[3] gain, *Simple English*[4] edo ingeles errazean idatzitako bertsioa dago eskuragarri (Yatskar et al., 2010; Woodsend & Lapata, 2011b; Coster & Kauchak, 2011a; Shardlow, 2013b). Frantseserako, Brouwers et al.- ek (2012) frantsesezko wikipedia erabiltzen dute jatorrizko testuak lortzeko eta Vikidia[5], 8-13 urte tartekoei zuzendutako entziklopedia, testu sinpleak lortzeko.

Medikuntza arloan eta zehatzago esanda erradiologiako txostenetan oinarrituz, Kvistab-ek & Velupillaia-k (2013) suedierako corpusa osatu eta aztertu dute.

Corpusak izan gabe, sinplifikazioa egiten duten tresnak eta sistemak garatzeko eta ebaluatzeko datu-multzoak garatu dira. Zhu-k, Bernhardek, & Gurevych-ek (2010) *Wikipedia*tik eta *Simple Wikipedia*tik hartutako 65.133 artikuluz osatu duten datu-multzoa eraiki dute. Artikuluak parekatzeko *language link* edo hizkuntzen arteko esteka jarraituz, Wikimediako *dump files* erabili dute.

Halaber, Wikipedia eta Simple Wikipedian oinarrituta, 137 k jatorrizko/sinplifikatutako esaldi lerratu dituzte Coster-ek & Kauchak-ek (2011b). 2007ko SemEval-eko lexikoaren ordezkapena (*Lexical substitution*) atazarako sortu zen datumultzoa oinarri hartuta anotazio prozesua eta anotatzaileen arteko adostasun emaitzak azaldu dituzte De Belder-ek & Moens-ek (2012).

Bestelako baliabideen artean Petersen-ek & Ostendorf-ek (2007) corpus bat osatu dute eta ikasketa automatikoaren bitartez eta banatu (*splitting*) behar diren esaldien azterketa egin dute, atzerriko hizkuntzen ikasketan laguntzeko. Štajner-ek & Saggion-ek (2013) ere sinplifikatu behar diren esaldien aukeraketa egiteko algoritmoa aurkezten dute. Algoritmo horrek esaldiak dauden bezala mantendu, banatu edo ezabatu behar diren adierazten du testu-genero eta helburu taldearen arabera.

3 Sinplifikazio motak eta metodoak

Esan bezala, HPan egiten diren testuen sinplifikazioak bi mota nagusikoak dira: sinplifikazio sintaktikoa eta sinplifikazio lexikala. Bi horiek TSko azpiatazatzat hartu izan ohi dira.

Sinplifikazio sintaktikoak testu baten konplexutasun gramatikala murriztea du helburu. Horretarako, egitura sintaktiko konplexuak sinpleagoez ordezkatzen dira (Siddharthan, 2006). Azken urte hauetan, lexiko mailako sinplifikazioak ere bere tokia hartu du eta bere helburua hitzen ulergarritasuna areagotzea da, hitz konplexuak edo maiztasun gutxikoak baliokide diren hitz ezagunagoekin, sinonimoekin edo sintagmekin ordezkatuz (Specia, Jauhar, & Mihalcea, 2012). Badaude sinplifikazio sintaktikoa eta lexikala uztartzen dituzten sistemak ere.

Testuak automatikoki sinplifikatzeko metodoei eta teknikei dagokienez, HPko beste atazetan bezala, hiru multzo nagusi bereizi behar dira: eskuzko erregeletan oinarritutakoak, estatistikan eta ikasketa automatikoan (datuetan) oinarritutakoak eta aurreko biak batzen dituzten sistema hibridoak. Azken urteotan estatistikan oinarritutako metodoek lekua irabazi diete eskuz idatzitako erregeletan oinarritutako sistemei, erregeletan oinarritutako sistemak eraikitzeak denbora eskatzen baitu.

Hiru multzo horietaz gain, badira TSa itzulpen prozesu bat bezala ulertzen dutenak, hau da, jatorrizko testuak testu sinplifikatu bihurtzea itzulpen bat balitz bezala kontsideratzen da, bai estatistikoki, bai eskuzko erregeletan oinarritu-

[3] `http://en.wikipedia.org/wiki/Main_Page` (2013ko irailean atzituta)
[4] `http://simple.wikipedia.org/wiki/Main_Page` (2013ko irailean atzituta)
[5] `http://fr.vikidia.org/wiki/Accueil` (2013ko irailean atzituta)

ta. Horrela, jatorrizko testuaren hizkuntza *A iturri hizkuntza*ren pareko izango litzateke eta, testu sinplearen hizkuntza *B helburu hizkuntza*rena.

Sistemen arkitekturei dagokienez, sistema gehienek modulu hauek dituzte: i) analizatzailea: bertan analisia egiten da, sinplifikazioaren aurreprozesua dena; izan ere, testua analizatu gabe ezin da testua sinplifikatu. Analisi hori gauzatzeko, analizatzaile sintaktikoak osagai-ereduan edo dependentzia-ereduan oinarritzen dira. ii) sinplifikatzailea edo transformatzailea: testuak sinplifikatzeaz arduratzen den modulua, bere gain metodo eta teknikak hartzen ditu eta batzuetan iii) testuen kohesioa bermatzen duen modulua.

Hurrengo azpiataletan, sinplifikazio mota bakoitza egiteko erabili diren tekniken eta metodoen laburpena egingo dugu. Azaldutako lan guztiak 1. eta 2. tauletan sailkaturik ikusiko ditugu, 1. taulan sinplifikazio sintaktikoa edo lexikala egiten dituztenak eta 2. taulan bi sinplifikazioak edo bestelako sinplifikazioak egiten dituztenak. Taula horietako lehenengo zutabean, hizkuntzak eta sistemak aurkeztuko dira lanei erreferentzia eginez, eta hurrengo zutabeetan egin duten sinplifikazio mota adieraziko da. Sistemak hizkuntzaren, kronologiaren eta alfabetoaren arabera zerrendatu dira.

3.1 Sinplifikazio sintaktikoa

Sinplifikazio sintaktikoa izan zen hasiera batean testuen sinplifikazio automatikoan garrantzi gehien eman zitzaion sinplifikazio mota edo azpiataza. Hasierako lan gehienak eskuzko erregeletan oinarritutakoak ziren, baina azken urteetan Wikipedia bezalako baliabideei esker, metodo estatistikoak ugaritu egin dira.

3.1.1 Eskuzko erregeletan oinarritutakoak

Sinplifikazio sintaktikoa aurkeztu zen lehendabiziko lanean (Chandrasekar, Doran, & Srinivas, 1996), bi metodo aurkeztu ziren: lehenengoa *chunk*etan oinarrituta eta bigarrena dependentzietan. Hurrengo lanean, erregela horiek ikasketa automatikoaren bitartez erauztea proposatu zuten (Chandrasekar & Srinivas, 1997).

Lehendabiziko lan horretan TSak bi helburu talde zituela ere azaldu zen: gizakiak eta tresnak. Banaketa horri jarraiki azalduko ditugu egin diren lanak.

Gizakiekin hasiz, afasikoei egunkariko testuak irakurterrazagoak egiteko sistema proposatu dute aipatutako PSET proiektuan. Sistema anaforaren tratamenduan oinarritzen da eta 3 modulu ditu: i) anaforaren ebazpena ii) sinplifikazioa eta iii) lehendabiziko moduluak detektatutako izenordainei dagozkien izen sintagmen ordezkapena (Canning & Tait, 1999).

Max-ek (2005; 2006) sinplifikazio motorra testu editore batean integratzen du gaixotasun kognitiboak dituzten pertsonei testuak egokitzen dizkieten autoreei laguntzeko.

Brasilen alfabetatze arazo larriak dituztenez, Brasilgo portugesean linguistikoki zailak diren fenomenoak aztertu ondoren, sinplifikazio proposamenak eman zituzten Aluísio et al.-ek (2008). Sinplifikazioa gauzatzeko erregelak eskuliburu batean deskribatu zituzten (Specia, Aluísio, & Pardo, 2008). Candido et al.-ek (2009) Siddharthan-en (2002) sintaxia sinplifikatzeko hiru mailako arkitekturan oinarrituz zazpi eragiketa azaltzen dituzte. Eragiketa horiek esaldiak banatzea, diskurtso markatzailea aldatzea, ahots pasiboa aktibo bihurtzea, perpausen hurrenkera aldatzea, hurrenkera kanonikoa errespetatzea, adizlagunak galdegai/ez-galdegai bihurtzea eta ez sinplifikatzea dira. Sistema horrek sinplifikatu aurretik testuak konplexuak diren ala ez ebaluatzen du (Gasperin et al., 2009).

Suedierarako garatu duten CogFLUX sistemak (Rybing, Smith, & Silvervarg, 2010) irakurketa errazeko eta testu normaletako corpus azterketa batean oinarritutako 25 transformazio erregela erabiltzen ditu. Erregela horiek bi motatakoak dira: esaldietan sintagmak ezabatu edo ordezkatzen dituztenak eta informazio sintaktikoa gehitzen dutenak.

Haurrek ipuinetako gertaerak hobeto uler ditzaten, ERNESTA *(Enhanced Readability through a Novel Event-based Simplification Tool)* (Barlacchi & Tonelli, 2013) izeneko sistema garatu dute italierarako. Anafora ebatzi ondoren, eta gertaerak kontuan hartuta, esaldiak sinplifikatzen ditu informazio psikolinguistikoan oinarrituz.

Entzumen arazoak dituztenak helburu izanik, bulgarierarako Lozanova et al.-ek (2013) multzo ezberdinetan (anaforaren ebazpena, perpaus mugak, subjektuen berreskurapena. etab.) banatzen diren 23 erregeletako sistema aurkezten dute. Aplikatzen dituzten eragiketak dira esaldiak banatzea, esaldi konplexuen sinplifikazioa, anaforaren ebazpena, subjektuen berreskurapena, perpausen hurrenkera zehaztea eta sintagma osagarrien txertatzea.

Gorrei koreerazko albisteen ulermena errazteko, Chung et al.-ek (2013) esaldi laburrago eta sinpleagoak sortzeaz gain errepresentazio grafikoak erabiltzen dituzte. Sinplifikatzeko esaldiak banatzen dituzte eta argumentuak tokiz aldatzen dituzte dagokien aditzetik gertuago egon daite-

zen.

HPko tresnak helburu dituzten sistemei dagokienez, tresnek informazioa eraginkortasun handiz prozesa dezaten, Beigman Klebanov-ek, Knight-ek, & Marcu-k (2004) *Easy Access Sentences* kontzeptua proposatu zuten. Sortzen diren esaldi horiek jatorrizko testuaren informazioa mantendu behar dute, eta aditz jokatu bakarraz eta entitate batez osatuak izan behar dira. Gainera, esaldi horiek gramatikalak izan behar dira. Horrelako esaldiak erabilita tresnek errazago aurkitu eta prozesatuko dute informazioa.

Azpitituluak egitean esaldiak sinplifikatzeko Daelemans-ek, Höthker-ek, & Sang-ek (2004) bi hurbilpen aurkezten dituzte, bata aurrerago azalduko dugun ikasketa automatikoan oinarritutakoa eta bestea erregeletan oinarritutakoa. Azken honetan nederlandera eta ingelesa sinplifikatzeko erregelak konpilatu dituzte. Bi faseetan egiten dute sintagmen ezabatzea: lehendabizi, erredundanteak diren esaldiak aukeratzen dituzte eta bigarrenik, trinkotasun maila bateko esaldiak ezabatzen dituzte. Ezabatzeko hautagaiak adberbioak, adjektiboak, izen propioak, sintagma preposizionalak, egitura parentetikoak, erlatibozko perpausak, zenbakiak eta denbora-adierazpenak dira.

Biomedikuntzako artikuluen laburpenetako analizatzaile sintaktikoaren emaitzak hobetzeko eta testu horietako informazioa erauzteko algoritmoa garatu dute Jonnalagadda et al.-ek (2009). Algoritmo horrek *Link Grammar* analizatzaile sintaktikoaren bitartez hitz pareen arteko erlazio gramatikal jakinak eta puntuazioa erabiltzen ditu; horrela, esaldiak perpausetan banatzen dituzte. BioSimplify sisteman (Jonnalagadda & Gonzalez, 2010a) izen-sintagmen ordezkatzeak ere gehitzen dituzte.

Biomedikuntzarekin jarraituz, iSimp sistemak (Peng et al., 2012) alor horretako testu zientifikoen laburpenak sinplifikatzen ditu testu meatzaritza egiteko helburuarekin. Patroiak erabiliz koordinazioa, erlatibozko perpausak eta aposizioak tratatzen ditu.

Itzulpen automatikoari laguntzeko Poornima et al.-ek (2011) esaldiak sinplifikatzeko bi metodo proposatzen dituzte: perpaus koordinatuak eta mendeko perpausak banatzea eta erlatibozko izenordaina dutenak sinplifikatzea.

Ahozko hizkuntza ulertzen duten sistemen errendimendua hobetzeko Tur et al.-ek (2011) esaldiak sinplifikatzen dituzte. Beraien helburua sailkatzaile bat denez, eta ez gizakiak, sortzen dituzten esaldiek ez dute zertan gramatikalak izan.

Laburpen automatikoak egiteko, Bawakid-ek & Oussalah-ek (2011) Tregex patroiak erabiliz esaldiak banatzeko erregelak aplikatzen dituzte. Algoritmo baten bitartez esaldiak trinkotzen (*sentence compression*) dituzte laburpena egin aurretik.

Laburpen automatikoak egiten dituzten sistemekin jarraituz, sistema batean integratuta dagoen bi mailako (analisia eta transformazioa) sinplifikatzailea aurkezten dute Silveira Botelho-k & Branco-k (2012) portugeserako. Erlatibozko perpausak, aposizioak eta egitura parentetikoak lantzen dituzte horiek testutik ezabatuz.

Corpus paraleloen hitzak lerratzeko, gakohitzetan oinarritutako zerrendak erabiliz banatzen dituzte esaldiak Srivastava-k & Sanyal-ek (2012). Zerrenda horiek ingeleserako eta Hindirako osatu dituzte.

Helburu taldea irekita uzten duten lanak ere badira. Esaterako, sintaxia sinplifikatzearekin batera kohesioari garrantzia emanez hiru mailako exekuzio-hodia duen arkitektura proposatzen du sintaxia sinplifikatzeko Siddharthan-ek (2006). Sintaxiko hiru maila horiek analisia, transformazioa eta birsorkuntza dira. Azkeneko maila horretan bermatzen da testu berriaren kohesioa.

Sistema irekita uzten dute ere Aranzabek, Díaz de Ilarrazak, & Gonzalez-Diosek (2012) corpus azterketan oinarrituz proposatzen dituzten erregeletarako. Dependentzia-zuhaitzetan oinarrituz esaldi konplexuetatik esaldi sinpleen zuhaitzen transformazioak egiten dituzte (Aranzabe, Díaz de Ilarraza, & Gonzalez-Dios, 2013).

3.1.2 Datuetan oinarritutakoak

Chandrasekar-en, Doran-en, & Srinivas-en (1996) lanari jarraituz, erregelak ikasketa automatikoaren bitartez ikastea proposatzen dute Chandrasekar & Srinivas-ek (1997). Une bakoitzean esaldi bana prozesatzen duen bi mailako arkitektura aurkezten dute: analisia eta transformazioak. Analisiak osagaien eta dependentzien informazioa erabiltzen du, eta transformazioak egiteko, erregelak automatikoki erauztea proposatzen dute domeinuetara errazago egokitzeko.

Azpitituluak egiteko Daelemans-en, Höthker-en, & Sang-en (2004) bigarren hurbilpena (3.1.1 azpiatalean aurkeztu dugu lehenengoa) ikasketa automatikoan oinarritzen da. Esaldiak sinplifikatzeko prozesua hitzen transformazio ataza bezala ulertzen dute; prozesu horretan testuko hitzak kopiatu, ezabatuak edo ordezkatu egiten dira.

Medero-k & Ostendorf-ek (2011) testuen sinplifikazioan egindako aldaketa sintaktikoak identifikatu eta deskribatzen dituen sistema bat aurkezten dute ingelesa irakurterrazagoa izateko.

Estatistika erabiliz ere, Bach et al.-ek (2011) *log-linear* ereduan oinarritutako sistema eraiki dute. Metodo horrek ezaugarri sorta baten gainean lan egiten duen *margin-based discriminative learning* algoritmo bat erabiltzen du. *Stack decoding* algoritmo batekin sinplifikazio hipotesiak sortu eta bilatzen dituzte.

Biomedikuntzako testuetan gertaeren erauzketa helburu izanik, Minard-ek, Ligozat-ek, & Grau-k (2012) ataza horretarako beharrezkoa den informazioarekin geratzeko sinplifikatzen dituzte esaldiak. Horretarako, corpus txiki baten etiketazioan oinarrituta CRF (*Conditional Random Fields*) sailkatzailea erabiltzen dute etiketatzeko, SVMen (*Super Vector Machine*) sarrera izango direnak. Esaldiak etiketatu ahala, corpusa handitzen joaten dira.

Danierarako, corpus handiak erabili gabe, Klerke-k & Søgaard-ek (2013) azpiesaldien aukeraketa eginez esaldi sinpleak lortzen dituzte. Azpiesaldi hautagaiak lortzeko, dependentzia analizatzaile sintaktiko batean oinarritzen diren heuristikoak, esaldiaren osagaiak mantentzeko erabiltzen dituztenak, ezabatzeko ausazko prozedura batekin konbinatzen dituzte. Hautagaien artean aukeratzeko funtzio-galera bat erabiltzen dute.

3.1.3 Hibridoak

Rol semantikoak etiketatzeko, esaldia banatzen dute Vickrey-k & Koller-ek (2008). Horretarako, eskuzko erregela sintaktikoak aplikatzen dituzte eta ondoren ikasketa automatikoaren bidez zein erregela aplikatu erabakitzen dute.

Koordinazioaren fenomenoan zentratuz Evans-ek (2011) esaldi koordinatuen anotazio sakon batean oinarrituta 4 sailkatzaile probatzen ditu. Esaldiak berridazteko corpus azterketan oinarritutako eskuzko erregelak erabiltzen ditu.

Frantseseko egitura sintaktikoak sinplifikatzeko arauak erauztean erabiltzen den eskuzko metodoa osatzeko, Seretan-ek (2012) estatistikoki nabarmentzen diren egitura linguistikoak proposatzen ditu etiketatzaileei laguntzeko.

Frantseserako ere modulu batek corpus azterketan lortutako erregelak programazio linealaren bitartez aukeratzen ditu (Brouwers et al., 2012).

3.2 Sinplifikazio lexikala

Azkeneko urte hauetan lexiko mailako sinplifikazioak ere bere tokia hartu du. Bere helburua hitzen ulergarritasuna areagotzea da, hitz konplexuak edo maiztasun gutxikoak hitz ezagunagoekin, sinonimoekin edo sintagmekin ordezkatuz. Hain garrantzitsua bihurtu da azken urte hauetan sinplifikazio lexikala non eta SemEval lehiaketan ataza bat antolatu zuten (Specia, Jauhar, & Mihalcea, 2012). Ataza horretan 5 sistemek hartu zuten parte eta teknika ezberdinak erabili zituzten (Amoia & Romanelli, 2012; Jauhar & Specia, 2012; Johannsen et al., 2012; Ligozat et al., 2012; Sinha, 2012).

Bi metodo nagusi erabil daitezke lexikoa sinplifikatzeko: hiztegietan eta datu-base lexikaletan oinarrituz ala estatistika erabiliz. Lan gehienek bien konbinazioak erabiltzen dituzte. Gehien erabili diren baliabideak, berriz, *WordNet* (Fellbaum, 2010) datu-base lexikala eta Wikipedia izan dira. Bi horien erabileraren arabera sailkatuko ditugu aurkeztuko ditugu lanak.

WordNet erabiliz, sinplifikazio lexikala aztertzen duten hainbat lan aurki ditzakegu, adibidez De Belder-en, Deschacht-en, & Moens-en (2010) metodoak bi motako hitz alternatiboen multzoak sortzea proposatzen du ordezkatu nahi den hitzari zuzenduak. Lehendabiziko hitz multzoa sinonimoak dituen hiztegi batetik edo WordNetetik lortzen da, eta bigarrena *Latent Words* hizkuntza-eredua erabiliz. Amaierako hitzaren aukeraketa probabilitate bidez kalkulatzen dute hiru baliabide hauetan oinarrituz: psikolinguistikako neurriak dituen datu-basea, testu errazen corpuseko unigramen probabilitatea eta silaba kopurua.

Bott et al.-ek (2012) ere bi mailatan oinarritzen den ordezkapen-eragiketak inplementatu dituzte gaztelaniako lexikoa sinplifikatzeko. Kasu honetan *OpenThesaurus* baliabide lexikalean oinarritzen dira eta ordezkapenerako hautagai hoberena aukeratzen dute hitzei sinpletasun neurri bat eman ondoren. WordNet eta *OpenThesaurus*en arteko konbinazioarekin ere esperimentuak egin dituzte (Saggion, Bott, & Rello, 2013).

WordNeten oinarrituz, Thomas-ek & Anderson-ek (2012) 6 algoritmo probatzen dituzte sinplifikazio lexikal egokiena lortzeko. Algoritmo horiek *Personalized Page Rank* eta informazio maximizazioaren printzipioak erabiltzen dituzte.

Nunes et al.-ek (2013) 4 pausotan banatutako metodoa aurkezten dute: kategoriak etiketatzea, sinonimoak identifikatzea, testuinguruaren maiztasunaren araberako ordezkapena eta esaldia zuzentzea. Erabiltzen dituzten baliabide lexikalak WordNet eta sinonimoen datu-base bat dira. Ordezkapenak egitean hitzen maiztasunak bilatzeko, haurren literaturako liburuekin osatutako hiztegi batean eta web bilatzaileetan oinarritzen dira.

Hiztegiak ere erabiliz, suedierarako Keskisärkkä-k (2012) sinonimoen ordezkapenaren bi-

tartez sinplifikatu du lexikoa. Ordezko sinonimoak aukeratzeko hiru estrategia erabili ditu: hitzen maiztasuna, hitzen luzera eta sinonimia.

Ingelesera itzuliz, medikuntzako lexikoa sinplifikatzeko, Leroy et al.-ek (2013) ordezkapen hitzak proposatzen dituen algoritmo bat testu editore batean integratzeko helburua dute, ondoren aditu batek balidazio edo gainbegiratze urratsean hitzik egokiena aukera dezan gramatikaltasuna bermatzearekin batera. Algoritmo horrek bi pausotan egiten du lan: lehenik, termino zailak identifikatzen ditu *Google Web Corpus*ean[6] n-grama kontaketak eginez eta agerpen gutxi dituztenak hitz zailagoak direla onartuz. Ondoren, termino horien hitz alternatibo errazak proposatzen ditu: sinonimoak eta hiperonimoak WordNetetik, definizioak eta mota semantikoak *Unified Medical Language System*etik[7] (UMLS), eta definizioak ingeles *Wiktionary*tik[8] eta *Simple Wiktionary*tik[9] erauziz, soilik kategoria gramatikal bera duten hitzak proposatzen dituelarik.

Aipatu beharreko beste baliabide bat Wikipedia da. Simple Wikipediako edizioen historiala erabiliz Yatskar et al.-ek (2010) bi hurbilpen proposatzen dituzte: i) edizio eragiketa guztiekin modelo probabilistiko bat sortzen dute eta ii) sinplifikazioak ez diren errebisioak iragazteko metadata erabiltzen dute.

Wikipediarekin jarraituz, Ligozat et al.-ek (2013) lexikoa sinplifikatzean kontuan hartzeko hiru irizpide aurkezten dituzte eta bakoitzari dagokion eredua aurkezten dute, Simple Wikipediako terminoen frekuentziak erabiliz, n-grametan oinarrituz eta kookurrentzien informazioa hartuz.

Ildo berean baina WordNet erabiliz, hitzen testuingurua kontuan hartzen duen bi mailako sistema proposatzen dute Biran-ek, Brody-k, & Elhadad-ek (2011). Lehenengo mailak erregelak erauzten ditu eta bigarrenak sinplifikazioa egiten du. Erregelak erauztean, lehenik, sinplifikatzeko hautagaiak diren edukizko hitz guztientzat (*stop words*, zenbakiak eta puntuazioa baztertuz) bektore bana eraikitzen dute eta, ondoren, hitz bakoitza ordezkatuko duen hautagaiak lortzeko WordNet erabiltzen dute. Sinplifikatzean jatorrizko esaldiaren testuinguruak bi modutan eragiten du: hitz-esaldien antzekotasuna eta testuinguruaren antzekotasuna. Egileen arabera sistema hori zazpi hitz baino gehiagoko esaldientzat da egokia.

Bayes teoreman oinarrituta, Shardlow-ek (2012) hitz bat testuinguru sinple batean agertzen den probabilitatea kalkulatzen du hitzen frekuentziaren gaineko kontaketak Wikipedia eta Simple Wikipediatik hartuz. Hitzak ordezkatzeko erabiltzen duten baliabide lexikala WordNet da. Shardlow-ek (2013a), berriz, sinplifikatzeko hautagaiak diren hitz konplexuak identifikatzeko metodoak azaltzen ditu.

Azpiatal honekin bukatzeko, Kauchak-ek (2013) hizkuntza eredu bat egokitzean sinplifikatu gabeko datuak (datu normalak) erabiltzearen eragina aztertu du eta ondorioztatu du datu normal gehigarriek ingeles errazaren hizkuntza-ereduen performantzia hobetzen dutela.

3.3 Bi sinplifikazioak

Atal honetan sinplifikazio sintaktikoa eta lexikala batera aztertzen dituzten lanak azalduko ditugu. Sistema horiek bi motatako sinplifikazioak egiten dituzte, hau da, testuaren konplexutasun lexikala eta sintaktikoa murrizten dute. Arkitektura orokorraren transformazio moduluan, beraz, sistema horiek bi sinplifikatzaile dituzte, sintaxia tratatzen duena eta lexikoa tratatzen duena.

3.3.1 Eskuzko erregeletan oinarritutakoak

Gizakia helburu duten lanekin hasiz, afasia dutenei albisteak egokitzeko Carroll et al.-en (1998) lanean bi mailatako arkitektura proposatzen da: analizatzailea eta sinplifikatzailea. Analizatzaile moduluak hiru azpimodulu ditu: lexiko etiketatzailea, analizatzaile morfologikoa eta analizatzaile sintaktikoa. Sinplifikatzaileak, berriz, bi atal dauzka: sintaxi sinplifikatzailea eta lexiko sinplifikatzailea. Sintaxi sinplifikatzaileak esaldi pasiboen aktiborako bihurketa, txertatutako perpausen erauzketa eta esaldien banaketa tratatzen ditu. Lexiko sinplifikatzaileak, aldiz, WordNeteko sinonimoak hartu eta Oxfordeko Psikolinguistikako datu-basean galdetuz, sinonimo bakoitzaren frekuentziak lortzen ditu. Ondoren, sinplifikazio mailaren arabera, sinonimo bat edo beste aukeratzen da. Sistema hori afasikoek dituzten fenomenoak kontuan izanda eraiki da.

Japonierarako eta jaiotzetiko entzumen arazoak dituzten pertsonei zuzenduta, 28.000 erregela baino gehiago inplementatu dituzte Inui et al.-ek (2003). Erregela horiek bai parafrasi lexikalak (sinonimoen ordezkapena), bai parafrasi sintaktikoak (banatutako egitura bat kendu, esaldiak banatu, e.a.) egiten dituzte.

[6] `http://catalog.ldc.upenn.edu/LDC2009T25` (2013ko urrian atzituta)

[7] `http://www.nlm.nih.gov/research/umls/` (2013ko irailean atzituta)

[8] `http://en.wiktionary.org/wiki/Wiktionary:Main_Page` (2013ko irailean atzituta)

[9] `http://simple.wiktionary.org/wiki/Main_Page` (2013ko irailean atzituta)

Medikuntzako literatura sinplifikatzeko asmoz, SIMTEXT sistemak (Damay et al., 2006; Ong et al., 2007), lexikoan sinonimoak ordezkatuz eta sintaxian perpausak banatzeko erregelak eta transformazio erregelak erabiliz osasun informazioa eskuragarriago egiten dute.

Biomedikuntzaren domeinuan, baina itzulpen automatikorako testuinguru baten barnean, medikuntza arloko testuak ingelesetik txinerara itzultzen dituen sisteman, eskuzko erregelen bitartez sinplifikatzen dute sintaxia Chen et al.-ek (2012) eta lexikoa terminoak ordezkatuz.

Analfabetismoari aurre egiteko, Brasilgo portugeserako 3.1.1. atalean aurkeztutako sistemari sinplifikazio lexikala gehituta sortu dute SIMPLIFICA sistema (Scarton et al., 2010), testu sinplifikatzailea integratuta duen testu-editorea. Lexikoa sinplifikatzeko, hitz konplexuak eta sinpleak dituzten hiztegietan oinarritzen dira.

Alfabetatze baxuaren arazoari aurka egiteko, Al-Baseet-ek (Al-Subaihin & Al-Khalifa, 2011), arabierarako sistemak, 4 moduluetako arkitektura proposatzen du: konplexutasuna ebaluatzea, lexikoa sinplifikatzea, sintaxia sinplifikatzea eta arabiar hizkuntzaren tipologia dela eta diakritizazioa. Lexiko mailan, sinonimoak bilatzeko WordNet proposatzen dute eta sintaxi mailan, elipsia, subjektuen, objektuen eta aditzen banaketa eta esaldi pasiboak bezalako fenomeno konplexuak lantzea proposatzen dute.

Larrialdien kudeaketaren domeinuetako testuak ulerterrazagoak egiteko, Temnikova-k, Orasan-ek, & Mitkov-ek (2012) hizkuntza kontrolatu bat proposatzen dute eta instrukzioak dituzten testuak sinplifikatzeko 5 motatako erregelak ematen dituzte: orokorrak, formatuaren gainekoak, sintaktikoak, lexikalak eta puntuazioaren gainekoak. Sinplifikatutako testuek egitura jakin batzuk jarraitu behar dituzte larrialdi kasuetan eraginkorrak izan daitezen: izenburua, azpitituluak, baldintzak, egin behar diren ekintzak (instrukzioak), oharrak (azalpenak) eta zerrendatzeak. Izenburuak eta instrukzioak ezinbestean azaldu behar badute ere, beste elementuak aukerakoak dira.

Helburu taldea irekia duen hiru mailako exekuzio-hodia duen arkitektura proposatzen du sintaxirako eta bikoa lexikorako Siddharthan-ek (2002). Sintaxiko hiru maila horiek analisia, transformazioa eta birsorkuntza dira eta lexikokoak parafrasiak eta sinonimoen ordezkapenak.

REGENT sistema (Siddharthan, 2011) dependentzia motatuetan oinarrituz, koordinazioa, mendeko perpausak, erlatibozko perpausak, eta aposizioak sinplifikatzeko eta ahots pasiboa aktibo bihurtzen dituen 63 erregelaz osatzen da. Esaldiak sortzeko bi aukera dauzka: i) transformatutako dependentzia grafoak hitzen hurrenkerarekin eta jatorrizko esaldiaren morfologiarekin lerratzea, *gen-light* eta ii) Stanfordeko dependentziak DSyntS errepresentazioak bihurtu ondoren esaldiak RealPro azalerako errealizatzailearekin sortzea, *gen-heavy*.

3.3.2 Datuetan oinarritutakoak

Tresnei begira, itzulpenaren kalitatea hobetzeko Doi-k & Sumita-k (2004) esaldiak banatzen dituzte. Bi pausotan egiten dute: lehendabizi hautagaiak lortzeko n-grametan oinarritutako hizkuntza-ereduak (NLM, *N-gram Language Model*) erabiltzen dituzte, ondoren hautagaien artean aukeratzeko NLMa eta esaldien antzekotasuna erabiltzen dituzte.

Estatistikan oinarritutako itzulpen automatikoko (SMT, *statistical machine translation*) teknikak oinarri hartuta zuhaitzen transformazioak egiten dituen eredua aurkezten dute Zhu-k, Bernhard-ek, & Gurevych-ek (2010). Eredu horrek 4 eragiketa egiten ditu: perpausak banatzea (*splitting*), hitzak ezabatzea edo erortzen uztea (*dropping*), ordenatzea (*reordering*) eta sintagma/hitz ordezkapena (*substitution*). Eredua iteratiboki entrenatzeko *expectation maximization* (EM) algoritmoa erabiltzen dute eta entrenatzeprozesu hori bizkortzeko hizkuntza bakarreko hitzen mapaketan oinarritutako metodoa aplikatzen dute. Azkenik, dekodifikatzaile bat erabiltzen dute esaldi sinplifikatuak sortzeko estrategia irenskorrak (*greedy*) erabiliz eta hizkuntza-ereduak integratuz.

Coster-ek & Kauchak-ek (2011a) itzulpen automatikoko hurbilpenak erabiltzen dituzte, baina sintagmetan oinarritutako aldaera gehitzen dute, itzulpen automatikoan aldaera horrek hobekuntzak lortu ez dituen arren, TSan sintaxian oinarritutakoak baino emaitza hobeak lortu ditu.

Wubben-ek, van den Bosch-ek, & Krahmerek (2012) ere esaldiak sinplifikatzeko sintagmetan oinarritutako itzulpen automatikoa erabiltzen dute, antzekotasun-ezan (*dissimilarity*) oinarritutako *re-ranking* heuristiko batekin areagotuz eta hizkuntza bakarreko corpus paralelo batean entrenatuz.

Portugueserako Specia-k (2010) ere SMT teknikak erabiltzen ditu. Metodo horrekin emaitza onak lortzen dira batez ere sinplifikazio lexikalean eta berridazketa sinpleetan.

Woodsend-ek & Lapata-k (2011a) jatorrizko eta helburu testuak kontuan izanik, berridazketa konplexuak egiten dituzten erregelak ikasten dituzte. Hurbilpen hori *quasi-synchronous gram-*

Hizkuntzak eta sistemak	**Sinplifikazio sintaktikoa**			**Sinpl.**
	Esk. erregelak	Datuak	Hibridoak	**lexikala**
Ingelesa				
(Chandrasekar, Doran, & Srinivas, 1996)	✔	-	-	-
(Chandrasekar & Srinivas, 1997)	-	✔	-	-
(Beigman Klebanov, Knight, & Marcu, 2004)	✔	-	-	-
(Daelemans, Höthker, & Sang, 2004)	✔	✔	-	-
(Doi & Sumita, 2004)	-	✔	-	-
(Max, 2005; Max, 2006)	✔	-	-	-
(Siddharthan, 2006)	✔	-	-	-
(Vickrey & Koller, 2008)	-	-	✔	-
(Jonnalagadda et al., 2009)	-	✔	-	-
(De Belder, Deschacht, & Moens, 2010)	-	-	-	✔
BioSimplify (Jonnalagadda & Gonzalez, 2010a)	✔	-	-	-
(Yatskar et al., 2010)	-	-	-	✔
(Bawakid & Oussalah, 2011)	✔	-	-	-
(Biran, Brody, & Elhadad, 2011)	-	-	-	✔
(Bach et al., 2011)	-	✔	-	-
(Evans, 2011)	-	-	✔	-
(Medero & Ostendorf, 2011)	-	✔	-	-
(Poornima et al., 2011)	✔	-	-	-
(Tur et al., 2011)	✔	-	-	-
(Amoia & Romanelli, 2012)	-	-	-	✔
(Jauhar & Specia, 2012)	-	-	-	✔
(Johannsen et al., 2012)	-	-	-	✔
(Ligozat et al., 2012)	-	-	-	✔
(Minard, Ligozat, & Grau, 2012)	-	✔	-	-
iSimp (Peng et al., 2012)	✔	-	-	-
(Shardlow, 2012)	-	-	-	✔
(Silveira Botelho & Branco, 2012)	✔	-	-	-
(Sinha, 2012)	-	-	-	✔
(Specia, Jauhar, & Mihalcea, 2012)	-	-	-	✔
(Srivastava & Sanyal, 2012)	✔	-	-	-
(Thomas & Anderson, 2012)	-	-	-	✔
(Nunes et al., 2013)	-	-	-	✔
(Kauchak, 2013)	-	-	-	✔
(Ligozat et al., 2013)	-	-	-	✔
(Shardlow, 2013a)	-	-	-	✔
Portugesa (Br eta Pt)				
PorSimples proiektua (Candido et al., 2009)	✔	-	-	-
(Silveira Botelho & Branco, 2012)	✔	-	-	-
Suediera				
CogFLUX (Rybing, Smith, & Silvervarg, 2010)	✔	-	-	-
(Keskisärkkä, 2012)	-	-	-	✔
Gaztelania				
(Bott et al., 2012; Saggion, Bott, & Rello, 2013)	-	-	-	✔
Frantsesa				
(Brouwers et al., 2012)	-	-	✔	-
(Seretan, 2012)	-	-	✔	-
Euskara				
(Aranzabe, Díaz de Ilarraza, & Gonzalez-Dios, 2012)	✔	-	-	-
Italiera				
ERNESTA (Barlacchi & Tonelli, 2013)	✔	-	-	-
Bulgariera				
(Lozanova et al., 2013)	✔	-	-	-
Koreera				
(Chung et al., 2013)	✔	-	-	-

1 taula: Sinplifikazio sintaktikoa edo lexikala egiten dituzten sistemak eta prototipoak hizkuntzaren, sinplifikazio motaren eta teknikaren arabera sailkatuta

*mar*ean (QG) oinarrituta dago. Programa lineal osoa bezala formulatuta dago eta QGa erabiltzen du berridazketa posible guztien espazioa harrapatzeko. Egileen arabera, eredu hori kontzeptualki sinplea eta konputazionalki efizientea da.

Feblowitz-ek & Kauchak-ek (2013) zuhaitzen transformazioa egiten dute lerratutako corpus baten analisiari probabilistikoki *synchronous tree substitution grammar* (STSG)-ekin ordezkapenak eginez. Hauek dira jarraitzen dituzten pausoak:

Hizkuntzak eta sistemak	Bi sinplifikazioak			Bestelakoak
	Esk. erregelak	Datuak	Hibridoak	
Ingelesa				
PSET, Systar (Carroll et al., 1998; Canning & Tait, 1999)	✔	-	-	-
(Siddharthan, 2002)	✔	-	-	-
SIMTEXT (Damay et al., 2006; Ong et al., 2007)	✔	-	-	-
(De Belder & Moens, 2010)	-	-	✔	-
(Kandula, Curtis, & Zeng-Treitler, 2010)	-	-	-	✔
(Siddharthan, 2010)	-	-	-	✔
(Zhu, Bernhard, & Gurevych, 2010)	-	✔	-	-
(Coster & Kauchak, 2011a)	-	✔	-	-
REGENT (Siddharthan, 2011)	✔	-	-	-
(Woodsend & Lapata, 2011a)	-	✔	-	-
(Chen et al., 2012)	✔	-	-	-
(Temnikova, Orasan, & Mitkov, 2012)	✔	-	-	-
(Wubben, van den Bosch, & Krahmer, 2012)	-	✔	-	-
(Feblowitz & Kauchak, 2013)	-	✔	-	-
(Paetzold & Specia, 2013)	✔	-	-	-
Japoniera				
(Inui et al., 2003)	✔	-	-	-
PorSimples, SIMPLIFICA (Gasperin et al., 2009)	✔	-	-	-
(Specia, 2010)	-	✔	-	-
Arabiera				
(Al-Subaihin & Al-Khalifa, 2011)	✔	-	-	-
Gaztelania				
Simplex proiektua (Bott, Saggion, & Figueroa, 2012)	-	-	✔	-
(Bautista et al., 2012)	-	-	-	✔
(Fajardo et al., 2013)	-	-	-	✔
Daniera				
(Klerke & Søgaard, 2013)	-	✔	-	-

2 taula: Bi sinplifikazioak eta bestelako sinplifikazioak egiten dituzten sistemak eta prototipoak hizkuntzaren, sinplifikazio motaren eta teknikaren arabera sailkatuta

i) gramatika ikasi zuhaitzen osagaiak lerratuz, ii) gramatika osatu informazio lexikala gehituz, iii) egoera finituko transduktore bat aplikatu, entrenatutako *log-linear* eredu batekin *n-best* sinplifikazio hoberenen zerrenda osatzeko eta iv) puntuazio altuena duena aukeratu.

Paetzold-ek & Specia-k (2013) Wikipediarekin ere zuhaitzen transformazioa egiten dituzten erregelak ikasten dituzte, bai sintaxia, bai lexikoa sinplifikatuko dituzten erregela bolumen handiak lortzeko. Hurbilpen horrek 3 osagai nagusi ditu: entrenatzeko modulua, sinplifikazio modulua eta *ranking* modulua. Erregelak ikasteko *Tree Transducer Toolkit* (T3) erabiltzen dute.

3.3.3 Hibridoak

Haurrei zuzendutako sistema garatzeko, lexikoa sinplifikatzeko WordNetetik lortutako hitz alternatiboak hizkuntza-eredu batekin konbinatzen dituzte De Belder-ek & Moens-ek (2010). Sintaxian aposizioak, erlatibozko perpausak, *Prefix subordination* (mendeko perpausak eta txertatzeko elementua gehituz) eta *infix coordination and subordination* (mendeko perpausak eta koordinatuak, soilik esaldiak banatuz) egiturak erregelen bidez sinplifikatzen dituzte eta ondoren programazio lineal osoaren bitartez testuan oro har izan dezaketen eragina kalkulatzen dute sinplifikazio hoberena aukeratuz.

Gizakiei testuak irisgarriagoak egiteko, gaztelaniako Bott-en, Saggion-en, & Figueroa-ren (2012) sistemak hiru pauso ditu: lehenik gramatika batek sinplifikatu behar diren egiturak bilatu eta etiketatzen ditu; bigarrenik, iragazki estatistiko batek berresten du ea benetan etiketatutako esaldiak sinplifikatu behar diren ala ez eta azkenik, manipulazio sintaktikoak egiten dituzte: ezabatzeak, txertatzeak eta nodo sintaktikoak kopiatzea. Erregelen bidez sinplifikatzen dituzten egiturak erlatibozko perpausak, gerundiozko eta partizipiozko egiturak, perpaus koordinatuak eta objektuen koordinazioa dira. Sistema horretan sinplifikazio lexikala eta materia sinplifikatzailea integratzeko asmoa dutela adierazten dute eta Drndarević et al.-en (2013) sintaxia eta lexikoa sinplifikatzen dituzten moduluak ebaluatzen dituzte.

3.4 Bestelako sinplifikazioak

Bestalde, badira ere esaldiak konektoreen erreformulazioen bitartez sinplifikatzen dituzten lanak (Siddharthan, 2010). Lan horretan, ingelesezko

because of lokailua duen esaldi batetik bi esaldi sortzeko erregeletan oinarritutako hiru hurbilpen ezberdinen konparazioa egiten da. Beste mota bateko sinplifikazioa zenbakidun adierazpenena da, esaterako gaztelaniazko *1,9 millones de hogares* kateari zenbaki bidezko adierazpenak eta lexikoa sinplifikatu ondoren *2 millones de casas* lortzea (Bautista et al., 2012). Zenbakidun adierazpenak sinplifikatzeko 5 eragiketa aurkezten dituzte: parentesi arteko zenbakiak ezabatzea, hizkiz daudenak zifraz ematea, kantitate handiak hitzen bitartez adieraztea, biribiltzea eta hamarrekoak ezabatuz biribiltzea.

Amaitzeko sintaktikoarekin batera semantikoa (Kandula, Curtis, & Zeng-Treitler, 2010) egiten duten lanak aurki ditzakegu. Semantikoa egiten duten lanak gutxiago dira HPan eta azalpenak gehitzea izango litzateke horien ataza, adibidez ingelesezko *Humerus* hitzari azalpena parentesi artean gehitzean *Humerus (a part of arm)* lortzea.

4 Ebaluaziorako metodoak

Testuen sinplifikazioa egiten duten sistemen ebaluazioa nola egin komunitatean irekita dagoen galdera bat da. Ataza horrentzat zehazki metrika edo metodoren bat proposatzen ez den bitartean, atal honetan aurkeztuko ditugu orain arte erabili diren metodoak. Metodo erabilienak itzulpen automatikoan erabiltzen diren neurriak, irakurketaren konplexutasun neurriak (*readability measures*) eta erabiltzaileei edo anotatzaileei galdeketak egitea dira. Normalean autoreek metodo bat baino gehiago erabiltzen dituzte sistemak ebaluatzeko.

4.1 Moduluz moduluko ebaluazioa

Hasierako lanetan moduluz modulu ebaluatzen ziren sistemak, analisia egiten zuen moduluari garrantzia emanez. Chandrasekar-ek, Doranek, & Srinivas-ek (1996) egin zuten ebaluazioan beren analisirako bi hurbilpenak (*chunk*etan eta dependentzietan oinarritutakoak) alderatu zituzten.

Siddharthan-ek (2002) moduluz moduluko ebaluazioa egiten du, baina etorkizuneko lanetan bi metodo proposatzen ditu sistemaren errendimendua oro har ebaluatzeko. Bi metodo horiek dira intrinsekoa (*intrinsically*), erabiltzaileen ebaluazioa erabiltzea, eta estrintsekoa (*extrinsically*), bere errendimendua beste sistema batean analizatzaile sintaktikoa, itzultzaile automatikoa) duen eragina neurtzea Jonnalagadda et al.-ek (2009) sinplifikatutako esaldiak eta jatorrizko esaldiak analizatzean egiten duten bezala. Siddharthan-ek (2006) erabiltzaileekin egindako ebaluazioan sinplifikatutako gramatikaltasuna eta antzekotasun semantikoa (ea jatorrizko esaldia eta sinplifikatutako esaldia baliokideak diren) neurtzen ditu.

Lan berriagoen artean, Aranzabek, Díaz de Ilarrazak, & Gonzalez-Diosek (2013) analisia eta esaldiak banatzen dituen modulua ebaluatzen dute sortu duten urre-patroi baten kontra konparatuz.

4.2 Erabiltzaileen bidezkoa

Ebaluatzeko beste metodo bat sistema hori helburu duen erabiltzaileekin ebaluatzea da. Carroll et al.-ek (1998) beraien sistema ebaluatzeko irakurketa esperimentuak egin dituzte ikusmen arazoak ez dituzten afasikoekin. Horrez gain, esaldien ulergarritasuna eta sistemaren baliagarritasuna aztertzeko subjektuak elkarrizketatu dituzte.

Begi mugimenduaren neurtzailea edo *Eyetracker*a erabiliz, hau da, begia eta buruaren arteko mugimendua eta begirada non kokatuta dagoen neurtzen duen tresna, gaztelanian sinplifikazio lexikalaren eragina aztertu da. Alde batetik, Rello et al.-ek (2013) lexikoa sinplifikatzeko bi estrategia probatu dituzte: hitzak sinonimo errazagoekin ordezkatzea eta sinonimo errazagoak eskaintzea hitz konplexuarekin batera. Bestetik, Rello-k, Baeza-Yates-ek, & Saggionek (2013) sinplifikazio lexikalaren eragina gaztelaniazko aditzen parafrasien bitartez (*confiar* eta *tener confianza* bezalako aditz eta kolokazio pareak kontrastatuz) neurtu dute. Rello et al.-ek (2013) adierazpen numerikoa letraz edo hitzen bitartez ematea aztertu zuten. Hiru esperimentu horiek dislexia diagnostikatuta duten pertsonekin egin zituzten, baina datuak kontrastatu ahal izateko kontrol talde bat ere osatu zuten.

Beste sistema batzuk ere erabiltzaileekin ebaluatuak izan dira, baina erabiltzaile horiek ez dira beti sistema garatzean izan zuten helburu taldekoak. De Belder-ek & Moens-ek (2010) ebaluazioa Wikipedia (jatorrizkoa eta sinplea) entziklopediako eta *Literacyworks*[10] web orriko testuekin eta *Amazon's Mechanical Turk crowdsourcing*erako[11] web zerbitzua erabiliz egin zuten. Sinplifikazio lexikoa ebaluatzeko, ordezkapena zuzena zen ala ez galdetzen zuten eta sinplifikazio sintaxikoa ebaluatzeko, esaldiak zuzenak

[10] `http://literacynet.org/cnnsf/` (2013ko urrian atzituta)

[11] `https://www.mturk.com/mturk/welcome` (2013ko irailean atzituta)

ziren ala ez galdetzen zuten.

Sistemak helburu talde jakinik ez duen kasuetan, Yatskar et al.-ek (2010) lexikoa sinplifikatzen duten sistema ezberdinak ebaluatzeko ingeleseko jatorrizko hiztunak eta jatorrizko hiztunak ez direnen anotazioak erabili dituzte. Hala, sistema horietako bakoitzak sortzen dituen ehun hitz-pare eta ausazko beste ehun hitz-pare hartu dituzte eta etiketatzaileei eskatu zaie bikote horietako bakoitzean adierazteko zein den sinpleagoa, zein konplexuagoa, antzekoak diren, erlaziorik gabekoak diren, zalantza sorrarazten dien edo erabakitzeko zaila gertatu zaien.

Mechanical Turk web zerbitzuaren bitartez Leroy et al.-ek (2013) medikuntza arlokoen testuen sinplifikazio lexikala ebaluatzeko bi parametro neurtu dituzte: nabaritutako zailtasuna eta benetako zailtasuna. Lehenengoa neurtzeko, 1-5 bitarteko Likert eskala erabiltzen dute, 1 oso erraza izanik eta 5 oso zaila. Bigarrena neurtzeko, hiru neurri erabiltzen dituzte: ulermena neurtzeko testuarekin batera agertzen diren 5 aukera anitzeko galdera, ikasketarako testurik gabeko beste 7 aukera anitzeko galdera eta informazioaren oroimena neurtzeko, 2 estaldura-galdera libre.

4.3 Ebaluazio automatikoa

Corpusaren kontrako neurketak egitea ebaluatzeko beste metodo bat da. Ebaluazio mota horrek eskatzen duen baliabidea da eskuzko corpus sinplifikatu bat izatea urre-patroi bezala erabiltzeko. Modu honetan sinplifikazio eragiketak eta erregelak ondo aplikatzen diren aztertu ohi da.

Candido et al.-ek (2009) eskuz sinplifikatutako corpus baten kontra eragiketa guztiak bananbanan ebaluatzen dituzte doitasuna, estaldura eta F neurria erabiliz. Horretaz gain, esaldiak zuzen sinplifikatuak izan diren ebaluatzeko eskuzko ebaluazioaren beharra aurreikusten dute esaldiak benetan sinplifikatu diren jakiteko, Aluísio et al.-en (2008) aipatu bezala. Sistema ebaluatzeko, 143 esaldiz osatutako erreferentzia corpusa eraiki dute eta bertan interesgarriak diren egitura sintaktikoak jaso dituzte Gasperin-ek, Maziero-k, & Aluisio-k (2010). Corpuseko esaldiak eskuz sinplifikatuak izan dira sinplifikazio erregelak jarraituz. Perpaus adberbialen sinplifikazio-erregelak ebaluatzeko erreferentzia corpuseko esaldi sinplifikatuen eta jatorrizko esaldien konparazioa bi etiketatzaileek egin zuten eta horiei hiru etiketa jartzea eskatu zieten: 0) esaldi sinplifikatuaren esanahia aldatzen da 1) esaldi sinplifikatuaren esanahia ez da aldatzen baina ez da irakurtzeko errazagoa 2) esaldi sinplifikatuaren esanahia ez da aldatzen eta irakurtzeko errazagoa da. Sinplifikatutako esaldiak ebaluatzeko, berriz, *Levenshtein* (edizio) distantzia erabiliz konparatzen dituzte erreferentzia corpuseko esaldi sinplifikatuekin. Erregelen aplikazio-hurrenkera ere ebaluatzen dute *Levenshtein* distantzia erabiliz.

Bott-ek, Saggion-ek, & Mille-k (2012) erregelak aplikatu diren kasu guztiak kontuan hartuz, erregela testuinguru egokian aplikatu den eta emaitza ona izan duen ebaluatzen dute. Ondoren, doitasuna, F neurria eta erregela aplikatu den maiztasuna kalkulatzen dituzte etiketatu dituzten 262 esaldietan.

Sinplifikazio lexikala ebaluatzeko Cohen-en *kappa* indizean oinarritutako anotatzaileen arteko adostasun neurria proposatzen dute Specia-k, Jauhar-ek, & Mihalcea-k (2012) bai anotatzailen arteko adostasuna kontrastatzeko, bai sistemen arteko konparazioa egiteko urre-patroiaren kontra.

4.4 Itzulpen automatikoko neurriak erabiliz

Testuen sinplifikazio automatikoa itzulpen prozesu bat bezala uler daitekeenez, itzulpen automatikoko sistemak bezala ere ebaluatua izan da arlo horretan erabiltzen diren metrikak aplikatuz.

Daelemans-ek, Höthker-ek, & Sang-ek (2004) ebaluazioak duen zailtasunari erreparatuta eta jakinda oso garestia dela eskuz ebaluatzea, itzulpen automatikoan erabiltzen den BLEU metrika proposatzen dute. Ildo horretatik jarraituz, Zhuk, Bernhard-ek, & Gurevych-ek (2010) ere MTko BLEU and NIST neurriak erabiltzen dituzte beraien sistemaz gain sortu dituzten beste 4 oinlerro sistema ebaluatzeko.

Specia-k (2010) ere neurri horiek erabiltzeaz gain kate-parekatzea eta eskuzko ebaluazioa egiten ditu. Horrela segmentuak egokiak eta naturalak diren eta espero zen sinplifikazioa egiten duten egiaztatzen du.

Coster-ek & Kauchak-ek (2011a) BLUEz gain testuen trinkotasuna neurtzeko erabiltzen diren beste bi neurri erabiltzen ditu: *Simple String Accuracy* (SSA) neurria eta hitzen gainean kalkulatutako F neurria.

Bach et al.-ek (2011) itzulpen eta laburpen automatikoetan erabiltzen diren AveF10, ROUGE-2 eta ROUGE-4 metrikekin batera Flesch-Kincaid konplexutasun neurria erabiltzen dute.

Barlacchi-k & Tonelli-k (2013) MTko neurria den TER erabiltzen dute, eta TER-Plus tresna.

4.5 Irakurketaren konplexutasun neurriak erabiliz

Aurreprozesu bezala erabiltzen diren konplexutasuna ebaluatzen duten sistemak ere erabili izan dira sinplifikatutako testu horien konplexutasun maila baxuagoa den aztertzeko. Adibidez, Siddharthan-ek (2006) Flesch konplexutasun neurriak erabiltzen ditu egunkarietako testuen sinplifikazioak ebaluatzeko.

Lehen aipatu dugun galdeketan metodoarekin batera konplexutasun formulak ere erabiltzen dituzte Drndarević et al.-ek (2013). Konplexutasun neurriak ausaz aukeratutako 100 testuri aplikatzen dizkiete; testu horiek hiru mailatan sinplifikatuak izan dira: lexikala, sintaktikoa eta biak. Galdeketetan, berriz, 25 etiketatzaileri hiru galdera erantzuteko eskatzen diete. Galdera horiek jatorrizko esaldien gramatikaltasunari, sinplifikatutako esaldien gramatikaltasunari eta jatorrizko esaldiaren eta sinplifikatutako esaldiaren esanahien arteko ezberdintasunei buruzkoa dira. Multzo bakoitzarentzat erdiko joera jakiteko batezbestekoa eta mediana kalkulatzen dituzte eta aldakortasunaren indikatzaile bezala frekuentzien distribuzioa.

Temnikova-k, Orasan-ek, & Mitkov-ek (2012) bi motako ebaluazioa egiten dute intrintsekoa eta estrintsekoa. Intrintsekoa konplexutasuna neurtzen duten neurrien bitartez egiten dute. Estrintsekoak, berriz, hiru modutan egiten du: irakurmenezko ulermenean duen eragina, eskuzko itzulpenean eta itzulpen automatikoan duen eragina eta amaierako erabiltzaileen onargarritasuna aztertzen ditu. Hirurak erabiltzaileekin egin zituzten; lehenengoa eta hirugarrena galdeketa bidez, eta bigarrena postedizio esfortzua automatikoki neurtuz (denbora, ikuspuntu teknikoa eta ikuspuntu kognitiboa).

Ebaluaziorako neurriak eta metodoak bateratzeko asmoarekin, eta goian aipatutako esperimentuetan oinarrituta, Temnikova-k & Manevak (2013) C-neurria *(Comprehension Score, C-score)* proposatzen dute testuen ulermena ebaluatzeko. C-neurria testuz testu kalkulatzen da eta hiru formula ditu: sinplea, osoa eta testu tamainakoa. Formula bakoitzak aldagai batzuk hartzen ditu kontuan, eta testuen tamaina ezberdinen arabera aplika daiteke bata edo bestea.

5 Ondorioak

Lan honetan testuen sinplifikazio automatikoaren arloaren egungo egoera aurkeztu dugu HPko ikerketa-lerro honen ikuspegi orokorra emateko. Hizkuntza eta helburu talde ezberdinetarako egin diren lanak aurkeztearekin batera sistemen deskribapenen laburpenak egin ditugu eta sistema horiek ebaluatzeko erabili diren metodoak azaldu ditugu. Orokorrean sistemek jarraitzen duten arkitektura deskribatu dugu. Sistema horiek kronologikoki egiten duten sinplifikazio motaren arabera (sintaktikoa, lexikala edo biak) eta metodo nagusiaren arabera (erregeletan, datuetan oinarrituta edo biak) sailkatu ditugu, 1. eta 2. tauletan laburburbildu dugun moduan. Horrez gain, ikerketa honetatik sortu diren tresnak eta baliabideak ere ezagutzera eman ditugu.

Deskribapen hori kontuan hartuta, testuen sinplifikazioaren bilakaera ikusi ahal izan dugu hasierako lanetatik azkeneko argitalpenetaraino. Nabarmentzekoa da azken urte hauetan ikerketalerro honek izan duen emankortasuna, bai hizkuntza gehiagotara zabaltzen ari delako, bai teknika eta metodo ugari esperimentatzen ari direlako. Ohartu gara ere, hizkuntza asko tipologiaren aldetik oso ezberdinak diren arren, sinplifikatzeko pausoak eta eragiketak oso antzekoak direla; izan ere, hizkuntza batentzat baliagarria dena beste batentzat ere baliagarria gertatu delako.

Amaitzeko, testu sinplifikatuen alorrean gero eta jende gehiago lan egiten ari dela ikus dezakegu, mota horretako testuek eskaintzen dituzten abantailak handiak baitira bai pertsonentzat, bai HPko tresnentzat.

Eskerrak

Itziar Gonzalez-Diosen lana Eusko Jaurlaritzak doktoreak ez diren ikertzaileak prestatzeko Doktoratu Aurreko Programako laguntza bati esker izan da. Ikerketa hau Eusko Jaurlaritzak IXA taldea, A motako ikertalde finkatua (IT344-10) eta MICCINek Hibrido Sint (TIN2010-20218) proiektuei emandako finantziazioagatik gauzatu da.

Erreferentziak

Al-Subaihin, Afnan A. & Hend S. Al-Khalifa. 2011. Al-Baseet: A proposed Simplification Authoring Tool for the Arabic Language. In *International Conference on Communications and Information Technology (ICCIT)*, pages 121–125, March.

Allen, David. 2009. A study of the role of relative clauses in the simplification of news texts for learners of English. *System*, 37(4):585–599.

Aluísio, Sandra M. & Caroline Gasperin. 2010. Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplifica-

tion of Portuguese Texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53. Association for Computational Linguistics.

Aluísio, Sandra M., Lucia Specia, Thiago A.S. Pardo, Erick G. Maziero, & Renata P.M. Fortes. 2008. Towards Brazilian Portuguese automatic text simplification systems. In *Proceedings of the eighth ACM symposium on Document engineering*, DocEng '08, pages 240–248, New York, NY, USA. ACM.

Amoia, Marilisa & Massimo Romanelli. 2012. SB: mmSystem-Using Decompositional Semantics for Lexical Simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 482–486. Association for Computational Linguistics.

Aranzabe, María Jesús, Arantza Díaz de Ilarraza, & Itziar Gonzalez-Dios. 2012. First Approach to Automatic Text Simplification in Basque. In Luz Rello & Horacio Saggion, editors, *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop (LREC 2012)*, pages 1–8.

Aranzabe, María Jesús, Arantza Díaz de Ilarraza, & Itziar Gonzalez-Dios. 2013. Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque. *Procesamiento de Lenguaje Natural*, 50:61–68.

Bach, Nguyen, Qin Gao, Stephan Vogel, & Alex Waibel. 2011. TriS: A Statistical Sentence Simplifier with Log-linear Models and Margin-based Discriminative Training. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 474–482.

Barlacchi, Gianni & Sara Tonelli. 2013. ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian. In *Computational Linguistics and Intelligent Text Processing*. Springer, pages 476–487.

Bautista, Susana, Biljana Drndarevic, Raquel Hervás, Horacio Saggion, & Pablo Gervás. 2012. Análisis de la Simplificación de Expresiones Numéricas en Español mediante un Estudio Empírico. *Linguamática*, 4(2):27–41.

Bautista, Susana, Raquel Hervás, & Pablo Gervás. 2012. Simplificación de textos centrada en la adaptación de expresiones numéricas. In *I Congreso Internacional Universidad y Discapacidad*, Madrid, 11/2012.

Bawakid, Abdullah & Mourad Oussalah. 2011. Sentences Simplification for Automatic summarization. In *Cybernetic Intelligent Systems (CIS), 2011 IEEE 10th International Conference on*, pages 59–64. IEEE.

Beigman Klebanov, Beata, Kevin Knight, & Daniel Marcu. 2004. Text Simplification for Information-Seeking Applications. *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*, pages 735–747.

Bernhard, Delphine, Louis De Viron, Véronique Moriceau, & Xavier Tannier. 2012. Question Generation for French: Collating Parsers and Paraphrasing Questions. *Dialogue and Discourse*, 3(2):43–74.

Biran, Or, Samuel Brody, & Noemie Elhadad. 2011. Putting it Simply: a Context-Aware Approach to Lexical Simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501, Portland, Oregon, USA, June. Association for Computational Linguistics.

Blake, Catherine, Julia Kampov, Andreas K Orphanides, David West, & Cory Lown. 2007. UNC-CH at DUC 2007: Query expansion, lexical simplification and sentence selection strategies for Multi-Document summarization. In *Proceedings of the Document Understanding Conference 2007*.

Bott, Stefan, Luz Rello, Biljana Drndarevic, & Horacio Saggion. 2012. Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In *Proceedings of COLING*, pages 357–373.

Bott, Stefan & Horacio Saggion. 2011. An Unsupervised Alignment Algorithm for Text Simplification Corpus Construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, MTTG '11, pages 20–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bott, Stefan & Horacio Saggion. 2012. Automatic simplification of spanish text for e-accessibility. In *Computers Helping People with Special Needs*. Springer, pages 527–534.

Bott, Stefan, Horacio Saggion, & David Figueroa. 2012. A Hybrid System for Spanish Text Simplification. In *Third Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 75–84, Montreal, Canada.

Bott, Stefan, Horacio Saggion, & Simon Mille. 2012. Text Simplification Tools for Spanish. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, & Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Brouwers, Laetitia, Delphine Bernhard, Anne-Laure Ligozat, & Thomas François. 2012. Simplification syntaxique de phrases pour le français. In *Actes de la Conférence Conjointe JEP-TALN-RECITAL, Montpellier, France*, pages 211–224.

Burstein, Jill. 2009. Opportunities for Natural Language Processing Research in Education. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volumen 5449 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pages 6–27.

Buyko, Ekaterina, Erik Faessler, Joachim Wermter, & Udo Hahn. 2011. Syntactic Simplification and Semantic Enrichment-Trimming Dependency Graphs for Event Extraction. *Computational Intelligence*, 27(4):610–644.

Candido, Jr., Arnaldo, Erick Maziero, Caroline Gasperin, Thiago A. S. Pardo, Lucia Specia, & Sandra M. Aluisio. 2009. Supporting the Adaptation of Texts for Poor Literacy Readers: a Text Simplification Editor for Brazilian Portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, EdAppsNLP '09, pages 34–42, Stroudsburg, PA, USA. Association for Computational Linguistics.

Canning, Yvonne & John Tait. 1999. Syntactic Simplification of Newspaper Text for Aphasic Readers. In *ACM SIGIR'99 Workshop on Customised Information Delivery*, pages 6–11. Citeseer.

Carroll, John, Guido Minnen, Yvonne Canning, Siobhan Devlin, & John Tait. 1998. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10. Citeseer.

Carroll, John, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, & John Tait. 1999. Simplifying text for language-impaired readers. In *Proceedings of EACL*, volumen 99, pages 269–270. Citeseer.

Caseli, Helena M., Tiago F. Pereira, Lucia Specia, Thiago. A. S. Pardo, Caroline. Gasperin, & Sandra Aluísio. 2009. Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *the Proceedings of CICLing*, pages 59–70.

Chandrasekar, Raman, Christine Doran, & Bangalore Srinivas. 1996. Motivations and Methods for Text Simplification. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 1041–1044, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chandrasekar, Raman & Bangalore Srinivas. 1997. Automatic induction of rules for text simplification. *Knowledge-Based Systems*, 10(3):183–190.

Chen, Han-Bin, Hen-Hsen Huang, Hsin-Hsi Chen, & Ching-Ting Tan. 2012. A Simplification-Translation-Restoration Framework for Cross-Domain SMT Applications. In *COLING*, pages 545–560.

Chung, Jin-Woo, Hye-Jin Min, Joonyeob Kim, & Jong C Park. 2013. Enhancing Readability of Web Documents by Text Augmentation for Deaf People. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, WIMS '13, pages 30:1–30:10, New York, NY, USA. ACM.

Coster, William & David Kauchak. 2011a. Learning to Simplify Sentences Using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, MTTG '11, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Coster, William & David Kauchak. 2011b. Simple English Wikipedia: A New Text Simplification Task. In *ACL (Short Papers)'11*, pages 665–669.

Crossley, Scott A, David Allen, & Danielle S McNamara. 2012. Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research*, 16(1):89–108.

Daelemans, Walter, Anja Höthker, & Erick Tjong Kim Sang. 2004. Automatic Sentence Simplification for Subtitling in Dutch and English. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1045–1048.

Damay, Jerwin Jan S., Gerard Jaime D. Lojico, Kimberly Amanda L. Lu, Dex B. Tarantan, & Ethel C. Ong. 2006. SIMTEXT. Text Simplification of Medical Literature. In *3rd National Natural Language Processing Symposium - Building Language Tools and Resources*.

De Belder, Jan, Koen Deschacht, & Marie-Francine Moens. 2010. Lexical Simplification. In *Proceedings of Itec2010: 1st International Conference on Interdisciplinary Research on Technology, Education and Communication*.

De Belder, Jan & Marie-Francine Moens. 2010. Text Simplification for Children. In *Proceedings of the SIGIR workshop on accessible search systems*, pages 19–26.

De Belder, Jan & Marie-Francine Moens. 2012. A Dataset for the Evaluation of Lexical Simplification. *Computational Linguistics and Intelligent Text Processing*, pages 426–437.

Dell'Orletta, Felice, Simonetta Montemagni, & Giulia Venturi. 2011. READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, SLPAT '11, pages 73–83, Stroudsburg, PA, USA. Association for Computational Linguistics.

Devlin, Siobhan & Gary Unthank. 2006. Helping Aphasic People Process Online Information. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, Assets '06, pages 225–226, New York, NY, USA. ACM.

Doi, Takao & Eiichiro Sumita. 2004. Splitting Input Sentence for Machine Translation Using Language Model with Sentence Similarity. In *Proc. of the 20th international conference on Computational Linguistics*.

Drndarević, Biljana, Sanja Štajner, Stefan Bott, Susana Bautista, & Horacio Saggion. 2013. Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules. In *Computational Linguistics and Intelligent Text Processing*. Springer, pages 488–500.

Evans, Richard J. 2011. Comparing methods for the syntactic simplification of sentences in information extraction. *Literary and linguistic computing*, 26(4):371–388.

Fajardo, Inmaculada, Gema Tavares, Vicenta Ávila, & Antonio Ferrer. 2013. Towards text simplification for poor readers with intellectual disability: When do connectives enhance text cohesion? *Research in Developmental Disabilities*, 34(4):1267–1279.

Feblowitz, Dan & David Kauchak. 2013. Sentence Simplification as Tree Transduction. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 1–10, Sofia, Bulgaria, August. Association for Computational Linguistics.

Fellbaum, Christiane. 2010. WordNet. In Roberto Poli, Michael Healy, & Achilles Kameas, editors, *Theory and Applications of Ontology: Computer Applications*. Springer Netherlands, pages 231–243.

Gasperin, Caroline, Erick Maziero, & Sandra M Aluisio. 2010. Challenging Choices for Text Simplification. In *Computational Processing of the Portuguese Language*. Springer, pages 40–50.

Gasperin, Caroline, Erick Maziero, Lucia Specia, Thiago A.S. Pardo, & Sandra M. Aluisio. 2009. Natural language processing for social inclusion: a text simplification architecture for different literacy levels. *the Proceedings of SEMISH-XXXVI Seminário Integrado de Software e Hardware*, pages 387–401.

Hancke, Julia, Sowmya Vajjala, & Detmar Meurers. 2012. Readability Classification for German using lexical, syntactic, and morphological features. In *COLING 2012: Technical Papers*, page 1063–1080.

Inui, Kentaro, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, & Tomoya Iwakura. 2003. Text Simplification for Reading Assistance: A Project Note. In *Proceedings of the second international workshop on Paraphrasing-Volume 16*, pages 9–16. Association for Computational Linguistics.

Jauhar, Sujay Kumar & Lucia Specia. 2012. UOW-SHEF: SimpLex–Lexical Simplicity Ranking based on Contextual and Psycholinguistic Features. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 477–481. Association for Computational Linguistics.

Johannsen, Anders, Héctor Martínez, Sigrid Klerke, & Anders Søgaard. 2012. EMNLP@

CPH: Is frequency all there is to simplicity? In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 408–412. Association for Computational Linguistics.

Jonnalagadda, Siddhartha & Graciela Gonzalez. 2010a. BioSimplify: an open source sentence simplification engine to improve recall in automatic biomedical information extraction. In *AMIA Annual Symposium Proceedings*, volumen 2010, page 351. American Medical Informatics Association.

Jonnalagadda, Siddhartha & Graciela Gonzalez. 2010b. Sentence Simplification Aids Protein-Protein Interaction Extraction. *Arxiv preprint arXiv:1001.4273*.

Jonnalagadda, Siddhartha, Luis Tari, Joerg Hakenberg, Chitta Baral, & Graciela Gonzalez. 2009. Towards Effective Sentence Simplification for Automatic Processing of Biomedical Text. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 177–180. Association for Computational Linguistics.

Kandula, Sasikiran, Dorothy Curtis, & Qing Zeng-Treitler. 2010. A Semantic and Syntactic Text Simplification Tool for Health Content. In *AMIA Annual Symposium Proceedings*, volumen 2010, page 366. American Medical Informatics Association.

Kauchak, David. 2013. Improving Text Simplification Language Modeling Using Unsimplified Text Data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria, August. Association for Computational Linguistics.

Keskisärkkä, Robin. 2012. Automatic Text Simplification via Synonym Replacement. Master's thesis, Linköping.

Klaper, David, Sarah Ebling, & Martin Volk. 2013. Building a German/Simple German Parallel Corpus for Automatic Text Simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19, Sofia, Bulgaria, August. Association for Computational Linguistics.

Klerke, Sigrid & Anders Søgaard. 2012. DSim, a Danish Parallel Corpus for Text Simplification. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, & Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 4015–4018, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Klerke, Sigrid & Anders Søgaard. 2013. Simple, readable sub-sentences. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 142–149, Sofia, Bulgaria, August. Association for Computational Linguistics.

Kvistab, Maria & Sumithra Velupillaia. 2013. Professional Language in Swedish Radiology Reports–Characterization for Patient-Adapted Text Simplification. In *Scandinavian Conference on Health Informatics 2013*, page 55.

Lal, Partha & Stefan Rüger. 2002. Extract-based Summarization with Simplification. In *Proceedings of the Workshop on Text Summarization at DUC 2002 In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization*.

Leroy, Gondy, James E Endicott, David Kauchak, Obay Mouradi, & Melissa Just. 2013. User Evaluation of the Effects of a Text Simplification Algorithm Using Term Familiarity on Perception, Understanding, Learning, and Information Retention. *Journal of medical Internet research*, 15(7).

Ligozat, Anne-Laure, Anne Garcia-Fernandez, Cyril Grouin, & Delphine Bernhard. 2012. ANNLOR: A Naïve Notation-system for Lexical Outputs Ranking. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 487–492. Association for Computational Linguistics.

Ligozat, Anne-Laure, Cyril Grouin, Anne Garcia-Fernandez, & Delphine Bernhard. 2013. Approches à base de fréquences pour la simplification lexicale. In *Actes TALN-RÉCITAL 2013*, pages 493–506. ATALA.

Lozanova, Slavina, Ivelina Stoyanova, Svetlozara Leseva, Svetla Koeva, & Boian Savtchev. 2013. Text Modification for Bulgarian Sign Language Users. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 39–48, Sofia, Bulgaria, August. Association for Computational Linguistics.

Max, Aurélien. 2005. Simplification interactive pour la production de textes adaptés aux personnes souffrant de troubles de la compréhension. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*.

Max, Aurélien. 2006. Writing for Language-Impaired Readers. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volumen 3878 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pages 567–570.

Medero, Julie & Mari Ostendorf. 2011. Identifying Targets for Syntactic Simplification. In *Proceedings of the SLaTE 2011 workshop*, pages 69–72.

Minard, Anne-Lyse, Anne-Laure Ligozat, & Brigitte Grau. 2012. Simplification de phrases pour l'extraction de relations. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 1–14, Grenoble, France, June. ATALA/AFCP.

Nunes, Bernardo Pereira, Ricardo Kawase, Patrick Siehndel, Marco A. Casanova, & Stefan Dietze. 2013. As Simple as It Gets - A Sentence Simplifier for Different Learning Levels and Contexts. In *Advanced Learning Technologies (ICALT), 2013 IEEE 13th International Conference on*, pages 128–132.

Oh, Sun Young. 2001. Two Types of Input Modification and EFL Reading omprehension: Simplification Versus Elaboration. *TESOL Quarterly*, 35(1):69–96.

Ong, Ethel, Jerwin Damay, Gerard Lojico, Kimberly Lu, & Dex Tarantan. 2007. Simplifying Text in Medical Literature. *J. Research in Science Computing and Eng*, 4(1):37–47.

Paetzold, Gustavo H & Lucia Specia. 2013. Text Simplification as Tree Transduction. In Sociedade Brasileira de Computaçao, editor, *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pages 116–125.

Peng, Yifan, Catalina O Tudor, Manabu Torii, Cathy H Wu, & K Vijay-Shanker. 2012. iSimp: A Sentence Simplification System for Biomedical Text. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, pages 1–6. IEEE.

Petersen, Sarah E. 2007. *Natural Language Processing Tools for Reading Level Assessment and Text Simplification for Bilingual Education*. Ph.D. thesis, Citeseer.

Petersen, Sarah E & Mari Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. In *In Proceedings of Workshop on Speech and Language Technology for Education. SLaTE*, pages 69–72. Citeseer.

Poornima, C., V. Dhanalakshmi, K.M. Anand, & KP Soman. 2011. Rule based Sentence Simplification for English to Tamil Machine Translation System. *International Journal of Computer Applications*, 25(8):38–42.

Rello, Luz, Ricardo Baeza-Yates, Stefan Bott, & Horacio Saggion. 2013. Simplify or Help? Text Simplification Strategies for People with Dyslexia. *Proc. W4A*, 13.

Rello, Luz, Ricardo Baeza-Yates, & Horacio Saggion. 2013. The Impact of Lexical Simplification by Verbal Paraphrases for People with and without Dyslexia. In *Computational Linguistics and Intelligent Text Processing*. Springer, pages 501–512.

Rello, Luz, Susana Bautista, Ricardo Baeza-Yates, Pablo Gervás, Raquel Hervás, & Horacio Saggion. 2013. One Half or 50%? An Eye-Tracking Study of Number Representation Readability. In *Proc. INTERACT*, volumen 13, pages 1–17.

Rybing, Jonas, Christian Smith, & Annika Silvervarg. 2010. Towards a Rule Based System for Automatic Simplification of texts. In *The Third Swedish Language Technology Conference (SLTC 2010)*, pages 17–18.

Saggion, Horacio, Stefan Bott, & Luz Rello. 2013. Comparing Resources for Spanish Lexical Simplification. In *SLSP 2013: 1st International Conference on Statistical Language and Speech Processing*, pages 1–12. Springer.

Saggion, Horacio, Elena Gómez-Martínez, Esteban Etayo, Alberto Anula, & Lorena Bourg. 2011. Text Simplification in Simplext: Making Text More Accessible. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 47:341–342.

Scarton, Carolina, Matheus de Oliveira, Arnaldo Candido Jr, Caroline Gasperin, & Sandra Maria Aluísio. 2010. SIMPLIFICA: a

tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments. In *Proceedings of the NAACL HLT 2010 Demonstration Session*, pages 41–44. Association for Computational Linguistics.

Seretan, Violeta. 2012. Acquisition of Syntactic Simplification Rules for French. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, & Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Shardlow, Matthew. 2012. Bayesian Lexical Simplification. Txosten teknikoa, Short Taster Research Project. The University of Manchester.

Shardlow, Matthew. 2013a. A Comparison of Techniques to Automatically Identify Complex Words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109, Sofia, Bulgaria, August. Association for Computational Linguistics.

Shardlow, Matthew. 2013b. The CW Corpus: A New Resource for Evaluating the Identification of Complex Words. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 69–77, Sofia, Bulgaria, August. Association for Computational Linguistics.

Siddharthan, Advaith. 2002. An Architecture for a Text Simplification System. In *Proceedings of the Language Engineering Conference (LEC'02)*, pages 64–71, Washington, DC, USA. IEEE Computer Society.

Siddharthan, Advaith. 2006. Syntactic Simplification and Text Cohesion. *Research on Language & Computation*, 4(1):77–109.

Siddharthan, Advaith. 2010. Complex Lexico-Syntactic Reformulation of Sentences using Typed Dependency Representations . In *Proceedings of the 6th International Natural Language Generation Conference*, pages 125–133. Association for Computational Linguistics.

Siddharthan, Advaith. 2011. Text Simplification using Typed Dependencies: A Comparison of the Robustness of Different Generation Strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 2–11. Association for Computational Linguistics.

Siddharthan, Advaith, Ani Nenkova, & Kathleen McKeown. 2004. Syntactic Simplification for Improving Content Selection in Multi-Document Summarization. In *Proceedings of the 20th international conference on Computational Linguistics*, page 896. Association for Computational Linguistics.

Silveira Botelho, Sara & António Branco. 2012. Enhancing Multi-document Summaries with Sentence Simplification. In *ICAI 2012: International Conference on Artificial Intelligence*.

Simensen, Aud Marit. 1987. Adapted Readers: How are they Adapted. *Reading in a Foreign Language*, 4(1):41–57.

Sinha, Ravi. 2012. UNT-SimpRank: Systems for Lexical Simplification Ranking. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 493–496. Association for Computational Linguistics.

Specia, Lucia. 2010. Translating from Complex to Simplified Sentences. *Computational Processing of the Portuguese Language*, pages 30–39.

Specia, Lucia, Sandra M. Aluísio, & Thiago A.S. Pardo. 2008. Manual de Simplificação Sintática para o Português. Txosten teknikoa NILC-TR-08-06, São Carlos-SP.

Specia, Lucia, Sujay Kumar Jauhar, & Rada Mihalcea. 2012. Semeval-2012 Task 1: English Lexical Simplification. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, pages 347–355.

Srivastava, Jyoti & Sudip Sanyal. 2012. Segmenting Long Sentence Pairs to Improve Word Alignment in English-Hindi Parallel Corpora. In *Advances in Natural Language Processing*. Springer, pages 97–107.

Štajner, Sanja & Horacio Saggion. 2013. Adapting Text Simplification Decisions to Different Text Genres and Target Users. *Procesamiento del Lenguaje Natural*, 51:135–142.

Temnikova, Irina & Galina Maneva. 2013. The C-Score – Proposing a Reading Comprehension Metrics as a Common Evaluation Measure for Text Simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 20–29, Sofia, Bulgaria, August. Association for Computational Linguistics.

Temnikova, Irina, Constantin Orasan, & Ruslan Mitkov. 2012. CLCM - A Linguistic Resource for Effective Simplification of Instructions in the Crisis Management Domain and its Evaluations. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, & Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3007–3014, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Thomas, S. Rebecca & Sven Anderson. 2012. WordNet-Based Lexical Simplification of a Document. In *Empirical Methods in Natural Language Processing*, pages 80–88.

Tur, Gokhan, Dilek Hakkani-Tur, Larry Heck, & S. Parthasarathy. 2011. Sentence Simplification for Spoken Language Understanding. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5628–5631. IEEE.

Vanderwende, Lucy, Hisami Suzuki, Chris Brockett, & Ani Nenkova. 2007. Beyond SumBasic: Task-focused Summarization with Sentence Simplification and Lexical Expansion. *Information Processing & Management*, 43(6):1606–1618.

Vickrey, David & Daphne Koller. 2008. Sentence Simplification for Semantic Role Labeling. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-2008: HLT)*, pages 344–352.

Woodsend, Kristian & Mirella Lapata. 2011a. Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Stroudsburg, PA, USA. Association for Computational Linguistics.

Woodsend, Kristian & Mirella Lapata. 2011b. WikiSimple: Automatic Simplification of Wikipedia Articles. In *Proceedings of the TwentyFifth AAAI Conference on Artificial Intelligence*, pages 927–932.

Wubben, Sander, Antal van den Bosch, & Emiel Krahmer. 2012. Sentence Simplification by Monolingual Machine Translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 1015–1024. Association for Computational Linguistics.

Yatskar, Mark, Bo Pang, Cristian Danescu-Niculescu-Mizil, & Lillian Lee. 2010. For the sake of simplicity: unsupervised extraction of lexical simplifications from Wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 365–368, Stroudsburg, PA, USA. Association for Computational Linguistics.

Young, Dolly N. 1999. Linguistic Simplification of SL Reading Material: Effective Instructional Practice? *The Modern Language Journal*, 83(3):350–366.

Zhu, Zhemin, Delphine Bernhard, & Iryna Gurevych. 2010. A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of The 23rd International Conference on Computational Linguistics*, pages 1353–1361. Association for Computational Linguistics.