Idiomatikotasunaren karakterizazio automatikoa: `izena+aditza` konbinazioak

Doktoregaia: Antton Gurrutxaga Hernaiz

Zuzendariak: Iñaki Alegria Loinaz, Xabier Artola Zubillaga

Lengoaia eta Sistema Informatikoak Saila
Euskal Herriko Unibertsitatea / Universidad del País Vasco

2014ko uztailaren 21a

urratsak egiten, pausoak ematen...

eu *urrats* = en *step*

urratsak egiten, pausoak ematen...

eu *urrats* = en *step*

$\Longrightarrow$

en *to* ***do/make/give****? steps*

urratsak egiten, pausoak ematen...

Martin Luther King

to pull sb's leg = adarra jo

I'm pulling your leg!

≠

Adarra jotzen ari naiz!

to pull sb's leg = adarra jo

≠

I'm pulling your leg!

Adarra jotzen ari naiz!

erabilera "idiomatikoa"

to put sb's leg = *norbaiti adarra jo*

(to tease, to joke) *(burla egin, iseka egin)*

Unitate fraseologikoez ari gara!

Unitate fraseologikoez ari gara!

hiztunok erabiltzen ditugun hitz-konbinazio "preferentzialak" edo unitate "aurrefabrikatuak"

konbinazio idiosinkratikoak, sistemaren gramatika-arauen zein semantikaren arabera soilik ezin aurreikus edo azal daitezkeenak

ezaugarri semantiko, morfosintaktiko eta lexikal bereziak dituztenak

Testuingurua

- Euskaraz, HAULen errepresentazioa/identifikazioa, termino-erauzketa eta entitateen arloa landu dira, batez ere

  $\rightarrow$ **Euskarazko UFak erauzteko eta karakterizatzeko teknologia garatu beharra**

- Testuinguru orokorra: fraseologia konputazionalean, UFak karakterizatzeko ohiko metodoa osagaien agerkidetzan oinarritua da

  $\rightarrow$ **Azkenaldiko teknika berriak esperimentatzeko eta garatzeko premia**

Helburuak

Helburu nagusia

Corpusetatik `izena+aditza` osaerako unitate fraseologikoak (UFak) automatikoki eskuratzeko eta haien idiomatikotasunaren arabera karakterizatzeko teknikak ikertzea, garatzea eta konbinatzea

Helburuak

Helburu zehatzak

1. UFen **idiomatikotasuna**ren definizio operatiboa: haren osagai diren propietate neurgarriak zehaztea eta UFen sailkapen-eredua lantzea
2. UFen erauzketa eta karakterizazio automatikoaren **ebaluazio-metodologia** zehaztea, eta behar diren baliabideak eratzea
3. Ikergaitzat ditugun euskarazko **`izena+aditza` osaerako unitate**en ezaugarriak deskribatzea
4. Idiomatikotasunaren propietate bakoitza neurtzeko **lan esperimentala**, eta horien emaitzak konbinatzea
5. Emaitzak analizatzea eta **ondorioak** ateratzea, batez ere, euskarazko **baliabide lexikal konputazionalak** eratzeari eta **hiztegigintza**ri begira

Helburuak

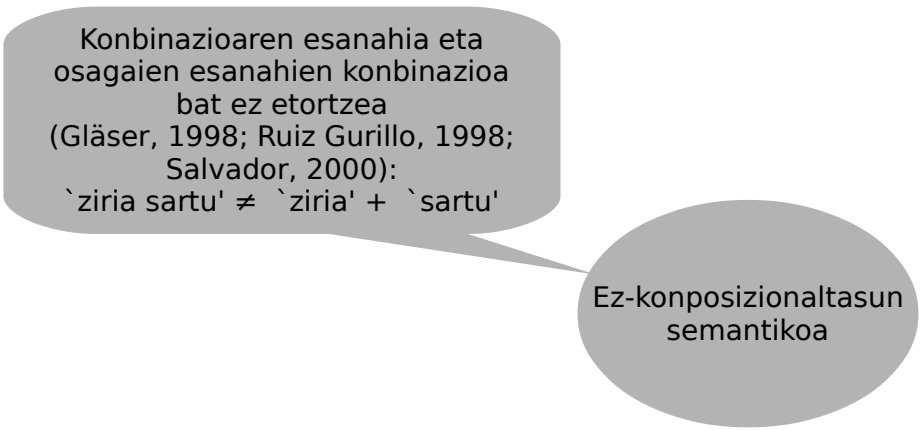
Galderak

- Zenbaterainoko korrelazioa dago idiomatikotasun-mailaren eta haren propietate bakoitzaren neurketen artean?
- Zenbateraino datoz bat UFen propietateen ebidentzia enpirikoak teoria fraseologikoak UFetarako oro har zein UF-kategoria bakoitzerako aurresaten duenarekin?
- Hobetu daiteke UFen karakterizazioa idiomatikotasunaren propietate bakunen kuantifikazioaren emaitzak konbinatuz?

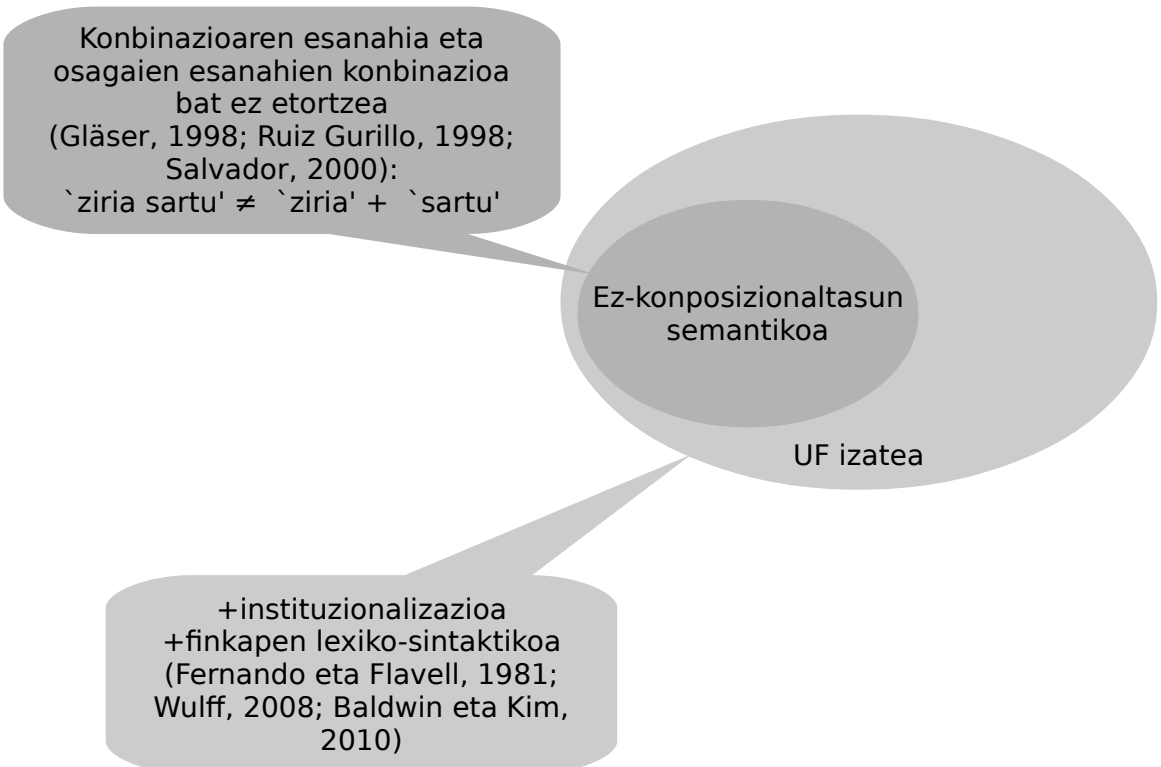
Aurkezpenaren eskema

Aurkezpenaren eskema

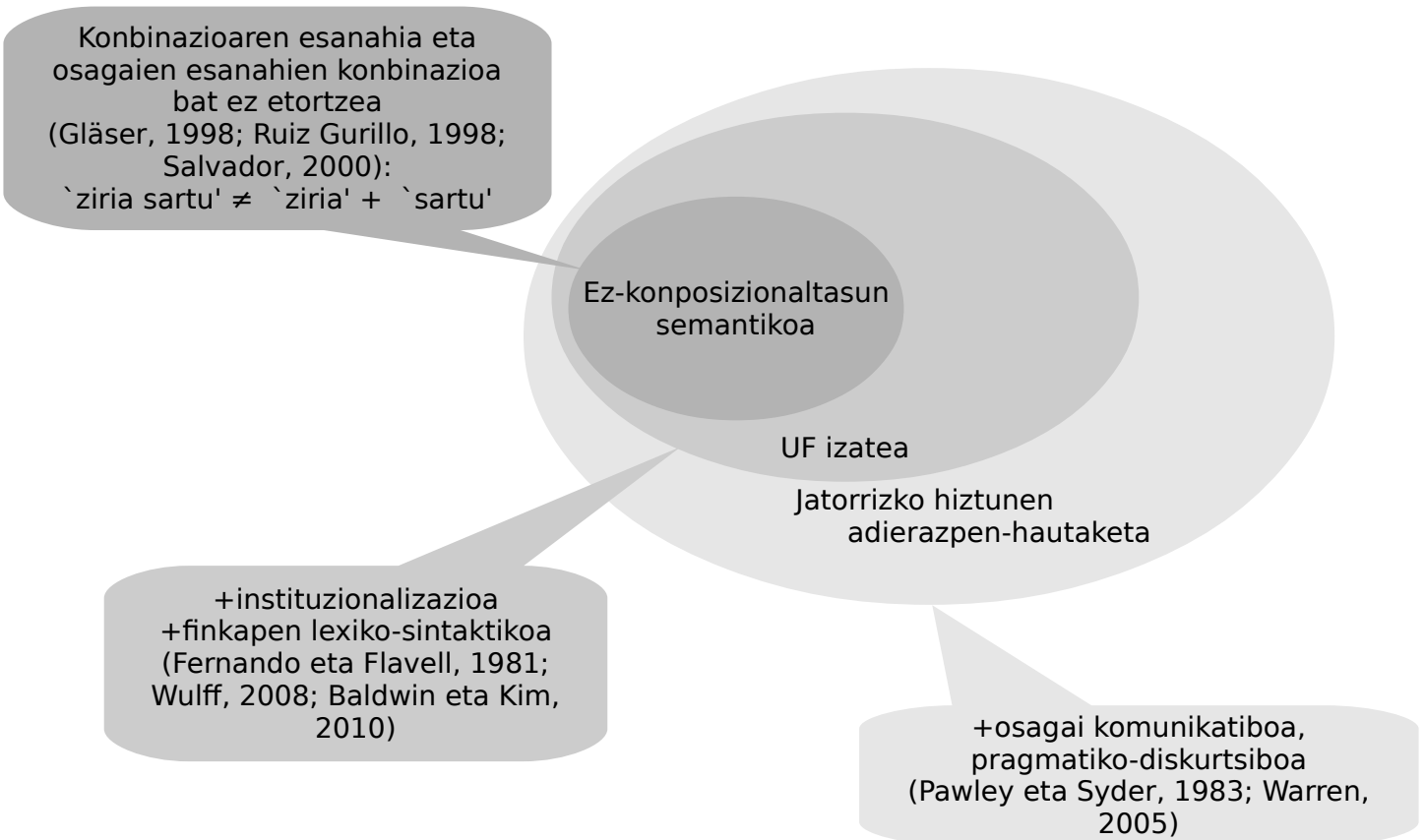
Idiomatikotasunaren definizioak



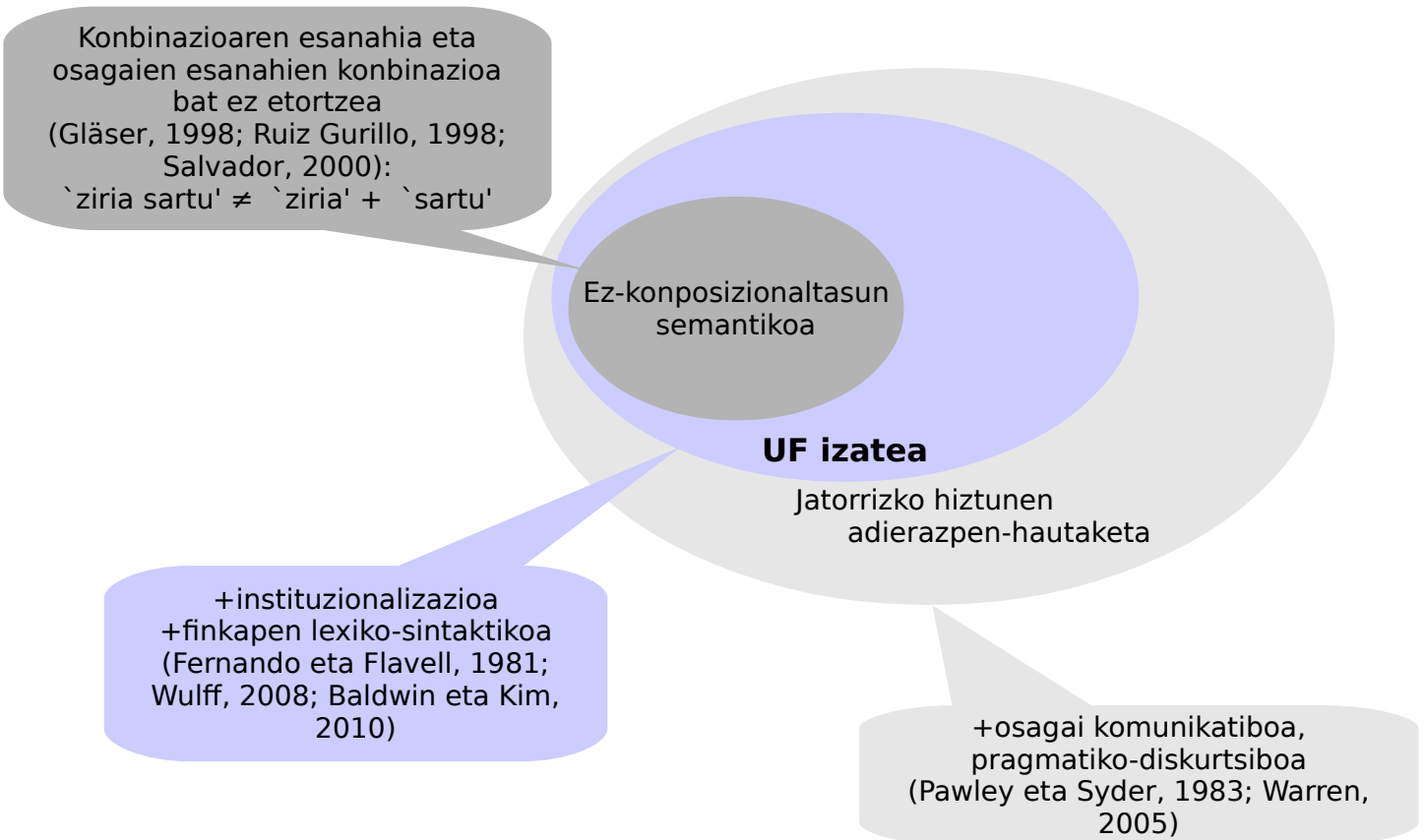
Idiomatikotasunaren definizioak



Idiomatikotasunaren definizioak



Idiomatikotasunaren definizioak



Idiomatikotasunaren definizioa

Gure hautua

- Idiomatikotasuna konbinazio bat UF izatea determinatzen duen propietatea da, muineko ezaugarritzat **idiosinkrasia** duena
- Idiomatikotasuna konplexua eta graduala da, eta bere barnean zenbait propietate hartzen ditu: **instituzionalizazioa**, **ez-konposizionaltasun semantikoa** (osoa edo partziala), eta **finkapena** (morfosintaktikoa zein lexikala)

Idiomatikotasunaren osagaiak

- **Instituzionalizazioa**: adiera hertsian, idiosinkrasia estatistikoa (konbinazioa ausazkoa ez izatea)
  - behatutako maiztasuna > osagaiak ausaz konbinatuz itxaron daitekeen maiztasuna
- Ez-konposizionaltasuna: konbinazioaren esanahia $\neq$ osagaien esanahien konbinazioa. Konposizionaltasun-mailak:
  - ‘adarra jo’ $\neq$ ‘adarra’ + ‘jo’
  - ‘zubiak eraiki’ $\leftarrow\!-$ ‘zubiak’ + ‘eraiki’
  - ‘atentzioa eman’ = ‘atentzioa’ + ‘$\text{eman}_{\text{atentzio}}$’
- Finkapen morfosintaktikoa: aldakuntza batzuk izateko murriztapenak
  - **adarra ipini / adarrak ipini*; **adar ederra jo nion / adarra ederki jo nion*
- Finkapen lexikala: osagaiak sinonimoez ordezkatzeko murriztapenak
  - **meza aditu / meza entzun*; *hanka sartu $\neq$ zangoa sartu*

Idiomatikotasunaren osagaiak

- Instituzionalizazioa: adiera hertsian, idiosinkrasia estatistikoa (konbinazioa ausazkoa ez izatea)
  - behatutako maiztasuna > osagaiak ausaz konbinatuz itxaron daitekeen maiztasuna
- **Ez-konposizionaltasuna**: konbinazioaren esanahia $\neq$ osagaien esanahien konbinazioa. Konposizionaltasun-mailak:
  - ‘adarra jo’ $\neq$ ‘adarra’ + ‘jo’
  - ‘zubiak eraiki’ $\leftarrow$ ‘zubiak’ + ‘eraiki’
  - ‘atentzioa eman’ = ‘atentzioa’ + ‘$\text{eman}_{\text{atentzio}}$’
- Finkapen morfosintaktikoa: aldakuntza batzuk izateko murriztapenak
  - **adarra ipini / adarrak ipini*; **adar ederra jo nion / adarra ederki jo nion*
- Finkapen lexikala: osagaiak sinonimoez ordezkatzeko murriztapenak
  - **meza aditu / meza entzun*; *hanka sartu $\neq$ zangoa sartu*

Idiomatikotasunaren osagaiak

- Instituzionalizazioa: adiera hertsian, idiosinkrasia estatistikoa (konbinazioa ausazkoa ez izatea)
  - behatutako maiztasuna > osagaiak ausaz konbinatuz itxaron daitekeen maiztasuna
- Ez-konposizionaltasuna: konbinazioaren esanahia $\neq$ osagaien esanahien konbinazioa. Konposizionaltasun-mailak:
  - ‘adarra jo’ $\neq$ ‘adarra’ + ‘jo’
  - ‘zubiak eraiki’ $\leftarrow\!-$ ‘zubiak’ + ‘eraiki’
  - ‘atentzioa eman’ = ‘atentzioa’ + ‘$\text{eman}_{\text{atentzio}}$’
- **Finkapen morfosintaktikoa**: aldakuntza batzuk izateko murriztapenak
  - **adarra ipini / adarrak ipini*; **adar ederra jo nion / adarra ederki jo nion*
- Finkapen lexikala: osagaiak sinonimoez ordezkatzeko murriztapenak
  - **meza aditu / meza entzun*; *hanka sartu $\neq$ zangoa sartu*

Idiomatikotasunaren osagaiak

- Instituzionalizazioa: adiera hertsian, idiosinkrasia estatistikoa (konbinazioa ausazkoa ez izatea)
  - behatutako maiztasuna > osagaiak ausaz konbinatuz itxaron daitekeen maiztasuna
- Ez-konposizionaltasuna: konbinazioaren esanahia $\neq$ osagaien esanahien konbinazioa. Konposizionaltasun-mailak:
  - ‘adarra jo’ $\neq$ ‘adarra’ + ‘jo’
  - ‘zubiak eraiki’ $\leftarrow\!-$ ‘zubiak’ + ‘eraiki’
  - ‘atentzioa eman’ = ‘atentzioa’ + ‘$\text{eman}_{\text{atentzio}}$’
- Finkapen morfosintaktikoa: aldakuntza batzuk izateko murriztapenak
  - **adarra ipini / adarrak ipini*; **adar ederra jo nion / adarra ederki jo nion*
- **Finkapen lexikala**: osagaiak sinonimoez ordezkatzeko murriztapenak
  - **meza aditu / meza entzun*; *hanka sartu $\neq$ zangoa sartu*

Idiomatikotasun-continuuma

- Idiomatikotasunaren propietateak hein desberdinean konbinatzen dira, eta UFak sailkatzeko irizpideak ezartzearen emaitza graduazio moduko continuum bat da (Sinclair, 1996; Ruiz Gurillo, 1998; Bannard et al., 2003; Katz eta Giesbrecht, 2006; Wulff, 2008)

Idiomatikotasun-continuuma

- Idiomatikotasunaren propietateak hein desberdinean konbinatzen dira, eta UFak sailkatzeko irizpideak ezartzearen emaitza graduazio moduko continuum bat da (Sinclair, 1996; Ruiz Gurillo, 1998; Bannard et al., 2003; Katz eta Giesbrecht, 2006; Wulff, 2008)

  BAINA

- Continuum horretan eremu edo zona desberdinak proposatu dira:

  'Collocations' and 'idioms' represent two large and amorphous subgroups of FEIs on a continuum. (Moon, 1998)

Sintagma-erako UFak sailkatzeko proposamen batzuk

Author	Opaque, invariable unit	Partially motivated unit	Phraseologically bound unit
Vinogradov (1947)	Phraseological function	Phraseological unity	Phraseological communication
Amosova (1963)	Idiom	Idiom (nor differentiated)	Phraseme, or phraseoloid
Cowie (1981)	Pure idiom	Figurative idiom	Restricted collocation
Mel'čuk (1988)	Idiom	Idiom (nor differentiated)	Collocation
Gläser (1988)	Idiom	Idiom (nor differentiated)	Restricted collocation
Howarth (1996)	Pure idiom	Figurative idiom	Restricted collocation
Corpas Pastor (1996)	Locución	Locución	Colocación
Burger (1998)	Idiom	Partial idiom	Collocation

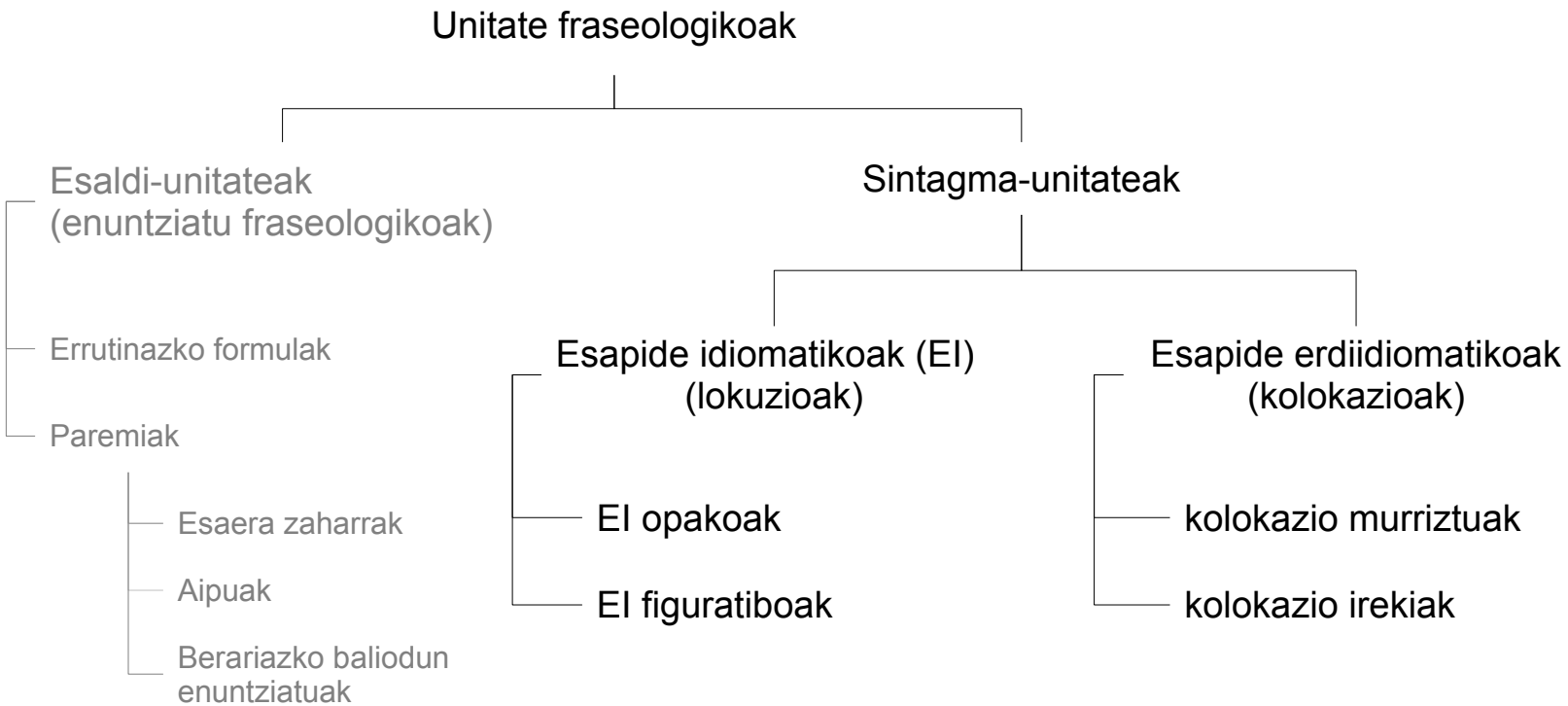
Sintagma-erako UFak sailkatzeko proposamen batzuk

Author	Opaque, invariable unit	Partially motivated unit	Phraseologically bound unit
Vinogradov (1947)	Phraseological function	Phraseological unity	Phraseological communication
Amosova (1963)	Idiom	Idiom (nor differentiated)	Phraseme, or phraseoloid
Cowie (1981)	Pure idiom	**Figurative idiom**	Restricted collocation
Mel'čuk (1988)	Idiom	Idiom (nor differentiated)	Collocation
Gläser (1988)	Idiom	Idiom (nor differentiated)	Restricted collocation
Howarth (1996)	Pure idiom	**Figurative idiom**	Restricted collocation
Corpas Pastor (1996)	Locución	Locución	Colocación
Burger (1998)	Idiom	**Partial idiom**	Collocation

Sintagma-erako UFak sailkatzeko proposamen batzuk

Author	Opaque, invariable unit	Partially motivated unit	Phraseologically bound unit
Vinogradov (1947)	Phraseological function	Phraseological unity	Phraseological communication
Amosova (1963)	Idiom	Idiom (nor differentiated)	Phraseme, or phraseoloid
Cowie (1981)	Pure idiom	Figurative idiom	**Restricted collocation**
Mel'čuk (1988)	Idiom	Idiom (nor differentiated)	Collocation
Gläser (1988)	Idiom	Idiom (nor differentiated)	**Restricted collocation**
Howarth (1996)	Pure idiom	Figurative idiom	**Restricted collocation**
Corpas Pastor (1996)	Locución	Locución	Colocación
Burger (1998)	Idiom	Partial idiom	Collocation

UFen sailkapen-eredua



Non da esapide idiomatikoen eta kolokazioen arteko muga, edo zertan da bien arteko alde zehatza?

Esapide idiomatikoak (lokuzioak)

Ezaugarri nagusiak

- Idiosinkrasia estatistikoa: oro har, **maiztasun txikiagokoak kolokazioak baino** (Moon, 1998)
- Ez-konposizionaltasuna: erabatekoa edo handia
  - Opakoak (deskodetzeko arazoak): *adarra jo, ziria sartu, erreka jo, gorriak ikusi*
  - Figuratiboak (deskodetzeko aukera; deskonposagarriak): *burua hautsi, garunak urtu, zubiak eraiki*
- Finkapen morfosintaktikoa: aski handia; batzuk aldaezinak (*alde egin*); beste batzuk, gehienak, malgutasun gutxikoak (*adarra jo, gerrikoa estutu*)
- Finkapen lexikala: aski handia, osagaiak ordezkatzeko murriztapenak; batez ere, opakoetan (*hanka sartu* $\neq$ *zangoa sartu*); hala ere: *burua makurtu/apaldu*; *hautsak harrotu/eraiki*

Esapide idiomatikoak (lokuzioak)

Ezaugarri nagusiak

- Idiosinkrasia estatistikoa: oro har, maiztasun txikiagokoak kolokazioak baino (Moon, 1998)
- Ez-konposizionaltasuna: **erabatekoa edo handia**
  - Opakoak (deskodetzeko arazoak): *adarra jo, ziria sartu, erreka jo, gorriak ikusi*
  - Figuratiboak (deskodetzeko aukera; deskonposagarriak): *burua hautsi, garunak urtu, zubiak eraiki*
- Finkapen morfosintaktikoa: aski handia; batzuk aldaezinak (*alde egin*); beste batzuk, gehienak, malgutasun gutxikoak (*adarra jo, gerrikoa estutu*)
- Finkapen lexikala: aski handia, osagaiak ordezkatzeko murriztapenak; batez ere, opakoetan (*hanka sartu* $\neq$ *zangoa sartu*); hala ere: *burua makurtu/apaldu*; *hautsak harrotu/eraiki*

Esapide idiomatikoak (lokuzioak)

Ezaugarri nagusiak

- Idiosinkrasia estatistikoa: oro har, maiztasun txikiagokoak kolokazioak baino (Moon, 1998)
- Ez-konposizionaltasuna: erabatekoa edo handia
  - Opakoak (deskodetzeko arazoak): *adarra jo, ziria sartu, erreka jo, gorriak ikusi*
  - Figuratiboak (deskodetzeko aukera; deskonposagarriak): *burua hautsi, garunak urtu, zubiak eraiki*
- Finkapen morfosintaktikoa: **aski handia**; batzuk aldaezinak (*alde egin*); beste batzuk, gehienak, malgutasun gutxikoak (*adarra jo, gerrikoa estutu*)
- Finkapen lexikala: aski handia, osagaiak ordezkatzeko murriztapenak; batez ere, opakoetan (*hanka sartu* $\neq$ *zangoa sartu*); hala ere: *burua makurtu/apaldu*; *hautsak harrotu/eraiki*

Esapide idiomatikoak (lokuzioak)

Ezaugarri nagusiak

- Idiosinkrasia estatistikoa: oro har, maiztasun txikiagokoak kolokazioak baino (Moon, 1998)
- Ez-konposizionaltasuna: erabatekoa edo handia
  - Opakoak (deskodetzeko arazoak): *adarra jo, ziria sartu, erreka jo, gorriak ikusi*
  - Figuratiboak (deskodetzeko aukera; deskonposagarriak): *burua hautsi, garunak urtu, zubiak eraiki*
- Finkapen morfosintaktikoa: aski handia; batzuk aldaezinak (*alde egin*); beste batzuk, gehienak, malgutasun gutxikoak (*adarra jo, gerrikoa estutu*)
- Finkapen lexikala: **aski handia**, osagaiak ordezkatzeko murriztapenak; batez ere, opakoetan (*hanka sartu* $\neq$ *zangoa sartu*); hala ere: *burua makurtu/apaldu*; *hautsak harrotu/eraiki*

Esapide erdiidiomatikoak (kolokazioak)

Ezaugarri nagusiak

- Idiosinkrasia estatistikoa: oro har, esapide idiomatikoak baino **maiztasun handia**gokoak; eskola britainiarrean, ezaugarri zentrala da
- Ez-konposizionaltasuna: oro har, erdikonposizionalak. Polartasuna (Hausmann, 1989):
  - Oinarriak “ohiko” edo oinarrizko adiera atxikitzen du
  - Kolokatiboa: ahula da (euskarri-aditzak: *eztula egin, lotsa eman*), edo kolokazioari espezifikoa zaion adiera du (*gola sartu, zarata atera*)
- Finkapen morfosintaktikoa: apala, ez da ezaugarri nabarmena (Heid, 1994)
- Finkapen lexikala: ezaugarri bereizgarria (Melčuk, 1994); izaera polarrarekin erlazionatua
  - Kolokazio murriztuetan, kolokatiboa ezin da ordezkatu, edo oso aukera gutxi daude
  - Kolokazio irekiaren kategoria eztabaidatua da (Cowie, 1998: *pay one’s respects/a compliment/court to sb*); hautatze-murrizketak (Fontenelle, 1998)

Esapide erdiidiomatikoak (kolokazioak)

Ezaugarri nagusiak

- Idiosinkrasia estatistikoa: oro har, esapide idiomatikoak baino maiztasun handiagokoak; eskola britainiarrean, ezaugarri zentrala da
- Ez-konposizionaltasuna: oro har, **erdikonposizionalak**. Polartasuna (Hausmann, 1989):
  - Oinarriak “ohiko” edo oinarrizko adiera atxikitzen du
  - Kolokatiboa: ahula da (euskarri-aditzak: *eztula egin, lotsa eman*), edo kolokazioari espezifikoa zaion adiera du (*gola sartu, zarata atera*)
- Finkapen morfosintaktikoa: apala, ez da ezaugarri nabarmena (Heid, 1994)
- Finkapen lexikala: ezaugarri bereizgarria (Melčuk, 1994); izaera polarrarekin erlazionatua
  - Kolokazio murriztuetan, kolokatiboa ezin da ordezkatu, edo oso aukera gutxi daude
  - Kolokazio irekiaren kategoria eztabaidatua da (Cowie, 1998: *pay one’s respects/a compliment/court to sb*); hautatze-murrizketak (Fontenelle, 1998)

Esapide erdiidiomatikoak (kolokazioak)

Ezaugarri nagusiak

- Idiosinkrasia estatistikoa: oro har, esapide idiomatikoak baino maiztasun handiagokoak; eskola britainiarrean, ezaugarri zentrala da
- Ez-konposizionaltasuna: oro har, erdikonposizionalak. Polartasuna (Hausmann, 1989):
  - Oinarriak “ohiko” edo oinarrizko adiera atxikitzen du
  - Kolokatiboa: ahula da (euskarri-aditzak: *eztula egin, lotsa eman*), edo kolokazioari espezifikoa zaion adiera du (*gola sartu, zarata atera*)
- Finkapen morfosintaktikoa: **apala**, ez da ezaugarri nabarmena (Heid, 1994)
- Finkapen lexikala: ezaugarri bereizgarria (Melčuk, 1994); izaera polarrarekin erlazionatua
  - Kolokazio murriztuetan, kolokatiboa ezin da ordezkatu, edo oso aukera gutxi daude
  - Kolokazio irekiaren kategoria eztabaidatua da (Cowie, 1998: *pay one’s respects/a compliment/court to sb*); hautatze-murrizketak (Fontenelle, 1998)

Esapide erdiidiomatikoak (kolokazioak)

Ezaugarri nagusiak

- Idiosinkrasia estatistikoa: oro har, esapide idiomatikoak baino maiztasun handiagokoak; eskola britainiarrean, ezaugarri zentrala da
- Ez-konposizionaltasuna: oro har, erdikonposizionalak. Polartasuna (Hausmann, 1989):
  - Oinarriak “ohiko” edo oinarrizko adiera atxikitzen du
  - Kolokatiboa: ahula da (euskarri-aditzak: *eztula egin, lotsa eman*), edo kolokazioari espezifikoa zaion adiera du (*gola sartu, zarata atera*)
- Finkapen morfosintaktikoa: apala, ez da ezaugarri nabarmena (Heid, 1994)
- Finkapen lexikala: **ezaugarri bereizgarria** (Melčuk, 1994); izaera polarrarekin erlazionatua
  - Kolokazio murriztuetan, kolokatiboa ezin da ordezkatu, edo oso aukera gutxi daude
  - Kolokazio irekiaren kategoria eztabaidatua da (Cowie, 1998: *pay one’s respects/a compliment/court to sb*); hautatze-murrizketak (Fontenelle, 1998)

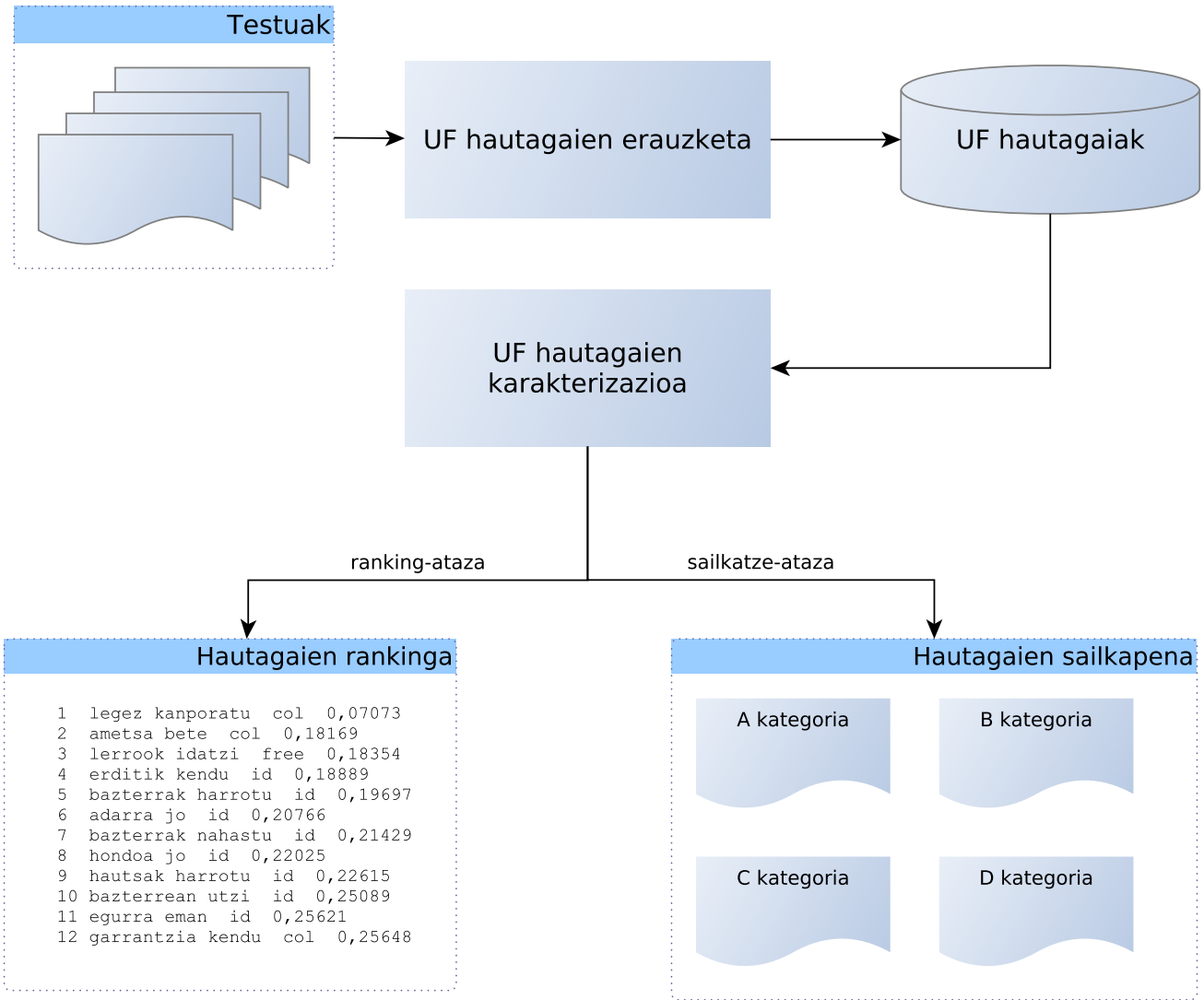
Aurkezpenaren eskema

UFen erauzketa, fraseologia konputazionalaren egitekoetako bat

Atazak (Urizar, 2012)

- **Erauzketa**: testuetatik UFak eskuratzea, eskuarki lexikoiak edo hiztegiak elikatzeko
- Identifikazioa: UFen testuetako banakako agerpenak atzematea
- Interpretazioa: UFen barne-sintaxia eta semantika desanbiguatzea
- Deskribapena: UF bakoitzak izan ditzakeen bariazio morfosintaktikoak zehaztea

Urratsak


Testuak
UF hautagaien erauzketa
UF hautagaiak
UF hautagaien karakterizazioa
ranking-ataza
sailkatze-ataza
Hautagaien rankinga
1 legez kanporatu col 0,07073
2 ametsa bete col 0,18169
3 lerrook idatzi free 0,18354
4 erditik kendu id 0,18889
5 bazterrak harrotu id 0,19697
6 adarra jo id 0,20766
7 bazterrak nahastu id 0,21429
8 hondoa jo id 0,22025
9 hautsak harrotu id 0,22615
10 bazterrean utzi id 0,25089
11 egurra eman id 0,25621
12 garrantzia kendu col 0,25648
Hautagaien sailkapena
A kategoria
B kategoria
C kategoria
D kategoria

Hautagaien erauzketa

Konbinazioak sortzeko zenbait irizpide

- Konbinazioaren osaera
  - Lemak, hitz-formak, bestelako informazio morfosintaktikoa (kasua, mugatasuna...)
- Konbinazioak osatzeko metodoa
  - Distantzia: hitz baten inguruko “leiho-zabalera” (*window span*) baten barneko hitzekin osatzen dira konbinazioak
  - Erlazio morfosintaktiko jakin bat duten hitzen arteko konbinazioak

Hautagaien erauzketa

Konbinazioak sortzeko zenbait irizpide

- Konbinazioaren osaera
  - Lemak, hitz-formak, bestelako informazio morfosintaktikoa (kasua, mugatasuna...)
- Konbinazioak osatzeko metodoa
  - Distantzia: hitz baten inguruko "leiho-zabalera" (*window span*) baten barneko hitzekin osatzen dira konbinazioak
  - Erlazio morfosintaktiko jakin bat duten hitzen arteko konbinazioak

Prozesamendu linguistikoa

Prozesamendu-mailak

- Prozesamendu estandar minimoa: lematizazioa eta etiketatze morfosintaktikoa
  - Lema eta kasu bereko forma flexionatuen ondoriozko aldakuntzak forma kanoniko berarekin lotzeko

    Adarra jotzeko *gogoz, ala? Ez* ***adarrik jo*** *niri.*
    Atzoko partidan hiru ***gol sartu*** *zituen.* ***Gola sartzen*** *ahalegindu da. Horrela jarraituz gero, ez du* ***golik sartuko****. Nork* ***sartu*** *ditu gaurko bi* ***golak****?*
  - Kategoria-osaera jakineko konbinazioak sortzeko: `izena+aditza`, `izena+izena`, `izena+izenondoa`...
- Prozesamendu aurreratua: azaleko edo sakoneko analisi sintaktikoa
  - Abantailak: doitasun handiagoko metodoa
  - Arazoak: analizatzaile sintaktikoetan hizkuntza bakoitzean erdietsi den kalitatearen mendekotasuna

Idiomatikotasuna karakterizatzeko teknikak

Oinarrizko karakterizazio-teknika: idiosinkrasia estatistikoa agerkidetzaren bidez neurtzea

- Osagaien agerkidetza-informazioa prozesatzea, elkartze-neurriak erabiliz (AM – *association measures*; Evert, 2005)
- Behatutako maiztasunak eredu desberdinen arabera estimatutako ausazko maiztasunekin konparatzea
- Ohiko AM batzuk: z neurria, t neurria, χ^2 (khi karratua), LLR (*log-likelihood ratio*, egiantz-arrazoiaren logaritmoa), Fisherren test zehatza, MI (*mutual information*, elkarrekiko informazioa), MI^3, f (maiztasuna)

Idiomatikotasuna karakterizatzeko teknikak

Karakterizazio-teknika aurreratuak

- Konposizionaltasuna: konbinazioaren eta osagaien arteko **antzekotasun distribuzionala**
  - WSM (word space model, *hitz-espazioaren eredua*): Fazly eta Stevenson (2007); Wulff (2008)
  - LSA (latent semantic analysis, *ezkutuko semantikaren analisia*): Katz eta Giesbrecht (2006); Krcmár et al. (2013)
  - Adiera-indukzioa: Korkontzelos eta Manandhar (2009)
- Malgutasun morfosintaktikoa: hautagairen portaera morfosintaktikoa batez besteko erreferentzia-portaera batekin konparatzea (banaketen arteko distantzia)
  - Portaera orokorra: `izena+aditza` osaerako konbinazioen batez besteko portaera (Barkema, 1994; Fazly eta Stevenson, 2007; Wulff, 2008)
  - Osagaien portaera: konbinazioaren osagai batek beste osagaiaren kategoriako edozein hitzekin osatutako konbinazioen batez besteko portaera (Bannard, 2007)
- Malgutasun lexikala: osagaien ordezkagarritasuna
  - Lin (1999); Fazly eta Stevenson (2007); Van de Cruys eta Moirón (2007)

Idiomatikotasuna karakterizatzeko teknikak

Karakterizazio-teknika aurreratuak

- Konposizionaltasuna: konbinazioaren eta osagaien arteko antzekotasun distribuzionala
  - WSM (word space model, *hitz-espazioaren eredua*): Fazly eta Stevenson (2007); Wulff (2008)
  - LSA (latent semantic analysis, *ezkutuko semantikaren analisia*): Katz eta Giesbrecht (2006); Krcmár et al. (2013)
  - Adiera-indukzioa: Korkontzelos eta Manandhar (2009)
- Malgutasun morfosintaktikoa: hautagairen portaera morfosintaktikoa batez besteko **erreferentzia-portaera batekin konparatzea** (banaketen arteko distantzia)
  - Portaera orokorra: `izena+aditza` osaerako konbinazioen batez besteko portaera (Barkema, 1994; Fazly eta Stevenson, 2007; Wulff, 2008)
  - Osagaien portaera: konbinazioaren osagai batek beste osagaiaren kategoriako edozein hitzekin osatutako konbinazioen batez besteko portaera (Bannard, 2007)
- Malgutasun lexikala: osagaien ordezkagarritasuna
  - Lin (1999); Fazly eta Stevenson (2007); Van de Cruys eta Moirón (2007)

Idiomatikotasuna karakterizatzeko teknikak

Karakterizazio-teknika aurreratuak

- Konposizionaltasuna: konbinazioaren eta osagaien arteko antzekotasun distribuzionala
  - WSM (word space model, *hitz-espazioaren eredua*): Fazly eta Stevenson (2007); Wulff (2008)
  - LSA (latent semantic analysis, *ezkutuko semantikaren analisia*): Katz eta Giesbrecht (2006); Krcmár et al. (2013)
  - Adiera-indukzioa: Korkontzelos eta Manandhar (2009)
- Malgutasun morfosintaktikoa: hautagairen portaera morfosintaktikoa batez besteko erreferentzia-portaera batekin konparatzea (banaketen arteko distantzia)
  - Portaera orokorra: `izena+aditza` osaerako konbinazioen batez besteko portaera (Barkema, 1994; Fazly eta Stevenson, 2007; Wulff, 2008)
  - Osagaien portaera: konbinazioaren osagai batek beste osagaiaren kategoriako edozein hitzekin osatutako konbinazioen batez besteko portaera (Bannard, 2007)
- Malgutasun lexikala: **osagaien ordezkagarritasuna**
  - Lin (1999); Fazly eta Stevenson (2007); Van de Cruys eta Moirón (2007)

Sailkapen automatikoa

- Elkartze-neurrien emaitzak konbinatzea. Pecina eta Schlesinger (2006); Lin et al. (2008): f, t neurria, LLR, MI eta χ^2 (ezaugarri guztien konbinazioa da onena, eta hurrena t neurria).
- Agerkidetzaren eta konposizionaltasunaren neurketak konbinatzea. Venkatapathy eta Joshi (2005): ekarpen handiena kolokazioaren izenari dagokion aditzaren eta kolokazioaren arteko antzekotasunak egiten du.
- Lau propietateen neurketak konbinatzea. Fazly eta Stevenson (2007): konbinazioaren aditza ere erabiltzen dute; ekarpen handiena malgutasun sintaktikoarekin eta lexikalarekin lotutako ezaugarriek egiten dute.

Aurkezpenaren eskema

Esperimentuen diseinuaren elementuak

1. Idiomatikotasunaren propietateekin lotutako **behagaiak** aukeratzea, eta horietako bakoitza neurtzeko estrategia zehaztea
2. Esperimentuetako karakterizazio-atazak deskribatzea
3. Ikergaia zehaztea: `izena+aditza` konbinazioak definitzea
4. Corpusetik `izena+aditza` konbinazio hautagaiak erauzteko prozesua
5. Ebaluazio-metodologia antolatzea, eta behar diren erreferentzia-baliabideak garatzea

Esperimentuen diseinuaren elementuak

1. Idiomatikotasunaren propietateekin lotutako behagaiak aukeratzea, eta horietako bakoitza neurtzeko estrategia zehaztea
2. Esperimentuetako **karakterizazio-atazak** deskribatzea
3. Ikergaia zehaztea: `izena+aditza` konbinazioak definitzea
4. Corpusetik `izena+aditza` konbinazio hautagaiak erauzteko prozesua
5. Ebaluazio-metodologia antolatzea, eta behar diren erreferentzia-baliabideak garatzea

Esperimentuen diseinuaren elementuak

1. Idiomatikotasunaren propietateekin lotutako behagaiak aukeratzea, eta horietako bakoitza neurtzeko estrategia zehaztea
2. Esperimentuetako karakterizazio-atazak deskribatzea
3. Ikergaia zehaztea: **`izena+aditza` konbinazioak** definitzea
4. Corpusetik `izena+aditza` konbinazio hautagaiak erauzteko prozesua
5. Ebaluazio-metodologia antolatzea, eta behar diren erreferentzia-baliabideak garatzea

Esperimentuen diseinuaren elementuak

1. Idiomatikotasunaren propietateekin lotutako behagaiak aukeratzea, eta horietako bakoitza neurtzeko estrategia zehaztea
2. Esperimentuetako karakterizazio-atazak deskribatzea
3. Ikergaia zehaztea: `izena+aditza` konbinazioak definitzea
4. Corpusetik `izena+aditza` konbinazio **hautagaiak erauzteko prozesua**
5. Ebaluazio-metodologia antolatzea, eta behar diren erreferentzia-baliabideak garatzea

Esperimentuen diseinuaren elementuak

1. Idiomatikotasunaren propietateekin lotutako behagaiak aukeratzea, eta horietako bakoitza neurtzeko estrategia zehaztea
2. Esperimentuetako karakterizazio-atazak deskribatzea
3. Ikergaia zehaztea: `izena+aditza` konbinazioak definitzea
4. Corpusetik `izena+aditza` konbinazio hautagaiak erauzteko prozesua
5. **Ebaluazio-metodologia** antolatzea, eta behar diren erreferentzia-baliabideak garatzea

Behagaiak

Hipotesia da behagai hauen bidez neur daitezkeela idiomatikotasunaren lau propietateak edo osagaiak:

Propietatea	**Behagaia**
Instituzionalizazioa (idiosinkrasia estatistikoa)	Agerkidetza
Konposizionaltasun-maila	Antzekotasun distribuzionala
Malgutasun morfosintaktikoa	Erreferentzia-portaera batekiko distantzia
Malgutasun lexikala	Osagaien ordezkagarritasuna

Karakterizazio-atazak

Ranking-ataza

- Continuumaren ideian oinarrituta, UF hautagaiak idiomatikotasun-mailaren arabera ordenatzea, rankingak antolatzea

  $\rightarrow$ Idiomatikotasunaren propietateen banakako neurketa

Sailkatze-ataza

- UF hautagaiak sailkapen-ereduan bereizi ditugun kategorietan banatzea

  $\rightarrow$ Propietateen banakako neurketen konbinazioa, ikasketa automatikoan

izena+aditza osaerako konbinazioen errepertorioa

Zabala (2004), Altzibar (2005) eta Urizarren (2012) lanetan oinarritua

1 izena$_{\text{subjektua}}$ + aditza

   1.1 Absolutibodun sintagma: *burua joan, eguzkia sartu, ardoa garraztu, esnea galdu*

   1.2 Ergatibodun sintagma: *gogoak eman, loak hartu, suak hartu, ilunak jo*

2 izena$_{\text{objektua}}$ + aditza: *hanka sartu, zubiak eraiki, lan egin, min hartu, bizarra egin, itxurak egin, aukera eman, erabakia hartu, zarata atera, eskerrak eman, urratsak egin, adostasuna lortu, gola sartu, elkartasuna adierazi*

3 izena$_{\text{subjektuaren pred.}}$ + aditza: *beldur izan, falta izan, giro egon*

4 izena$_{\text{objektuaren pred.}}$ + aditza: *atsegin ukan, damu ukan*

5 izena$_{\text{datiboa}}$ + aditza: *edanari eman, bideari ekin, lanari lotu*

6 izena$_{\text{adjuntua}}$ + aditza: *mendean hartu, borrokan sartu, martxan jarri, adarretatik heldu, larrutik ordaindu, burutik kendu, harira etorri, gogora ekarri, aurrera eraman, zerbitzura egon, sareetara bidali, muturreraino eraman, aurrez ikusi, arduraz jokatu, oinarritzat hartu*

UF hautagaiak erauztea

Corpus-baliabideak eta prozesamendua

- Kazetaritza-corpus bat (75 milioi hitz), bi iturritatik eratua
  - *Euskaldunon Egunkaria*: 2001-2002 (28 milioi hitz)
  - *Berria*: 2006-2010 (47 milioi hitz)
- Etiketatze linguistikoa: lematizazioa eta analisi morfologikoa
  - UPV/EHUko Ixa taldearen Eustagger etiketatzailea
- Eustaggerren irteeraren tratamendua
  - Izenen informazioa: `forma lema kategoria azpikategoria kasua mugatasuna`
  - Aditzen informazioa: printzipioz, `lema kategoria`; baina `PART` analisia dutenetan, analisi-kateko zenbait informazio atxikitzen ditugu:
    - *herrialde aurreratuak, gobernuaren aliatuak* modukoak ez erauzteko
    - Partizipio jokatugabearen gaineko erlatiboa bereiz tratatzeko (*hartutako erabakien ondorioak*)

UF hautagaiak erauztea

`izena+aditza` osaerako konbinazio hautagaiak lortzea

- Bigrama-sorkuntza
  - Ngram Statistics Package (NSP)[a] (Pedersen eta Banerjee, 2010)
  - Erauzte-parametroak
    - Maiztasun-atariak: 3, 10, 30, 50
    - Leiho-zabalera: $w = \pm 1$, $w = \pm 5$
- Forma kanonikoa esleitzea: bigramen normalizazioa
  - Aditz jakin batekin konbinaturik agertzen diren kasu bereko izen-formak normalizatu egin ditugu, maiztasun handieneko forma eta mugatasuna esleituz

[a] `http://www.d.umn.edu/~tpederse/nsp.html`

UF hautagaiak erauztea

egunkari_irakurri lema-bikotearen kasu eta mugatasun desberdineko konbinazio batzuk

egunkari**etan**_egunkari_IZE_ARR_**INE_NUMP**<>irakurri_ADI<>31 74 1956

egunkari**an**_egunkari_IZE_ARR_**INE_NUMS**<>irakurri_ADI<>43 76 1956

egunkari**ak**_egunkari_IZE_ARR_**ABS_NUMP**<>irakurri_ADI<>67 486 1956

egunkari**a**_egunkari_IZE_ARR_**ABS_NUMS**<>irakurri_ADI<>126 1003 1956

Normalizazioaren emaitza

egunkari**an**_egunkari_IZE_ARR_**INE_NUMS**<>irakurri_ADI<>**74 150** 1956

egunkari**a**_egunkari_IZE_ARR_**ABS_NUMS**<>irakurri_ADI<>**193 1489** 1956

Ebaluazio-prozedura

1. Irizpideak
2. Hiztegi-erreferentzia
3. Ebaluazio-erreferentzia sailkatua
4. Metrikak

Ebaluazioa: irizpideak

Kontuan hartu beharreko alderdi nagusiak

- UF kategoriak bereizi nahi ditugu $\rightarrow$ **UF bai/ez sailkapena aski ez**
- Euskaraz UF-kategoriak bereizten dituen iturri lexikografikorik ez dago $\rightarrow$ adituen eskulanaren bidezko ebaluazio-erreferentzia bat eratu behar
- Antzekotasun distribuzionalaren eta malgutasun morfosintaktikoaren neurketek hautagaien testuinguru-informazio aberatsa behar dute $\rightarrow$ hautagaien agerpen-kopurua txikiegia ez izatea (20-50)
- Ikasketa automatikoa: sistemak ikasteko adinako kopuruan izan behar ditu UF-kategorien instantziak $\rightarrow$ moduren bat behar dugu ebaluazio-erreferentzian UFak gutxiegi ez izateko

Ebaluazioa: irizpideak

Kontuan hartu beharreko alderdi nagusiak

- UF kategoriak bereizi nahi ditugu $\rightarrow$ UF bai/ez sailkapena aski ez
- Euskaraz UF-kategoriak bereizten dituen iturri lexikografikorik ez dago $\rightarrow$ **adituen eskulanaren bidezko ebaluazio-erreferentzia** bat eratu behar
- Antzekotasun distribuzionalaren eta malgutasun morfosintaktikoaren neurketek hautagaien testuinguru-informazio aberatsa behar dute $\rightarrow$ hautagaien agerpen-kopurua txikiegia ez izatea (20-50)
- Ikasketa automatikoa: sistemak ikasteko adinako kopuruan izan behar ditu UF-kategorien instantziak $\rightarrow$ moduren bat behar dugu ebaluazio-erreferentzian UFak gutxiegi ez izateko

Ebaluazioa: irizpideak

Kontuan hartu beharreko alderdi nagusiak

- UF kategoriak bereizi nahi ditugu $\rightarrow$ UF bai/ez sailkapena aski ez
- Euskaraz UF-kategoriak bereizten dituen iturri lexikografikorik ez dago $\rightarrow$ adituen eskulanaren bidezko ebaluazio-erreferentzia bat eratu behar
- Antzekotasun distribuzionalaren eta malgutasun morfosintaktikoaren neurketek hautagaien testuinguru-informazio aberatsa behar dute $\rightarrow$ **hautagaien agerpen-kopurua txikiegia ez izatea** (20-50)
- Ikasketa automatikoa: sistemak ikasteko adinako kopuruan izan behar ditu UF-kategorien instantziak $\rightarrow$ moduren bat behar dugu ebaluazio-erreferentzian UFak gutxiegi ez izateko

Ebaluazioa: irizpideak

Kontuan hartu beharreko alderdi nagusiak

- UF kategoriak bereizi nahi ditugu $\rightarrow$ UF bai/ez sailkapena aski ez
- Euskaraz UF-kategoriak bereizten dituen iturri lexikografikorik ez dago $\rightarrow$ adituen eskulanaren bidezko ebaluazio-erreferentzia bat eratu behar
- Antzekotasun distribuzionalaren eta malgutasun morfosintaktikoaren neurketek hautagaien testuinguru-informazio aberatsa behar dute $\rightarrow$ hautagaien agerpen-kopurua txikiegia ez izatea (20-50)
- Ikasketa automatikoa: sistemak ikasteko adinako kopuruan izan behar ditu UF-kategorien instantziak $\rightarrow$ moduren bat behar dugu **ebaluazio-erreferentzian UFak gutxiegi ez** izateko

Hiztegi-erreferentzia

Iturriak

- HB: Euskaltzaindiaren *Hiztegi Batua*[a]
- EH: Ibon Sarasolaren *Euskal Hiztegia*
- ELH: Elhuyar Fundazioaren *Euskara-Castellano/CastellanoVasco Hiztegia*[b]
- Intza: *Intza proiektua*[c]
- EDBL: Ixa taldearen EDBL datu-base lexikala[d]

[a] `http://www.euskaltzaindia.net/hiztegibatua`; `http://www.euskaltzaindia.net/eaeb` gunetik deskargatua (2010-06-04ko bertsioa).

[b] 2010-06-02ko bertsioa.

[c] `http://intza.armiarma.com` (2010-06-02ko bertsioa).

[d] `http://ixa2.si.ehu.es/edbl` (2010-06-22ko bertsioa).

Emaitza: `izena+aditza` osaerako 3 720 UF

Ebaluazio-erreferentzia

Laginaren hautaketa

- Hautatutako erauzketa: $f \geqslant 30$, $w = \pm 1$, bigramak normalizatuta
- UFen dentsitatea handitzeko: AMen araberako ranking bakoitzetik lehen n hautagaien multzoak hartu, eta horien bildura egin; $n = 2\,000$ hautatu dugu. Bigrama-kopurua: 4 334
- 1 200 bigramako ausazko azpimultzoa aukeratu dugu, aditu-talde batek eskuz sailka dezan

Sailkatze-lana

- Hiru hizkuntzalariz osatutako aditu-talde batek 5 kategoriatan sailkatu du (`id_op`, `id_fig`, `col_res`, `col_open`, `free`)
- Sailkatze-prozesurako irizpideak Krennen (2004) lanean inspiratuta eman ditugu
- `izena+aditza` ez diren konbinazioak (55) kenduta: 1 145eko lagina
- 5 kategoriako ereduari uko egin eta 3 kategoriakoa hobetsi dugu: `id` (`id_op`, `id_fig`), `col` (`col_res`, `col_open`), `free`. Arrazoiak:
  - Esapide idiomatiko opakoen kategoria oso instantzia gutxiri esleitu diete adituek
  - Kolokazio irekien kategoria oso labaina da, eta zalantza ugari sortu ditu sailkatzaileen artean
- ITA (etiketatzaileen arteko adostasuna): Fleiss $\kappa = 0{,}58$ (adostasun ertaina)
- Erreferentzia sailkatua: `id` 80; `col` 268; `free` 797

Sailkatze-lana

- Hiru hizkuntzalariz osatutako aditu-talde batek 5 kategoriatan sailkatu du (`id_op`, `id_fig`, `col_res`, `col_open`, `free`)
- Sailkatze-prozesurako irizpideak Krennen (2004) lanean inspiratuta eman ditugu
- `izena+aditza` ez diren konbinazioak (55) kenduta: 1 145eko lagina
- 5 kategoriako ereduari uko egin eta 3 kategoriakoa hobetsi dugu: **`id`** (`id_op`, `id_fig`), **`col`** (`col_res`, `col_open`), **`free`**. Arrazoiak:
  - Esapide idiomatiko opakoen kategoria oso instantzia gutxiri esleitu diete adituek
  - Kolokazio irekien kategoria oso labaina da, eta zalantza ugari sortu ditu sailkatzaileen artean
- ITA (etiketatzaileen arteko adostasuna): Fleiss $\kappa = 0{,}58$ (adostasun ertaina)
- Erreferentzia sailkatua: `id` 80; `col` 268; `free` 797

Erreferentzia sailkatuaren osaera

kategoria	Hiztegi-erreferentzia bai	ez	Totala
`id`	23	57	80
`col`	41	227	268
`free`	10	787	797
Totala	74	1 071	1 145

Eskuz sailkatutako ebaluazio-erreferentziako bigramen eta hiztegi-erreferentziaren arteko konparazioa.

Nabarmentzekoak

- Hiztegia aberasteko aukera handia (batez ere, `col`: % 15,3 daude hiztegietan)
- Predikatu konplexu batzuk (*jaramon egin, zor izan*) esapide erdiidiomatikoetan sailkatu dira, eta hori ez dator bat euskarazko fraseologian horrelakoak aditz-lokuziotzat jotzeko tradizioarekin

  $\rightarrow$ komenigarria da, kontrasterako, konbinazio horiek (17) esapide idiomatikotzat sailkatuta esperimentuen emaitzetan igarriko litzatekeen eragina aztertzea

Ebaluazio-metrikak

Ranking-ataza

- Ideia da nolabait konparatzea idealki ordenatutako rankinga neurri bakoitzak sortutako rankingarekin
  - Kendall τ_B heinen korrelazio-koefizientea, berdinketak kudeatzeko
- Batez besteko doitasunak ere kalkulatu dira (AP – *average precision*): $AP_{\texttt{UF}}$, $AP_{\texttt{id}}$ eta $AP_{\texttt{col}}$
  - Neurri batek UFekiko oro har eta UF-kategoria bakoitzarekiko duen portaera ikusteko

Sailkatze-ataza

- Wekak eskaintzen dituen neurri hauek erabili ditugu:
  - Zuzen sailkatutako instantzia-kopurua (*Correctly Classified Instances* – CCI) Zehaztasunaren baliokidea da (*accuracy*)
  - Kategoria bakoitzeko F neurriak: $F_{\texttt{id}}$, $F_{\texttt{col}}$, $F_{\texttt{free}}$
  - Batez besteko F neurri haztatua (*Weighted Average F-measure*) Mikrobatezbestekoa da (F_{mikro})

  Mikrobatezbestekoaz gain, makrobatezbestekoa ere (F_{makro}) kalkulatu dugu

Idiosinkrasia estatistikoaren neurketa

Agerkidetza-datuen prozesamendua: AMen kalkulua

- NSPren irteera UCS toolkitaren bidez prozesatzea[a] (Evert, 2005)
- Erabilitako AMak: z neurria, t neurria, χ^2 (khi karratua), LLR (*log-likelihood ratio*, egiantz-arrazoiaren logaritmoa), Fisherren test zehatza, MI (*mutual information*, elkarrekiko informazioa), MI^3, f (maiztasuna)

[a] http://www.collocations.de/software.html

Idiosinkrasia estatistikoaren neurketa

Agerkidetza-datuen prozesamendua: AMen kalkulua

- NSPren irteera UCS toolkitaren bidez prozesatzea[a] (Evert, 2005)
- Erabilitako AMak: z neurria, t neurria, χ^2 (khi karratua), LLR (*log-likelihood ratio*, egiantz-arrazoiaren logaritmoa), Fisherren test zehatza, MI (*mutual information*, elkarrekiko informazioa), MI^3, f (maiztasuna)

[a] http://www.collocations.de/software.html

	τ_B	AP_{UF}	AP_{id}	AP_{col}
ausazko rankinga	0,000	0,308	0,070	0,234
z neurria	(-0,038)	0,297	0,109	0,204
t neurria	0,197	0,455	0,084	0,383
χ^2	(-0,037)	0,302	0,119	0,206
LLR	0,156	0,427	0,100	0,335
Fisher	0,156	0,426	0,100	0,335
MI	(-0,121)	0,257	0,086	0,182
MI^3	0,103	0,389	0,107	0,291
f	0,189	0,436	0,074	0,379

Idiosinkrasia estatistikoaren neurketa

Agerkidetza-datuen prozesamendua: AMen kalkulua

- NSPren irteera UCS toolkitaren bidez prozesatzea[a] (Evert, 2005)
- Erabilitako AMak: z neurria, t neurria, χ^2 (khi karratua), LLR (*log-likelihood ratio*, egiantz-arrazoiaren logaritmoa), Fisherren test zehatza, MI (*mutual information*, elkarrekiko informazioa), MI^3, f (maiztasuna)

[a] http://www.collocations.de/software.html

	τ_B	AP_{UF}	AP_{id}	AP_{col}
ausazko rankinga	0,000	0,308	0,070	0,234
z neurria	(-0,038)	0,297	0,109	0,204
t neurria	**0,197**	**0,455**	0,084	**0,383**
χ^2	(-0,037)	0,302	0,119	0,206
LLR	0,156	0,427	0,100	0,335
Fisher	0,156	0,426	0,100	0,335
MI	(-0,121)	0,257	0,086	0,182
MI^3	0,103	0,389	0,107	0,291
f	0,189	0,436	0,074	0,379

Idiosinkrasia estatistikoaren neurketa

Agerkidetza-datuen prozesamendua: AMen kalkulua

- NSPren irteera UCS toolkitaren bidez prozesatzea[a] (Evert, 2005)
- Erabilitako AMak: z neurria, t neurria, χ^2 (khi karratua), LLR (*log-likelihood ratio*, egiantz-arrazoiaren logaritmoa), Fisherren test zehatza, MI (*mutual information*, elkarrekiko informazioa), MI^3, f (maiztasuna)

[a] http://www.collocations.de/software.html

	τ_B	AP_{UF}	AP_{id}	AP_{col}
ausazko rankinga	0,000	0,308	0,070	0,234
z neurria	(-0,038)	0,297	0,109	0,204
t neurria	**0,197**	**0,455**	0,084	**0,383**
χ^2	(-0,037)	0,302	**0,119**	0,206
LLR	0,156	0,427	0,100	0,335
Fisher	0,156	0,426	0,100	0,335
MI	(-0,121)	0,257	0,086	0,182
MI^3	0,103	0,389	0,107	0,291
f	0,189	0,436	0,074	0,379

Konposizionaltasun semantikoaren neurketa

Antzekotasun distribuzionalaren hipotesia:

- *Difference of meaning correlates with difference of distribution.* (Harris, 1954)

UFei aplikatua

- *The underlying hypothesis is that semantically idiomatic MWEs will occur in markedly different lexical contexts to their component words.* (Baldwin eta Kim, 2010)

UFentzako funtsezko prozedura

- UF izateko hautagai den hitz-konbinazioaren testuinguruak haren osagai bakunen testuinguruekin konparatzea
- Konparazio hori osagai bakoitzarekiko egin dugu, bakoitzaren eragina bereiz aztertzeko (Wulff, 2008); horrez gain, baterako neurria ere kalkulatu da, bien batezbestekoaren bidez

Konposizionaltasun semantikoaren neurketa

Testuinguru-sorkuntza

- Ondoz ondoko agerkidetzak hartu dira bigramatzat; arrazoia da erauzketan ezarritako irizpide bera erabiltzea ($w = \pm 1$)
- Osagai bakunen testuinguruetan konbinazioaren testuinguruak ez sartzea
  - *mahaia jaso* bigramaren agerpena atzematen denean, testuinguruko hitzek *mahaia jaso*ren testuinguru-dokumentua elikatu dute, eta ez dira *mahai* eta *jaso*ren testuingurutzat kontsideratu
- Testuinguru-dokumentuak elikatzeko: eduki-hitzak bakarrik (izen, aditz eta adjektiboak)
- Esaldi osoa hartu dugu testuingurutzat (Reddy et al., 2011)

Konposizionaltasun semantikoaren neurketa

Testuinguruen prozesamendua – WSM (*word space model*)

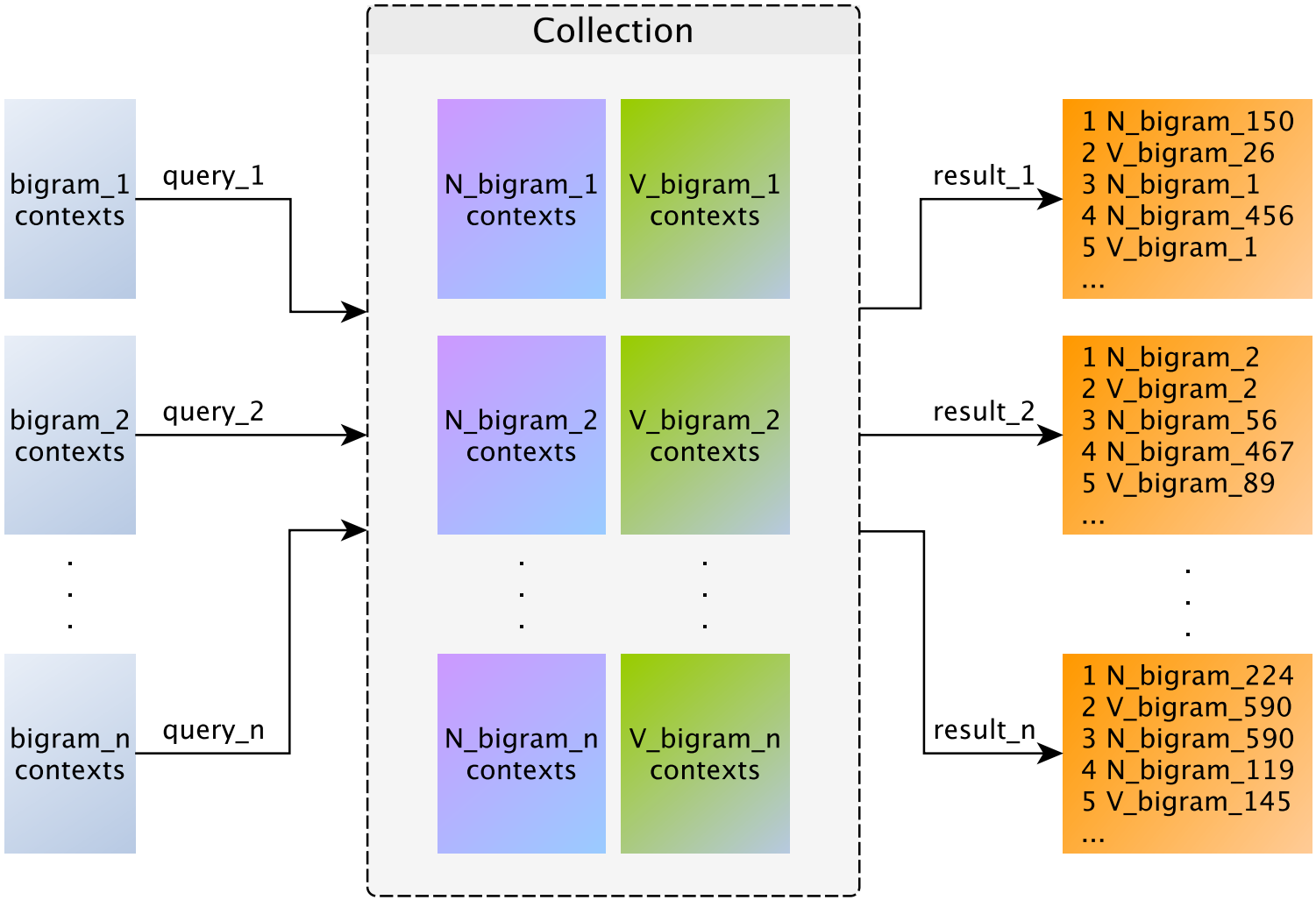
- Testuinguruak bektore gisa errepresentatuta
  - Bektoreetako balioak edo pisuak: f, t neurria, LLR, MI eta Fisherren test zehatza
  - Dimentsioak gutxitzeko:
    - Ehuneko jakin bat: rankingaren % 75 eta % 50
    - Kopuru mugatu bat: rankingeko lehen 3 000, 2 000 eta 1 000 elementuak (Reddy et al., 2011; Mitchell eta Lapata, 2008)
    - Maiztasun-atari bat ($f > 3$)
  - Corpuseko maiztasun handieneko 100 hitzak ez dira kontuan hartu (*stop word* gisa prozesatu dira)
- Antzekotasun-neurriak
  - Jaccard koefizientea, kosinua eta Jensen-Shannon dibergentzia
  - Berry-Roggheren (1974) R balioa (R_{BR}) eta Wulffen (2008) bi hedapenak (R_{W1} eta R_{W2})

Konposizionaltasun semantikoaren neurketa

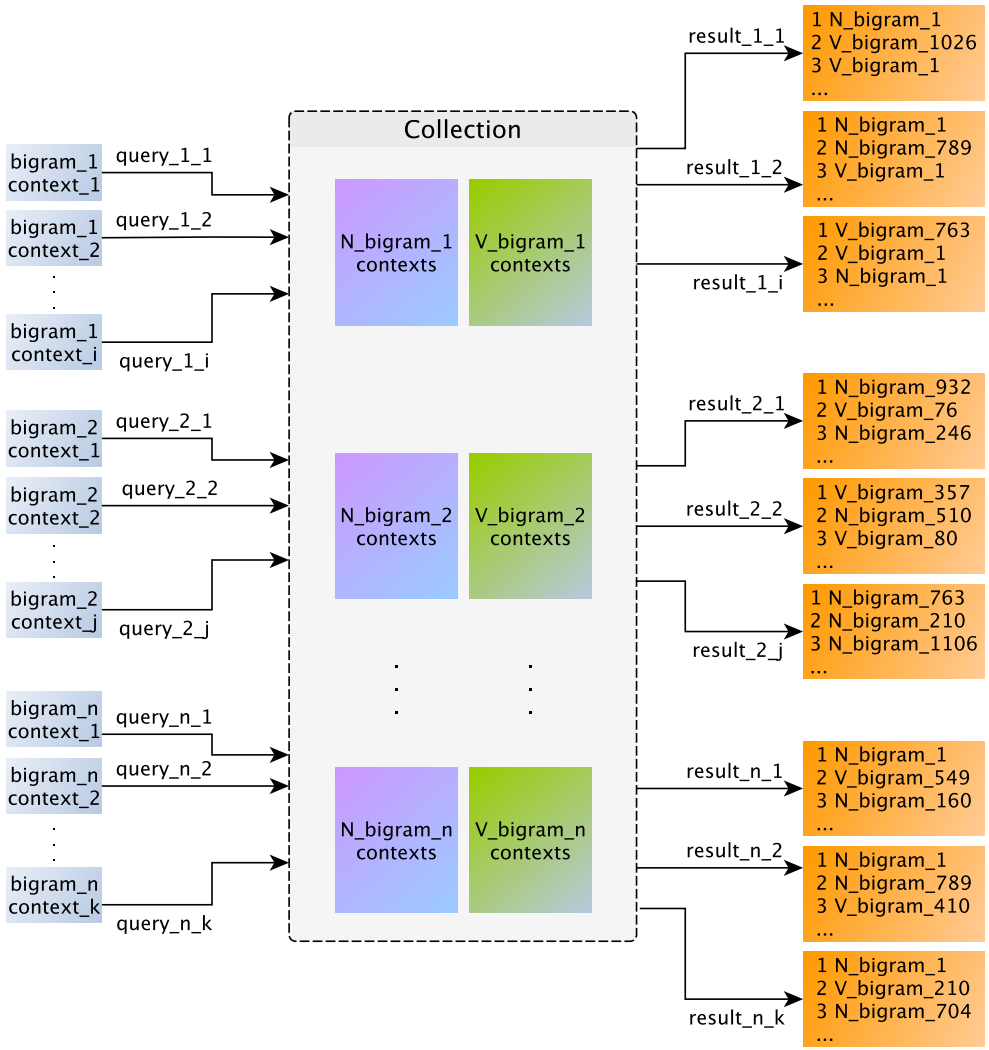
Testuinguruen prozesamendua – IR metodoak (Lemur)

- Bigramen testuinguruen dokumentuak *query* edo kontsulta gisa erabiltzea bigramen osagaien testuinguruen dokumentuek osatzen duten bildumaren kontra
- Bi modalitate:
  - L1: bigrama baten testuinguruak dokumentu bakarrean bildu dira, eta orobat osagaien testuinguruak
  - L2: bigramen testuinguru-esaldiak dokumentu banatan sartu dira; beraz, bigrama bakoitzerako agerkidetza adina kontsulta-dokumentu eratu dira
- Indizeak: Indri, KL, tf-idf, Okapi, cos (idf)

Lemur – L1 modalitatea



Lemur – L2 modalitatea



Malgutasun morfosintaktikoa

Prozeduraren oinarriak

- Bigramaren portaera bi erreferentzia-portaera mota hauekin konparatzea
  - Portaera orokorra
  - Osagaien portaera
- Konbinazio baten portaera = aldakuntza morfosintaktikoen banaketa
- Aldakuntzen hautaketa
  - EDBLko gauzatze-eskemak (Urizar, 2012) aztertu ondoren `izena` osagaiaren aldakuntzetan jarri dugu arreta; aditzaren aldakuntzak ez dira idiomatikotasunerako esanguratsuak
- Murriztapen-gramatika bat garatu dugu hautatutako aldakuntzak detektatzeko

Malgutasun morfosintaktikoa – Aztertutako aldakuntzak I

Izenaren hedapenak

- Determinatzailea: *liburu **bat** irakurri dut; **zenbat** liburu irakurri dituzu?*
- Izenondoa: *liburu **interesgarria** irakurri nuen*
- Izenlaguna: ***gustuko** liburuak irakurtzea*; ***italierazko** liburua irakurri*

Hedapen bat baino gehiago konbina daitezke aldakuntza berean: *liburu **interesgarri bat** irakurri dut; **lau** liburu **hauek** irakurri ditut; anaiak irakurritako **frantsesezko** liburu **eder batzuk***.

Erlatiboa

- ***irakurri dudan** liburua; anaiak **irakurritako** frantsesezko **liburu** eder batzuk*

Malgutasun morfosintaktikoa – Aztertutako aldakuntzak II

Mugatasuna: `NUMS, NUMP, MG, PH`

- *liburu interesgarri**a*** (`NUMS`) / *interesgarri**ak** irakurri nuen/nituen* (`NUMP`)
- *liburu **bat** irakurri dut* (`MG`)
- *ez dut liburu **hori** irakurri* (`NUMS`)
- ***hiru** liburu **hauek** irakurriko ditut* (`NUMP`)

Ordena: `IZE ADI` / `ADI IZE`

- ***liburu** interesgarri bat **irakurri** dut* / ***irakurri** dut gomendatu didazun **liburua***

Malgutasun morfosintaktikoa – Malgutasun-neurriak

Banaketen arteko distantziaren neurriak

- **RFR** (*relative frequency ratio*): aldakuntza bakoitzerako behatutako maiztasun erlatiboaren eta itxarondakoaren arteko zatiduren batura, Barkematik (1994) egokitu duguna
- **NSSD** (*normalized sum of squared deviations*): aurreko maiztasunen kenduren karratuen batura normalizatua (Wulff, 2008)
- H_{rel}: behatutako banaketaren entropia erlatiboa, sistemaren entropia maximoarekiko (Wulff, 2008)
- **KL**: Kullback-Leibler dibergentzia (Fazly and Stevenson, 2007)
- **CPMI** (*conditional pointwise mutual information*): elkarrekiko informazio puntual baldintzazkoa (Bannard, 2007)

Malgutasun lexikala – Ordezkagarritasuna

Prozeduraren oinarria

- Konbinazioaren osagaietako bakoitzaren ordezkagarritasuna konputatzea, ordezkotzat osagaiaren sinonimoak, kuasisinonimoak edo semantikoki erlazionatutako hitzak erabiliz

Ordezkoak eskuratzeko baliabideak

- `ELH_SK`: *Sinonimoen Kutxa* (Elhuyar Fundazioa)
- `EusWN 3.0`: WordNeten euskarazko bertsioa[a] (Ixa taldea)
- Baliabide konbinatuak eta hedatuak:
  - `ELHWN`: aurreko bien bildura
  - `WNhedap`: bigramaren osagai bakoitzak `EusWN 3.0`-n dituen ahaideak (*siblings*, edo hiperonimo bereko hitzak) gehituz osatutako baliabide hedatua
  - `ELHWNhedap`: `ELHWN` eta `WNhedap` baliabideen bildura
- `EB_antzdistr`: bigramen osagaien thesaurus distribuzional bat eratu dugu corpusetik, semantikoki antzekoenak diren hitzez osatua

[a] http://ixa2.si.ehu.es/cgi-bin/mcr/public/wei.consult.perl

Malgutasun lexikala – Ordezkagarritasuna

Ordezkagarritasuna neurtzeko baliabideen ezaugarriak: ordezkoen kopuruak

kategoria	ELH_SK	EusWN 3.0	ELHWN	WNhedap	ELHWNhedap
IZE	29 407	29 603	52 434	548 454	570 404
ADI	9 333	7 898	15 061	87 567	94 371
Totala	**38 740**	**37 669**	**67 495**	**636 021**	**664 775**

Adiera-bereizketa kontuan hartu gabe eratutako sinonimo-bikoteen bost bildumak

Ordezko konbinaziorik ez duten edo corpusean ordezkoaren agerpenik ez duten bigramen kopuruak

Baliabidea	**ordezkorik ez**
ELH_SK	674
EusWN 3.0	552
ELHWN	475
WNhedap	299
ELHWNhedap	238
EBantzdistr_20	207

Malgutasun lexikala – Ordezkagarritasuna

Ordezkagarritasunaren neurriak

- Van de Cruys eta Moirónen (2007) R indizea, bigramaren eta ordezko aldaeren arteko KL dibergentzian oinarritua
  - R_{nv} eta R_{vn}, hurrenez hurren, aditz batek izen jakin batekiko duen joera eta izen batek aditz batekiko duen joera dira, bere ordezkoek izen edo aditz horiekiko duten joerarekin konparatuta
- Fazly eta Stevensonen (2007) $\text{Fixedness}_{\text{lex}}$ neurria, bigramaren eta haren osagai bakoitza ordezkatuz sortutako aldaeren MI balioen arteko z neurria dena
  - Egileen baterako neurketaz gain (z_PMI_NV), z_PMI_V eta z_PMI_N ordezkagarritasunak bereiz kalkulatu ditugu

Esperimentu bakunen emaitza hautatuak

	neurria	τ_B	$AP_{\texttt{UF}}$	$AP_{\texttt{id}}$	$AP_{\texttt{col}}$
	ausaz	0,000	0,309	0,070	0,234
AM	t neurria	0,197	0,455	0,084	0,383
	χ^2	(-0,037)	0,302	0,119	0,206
DSim	L2_Indri_hit_erl_NV	0,322	0,566	0,294	0,354
	L2_KL_hit_erl_NV	0,309	0,551	0,320	0,332
	L2_Indri_rankV_hazt	0,320	0,552	0,130	0,431
MSFlex	CPMI_osag_mugat	0,154	0,432	0,113	0,329
	H_{rel}_mugat	0,119	0,434	0,134	0,313
	CPMI_orok_ADJ	(0,045)	0,387	0,202	0,237
	CPMI_izena_mugat	0,132	0,425	0,101	0,331
LFlex	z_PMI_V_WNhedap	0,110	0,365	0,120	0,259
	R_{nv}_EBantzdistr_20	0,066	0,381	0,066	0,323
	z_PMI_V_ELHWNhedap	0,108	0,357	0,122	0,250

Idiomatikotasunaren lau osagaiak neurtzeko egindako esperimentu bakunen emaitza onenen laburpena. Kendall τ_B koefizientea (idiomatikotasun-ranking ideal batekiko hein-korrelazioa); eta batez besteko doitasunak (AP), UFen rankingerako ($AP_{\texttt{UF}}$), oro har, eta AP espezifikoak esapide idiomatikoetarako eta kolokazioetarako.

Esperimentu bakunen emaitza hautatuak

	neurria	τ_B	$AP_{\texttt{UF}}$	$AP_{\texttt{id}}$	$AP_{\texttt{col}}$
	ausaz	0,000	0,309	0,070	0,234
AM	t neurria	0,197	0,455	0,084	0,383
	χ^2	(-0,037)	0,302	0,119	0,206
DSim	L2_Indri_hit_erl_NV	**0,322**	0,566	0,294	0,354
	L2_KL_hit_erl_NV	0,309	0,551	0,320	0,332
	L2_Indri_rankV_hazt	0,320	0,552	0,130	0,431
MSFlex	CPMI_osag_mugat	0,154	0,432	0,113	0,329
	H_{rel}_mugat	0,119	0,434	0,134	0,313
	CPMI_orok_ADJ	(0,045)	0,387	0,202	0,237
	CPMI_izena_mugat	0,132	0,425	0,101	0,331
LFlex	z_PMI_V_WNhedap	0,110	0,365	0,120	0,259
	R_{nv}_EBantzdistr_20	0,066	0,381	0,066	0,323
	z_PMI_V_ELHWNhedap	0,108	0,357	0,122	0,250

Idiomatikotasunaren lau osagaiak neurtzeko egindako esperimentu bakunen emaitza onenen laburpena. Kendall τ_B koefizientea (idiomatikotasun-ranking ideal batekiko hein-korrelazioa); eta batez besteko doitasunak (AP), UFen rankingerako ($AP_{\texttt{UF}}$), oro har, eta AP espezifikoak esapide idiomatikoetarako eta kolokazioetarako.

Esperimentu bakunen emaitza hautatuak

	neurria	τ_B	$AP_{\texttt{UF}}$	$AP_{\texttt{id}}$	$AP_{\texttt{col}}$
	ausaz	0,000	0,309	0,070	0,234
AM	t neurria	0,197	0,455	0,084	0,383
	χ^2	(-0,037)	0,302	0,119	0,206
DSim	L2_Indri_hit_erl_NV	**0,322**	**0,566**	0,294	0,354
	L2_KL_hit_erl_NV	0,309	0,551	0,320	0,332
	L2_Indri_rankV_hazt	0,320	0,552	0,130	0,431
MSFlex	CPMI_osag_mugat	0,154	0,432	0,113	0,329
	H_{rel}_mugat	0,119	0,434	0,134	0,313
	CPMI_orok_ADJ	(0,045)	0,387	0,202	0,237
	CPMI_izena_mugat	0,132	0,425	0,101	0,331
LFlex	z_PMI_V_WNhedap	0,110	0,365	0,120	0,259
	R_{nv}_EBantzdistr_20	0,066	0,381	0,066	0,323
	z_PMI_V_ELHWNhedap	0,108	0,357	0,122	0,250

Idiomatikotasunaren lau osagaiak neurtzeko egindako esperimentu bakunen emaitza onenen laburpena. Kendall τ_B koefizientea (idiomatikotasun-ranking ideal batekiko hein-korrelazioa); eta batez besteko doitasunak (AP), UFen rankingerako ($AP_{\texttt{UF}}$), oro har, eta AP espezifikoak esapide idiomatikoetarako eta kolokazioetarako.

Esperimentu bakunen emaitza hautatuak

	neurria	τ_B	$AP_{\texttt{UF}}$	$AP_{\texttt{id}}$	$AP_{\texttt{col}}$
	ausaz	0,000	0,309	0,070	0,234
AM	t neurria	0,197	0,455	0,084	0,383
	χ^2	(-0,037)	0,302	0,119	0,206
DSim	L2_Indri_hit_erl_NV	**0,322**	**0,566**	0,294	0,354
	L2_KL_hit_erl_NV	0,309	0,551	**0,320**	0,332
	L2_Indri_rankV_hazt	0,320	0,552	0,130	0,431
MSFlex	CPMI_osag_mugat	0,154	0,432	0,113	0,329
	H_{rel}_mugat	0,119	0,434	0,134	0,313
	CPMI_orok_ADJ	(0,045)	0,387	0,202	0,237
	CPMI_izena_mugat	0,132	0,425	0,101	0,331
LFlex	z_PMI_V_WNhedap	0,110	0,365	0,120	0,259
	R_{nv}_EBantzdistr_20	0,066	0,381	0,066	0,323
	z_PMI_V_ELHWNhedap	0,108	0,357	0,122	0,250

Idiomatikotasunaren lau osagaiak neurtzeko egindako esperimentu bakunen emaitza onenen laburpena. Kendall τ_B koefizientea (idiomatikotasun-ranking ideal batekiko hein-korrelazioa); eta batez besteko doitasunak (AP), UFen rankingerako ($AP_{\texttt{UF}}$), oro har, eta AP espezifikoak esapide idiomatikoetarako eta kolokazioetarako.

Esperimentu bakunen emaitza hautatuak

	neurria	τ_B	$AP_{\texttt{UF}}$	$AP_{\texttt{id}}$	$AP_{\texttt{col}}$
	ausaz	0,000	0,309	0,070	0,234
AM	t neurria	0,197	0,455	0,084	0,383
	χ^2	(-0,037)	0,302	0,119	0,206
DSim	L2_Indri_hit_erl_NV	**0,322**	**0,566**	0,294	0,354
	L2_KL_hit_erl_NV	0,309	0,551	**0,320**	0,332
	L2_Indri_rank**V**_hazt	0,320	0,552	0,130	**0,431**
MSFlex	CPMI_osag_mugat	0,154	0,432	0,113	0,329
	H_{rel}_mugat	0,119	0,434	0,134	0,313
	CPMI_orok_ADJ	(0,045)	0,387	0,202	0,237
	CPMI_izena_mugat	0,132	0,425	0,101	0,331
LFlex	z_PMI_V_WNhedap	0,110	0,365	0,120	0,259
	R_{nv}_EBantzdistr_20	0,066	0,381	0,066	0,323
	z_PMI_V_ELHWNhedap	0,108	0,357	0,122	0,250

Idiomatikotasunaren lau osagaiak neurtzeko egindako esperimentu bakunen emaitza onenen laburpena. Kendall τ_B koefizientea (idiomatikotasun-ranking ideal batekiko hein-korrelazioa); eta batez besteko doitasunak (AP), UFen rankingerako ($AP_{\texttt{UF}}$), oro har, eta AP espezifikoak esapide idiomatikoetarako eta kolokazioetarako.

Esperimentu bakunen emaitza hautatuak

	neurria	τ_B	$AP_{\texttt{UF}}$	$AP_{\texttt{id}}$	$AP_{\texttt{col}}$
	ausaz	0,000	0,309	0,070	0,234
AM	t neurria	0,197	0,455	0,084	0,383
	χ^2	(-0,037)	0,302	0,119	0,206
DSim	L2_Indri_hit_erl_NV	**0,322**	**0,566**	0,294	0,354
	L2_KL_hit_erl_NV	0,309	0,551	**0,320**	0,332
	L2_Indri_rank**V**_hazt	0,320	0,552	0,130	**0,431**
MSFlex	CPMI_osag_mugat	0,154	0,432	0,113	0,329
	H_{rel}_mugat	0,119	0,434	0,134	0,313
	CPMI_orok_ADJ	(0,045)	0,387	**0,202**	0,237
	CPMI_izena_mugat	0,132	0,425	0,101	0,331
LFlex	z_PMI_V_WNhedap	0,110	0,365	0,120	0,259
	R_{nv}_EBantzdistr_20	0,066	0,381	0,066	0,323
	z_PMI_V_ELHWNhedap	0,108	0,357	0,122	0,250

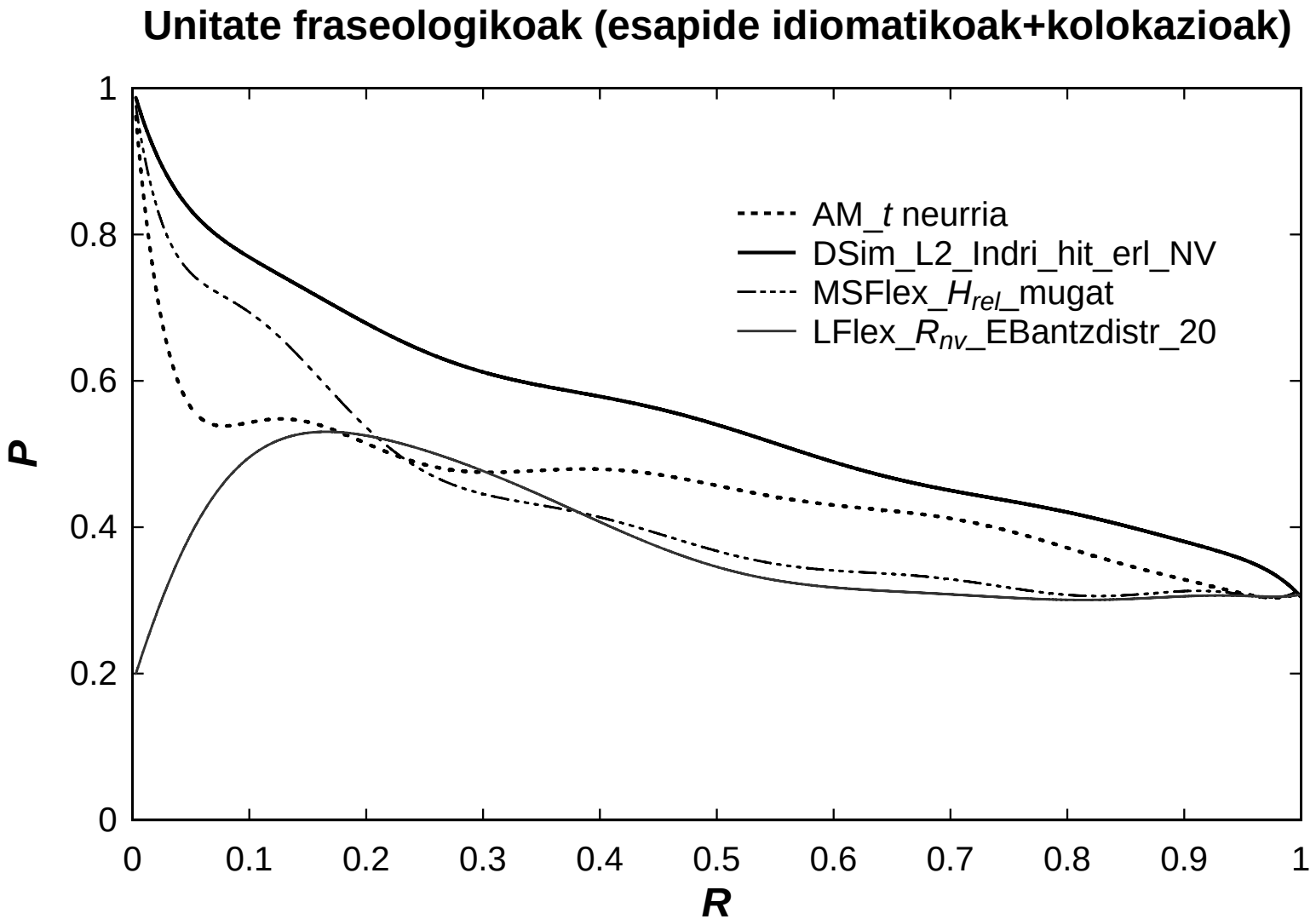
Idiomatikotasunaren lau osagaiak neurtzeko egindako esperimentu bakunen emaitza onenen laburpena. Kendall τ_B koefizientea (idiomatikotasun-ranking ideal batekiko hein-korrelazioa); eta batez besteko doitasunak (AP), UFen rankingerako ($AP_{\texttt{UF}}$), oro har, eta AP espezifikoak esapide idiomatikoetarako eta kolokazioetarako.

Esperimentu bakunen emaitza hautatuak

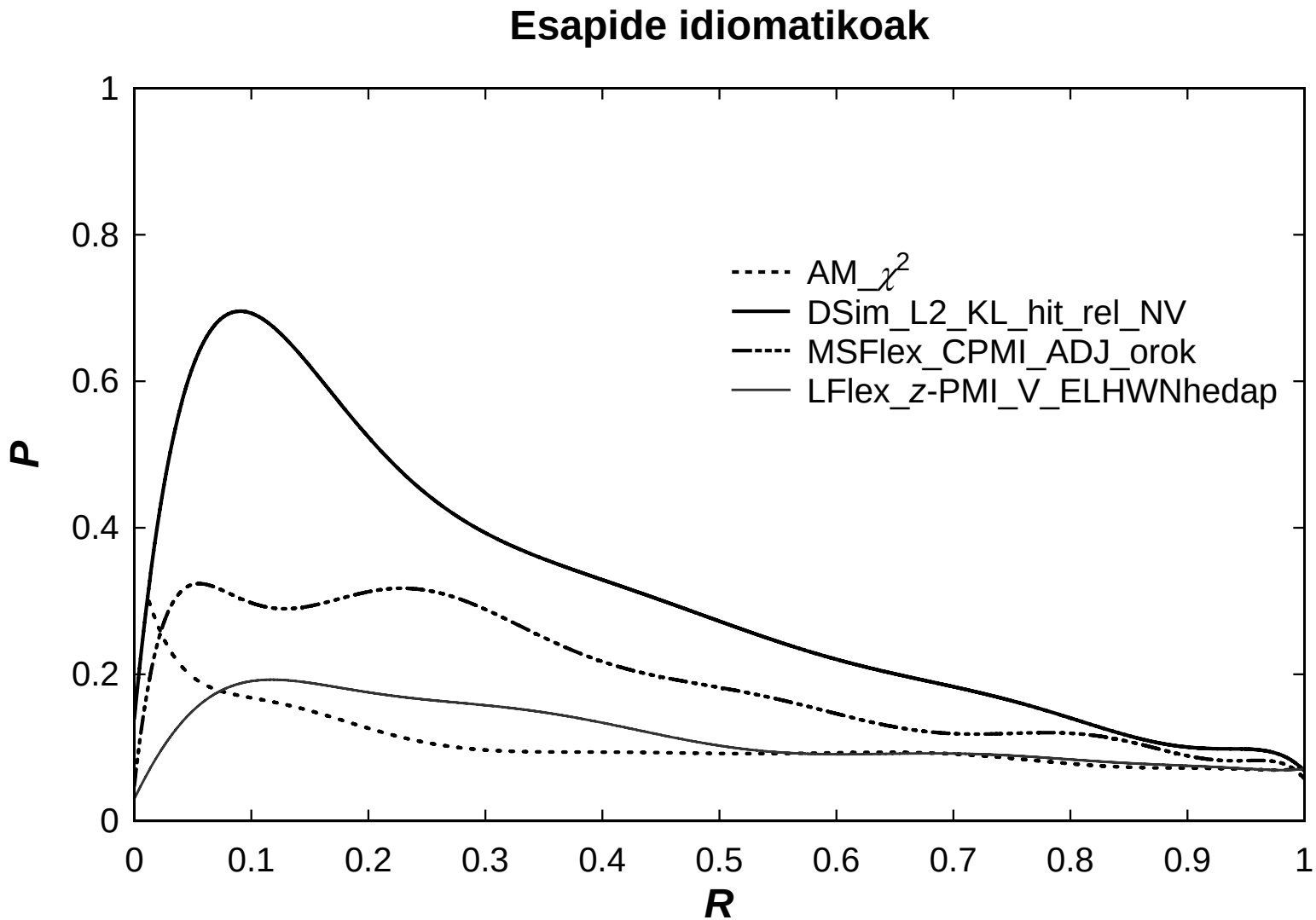
	neurria	τ_B	$AP_{\texttt{UF}}$	$AP_{\texttt{id}}$	$AP_{\texttt{col}}$
	ausaz	0,000	0,309	0,070	0,234
AM	t neurria	0,197	0,455	0,084	**0,383**
	χ^2	(-0,037)	0,302	0,119	0,206
DSim	L2_Indri_hit_erl_NV	**0,322**	**0,566**	0,294	0,354
	L2_KL_hit_erl_NV	0,309	0,551	**0,320**	0,332
	L2_Indri_rank**V**_hazt	0,320	0,552	0,130	**0,431**
MSFlex	CPMI_osag_mugat	0,154	0,432	0,113	0,329
	H_{rel}_mugat	0,119	0,434	0,134	0,313
	CPMI_orok_ADJ	(0,045)	0,387	**0,202**	0,237
	CPMI_izena_mugat	0,132	0,425	0,101	0,331
LFlex	z_PMI_V_WNhedap	0,110	0,365	0,120	0,259
	R_{nv}_EBantzdistr_20	0,066	0,381	0,066	0,323
	z_PMI_V_ELHWNhedap	0,108	0,357	0,122	0,250

Idiomatikotasunaren lau osagaiak neurtzeko egindako esperimentu bakunen emaitza onenen laburpena. Kendall τ_B koefizientea (idiomatikotasun-ranking ideal batekiko hein-korrelazioa); eta batez besteko doitasunak (AP), UFen rankingerako ($AP_{\texttt{UF}}$), oro har, eta AP espezifikoak esapide idiomatikoetarako eta kolokazioetarako.

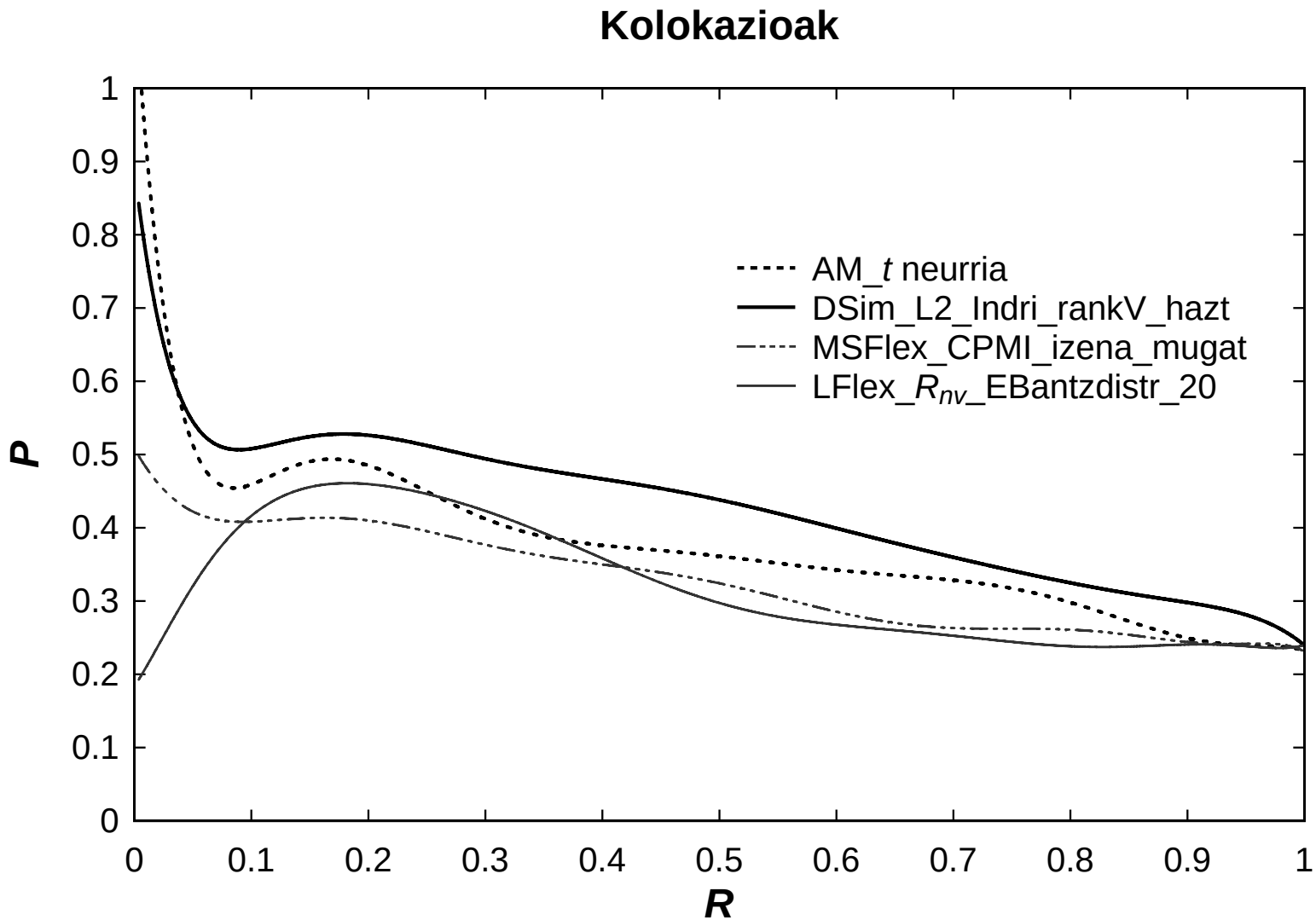
Emaitzak – P/R kurbak – UFak



Emaitzak – P/R kurbak – esapide idiomatikoak



Emaitzak – P/R kurbak – kolokazioak



Esperimentu bakunen emaitzen analisia

Emaitzen alderdi nagusiak

- Antzekotasun distribuzionaleko neurriak (DSim) dira idiomatikotasun-mailarekin τ_B korrelazio onena dutenak. Emaitza onenak IR tresna batekin lortu dira, L2 modalitateko esperimentuetan
- Esapide idiomatikoen erauzketan ($AP_{\texttt{id}}$) nabari da batez ere DSim neurrien gailentasuna
- Kolokazio-erauzketan, emaitza onena aditzaren semantika neurtzen duen Indri indize batek eman du (L2_Indri_rankV_hazt), kolokazioen ezaugarri aipatuenetakoa den idiosinkrasia estatistikoa neurtzen duten AMen gainetik
- AMak: emaitza hobeak kolokazioekin esapide idiomatikoekin baino
- MSFlex neurketak, oro har, bigarren onenak dira $AP_{\texttt{id}}$-ren emaitzetan, baina $AP_{\texttt{col}}$-en emaitzak balio apalenen artean daude
- LFlex: emaitza eskasenak, espero baino txarragoak

Ikasketa automatikoa

Esperimentuen diseinua

- **Sei metodo** (Weka toolkit): Naive Bayes, j48 (C4.5 erabaki-zuhaitza), Random Forest (RF), PART, SMO (SVM edo sostengu-bektoreen makinen inplementazio bat) eta Logistic Regression (LR)
- Ereduaren balidazioa: ebaluazio-erreferentzia ez da oso handia (1 145 instantzia) $\rightarrow$ balidazio gurutzatua (*cross-validation*)

Ikasketa automatikoa

Esperimentuen diseinua

- **Sei metodo** (Weka toolkit): Naive Bayes, j48 (C4.5 erabaki-zuhaitza), **Random Forest** (RF), PART, **SMO** (SVM edo sostengu-bektoreen makinen inplementazio bat) eta **Logistic Regression** (LR)
- Ereduaren balidazioa: ebaluazio-erreferentzia ez da oso handia (1 145 instantzia) $\rightarrow$ balidazio gurutzatua (*cross-validation*)

Ikasketa automatikoa

Esperimentuen diseinua

- Sei metodo (Weka toolkit): Naive Bayes, j48 (C4.5 erabaki-zuhaitza), Random Forest (RF), PART, SMO (SVM edo sostengu-bektoreen makinen inplementazio bat) eta Logistic Regression (LR)
- Ereduaren balidazioa: ebaluazio-erreferentzia ez da oso handia (1 145 instantzia) $\rightarrow$ **balidazio gurutzatua** (*cross-validation*)

Datu-multzoak

- `AM`, `DSim`, `MSFlex` eta `LFlex`: propietate bakoitzaren neurketak
  Atributu-kopuruak: AM 7; DSim 40; MSFlex 34; LFlex 20
- `4 osag.`: propietate guztien neurketak, batera
- `4 osag.+ad.`: propietate guztien neurketak, eta konbinazioaren aditza, Fazlyri (2007) jarraituz, eta korrelazioa egiaztatu ondoren
- Atributu-hautaketa automatikoa
  - `CS-BF`: Wekaren *AttributeSelectedClassifier* metasailkatzailean, osagai hauek dituen iragazkia: `CfsSubsetEval1` ebaluatzailetzat, eta `BestFirst` bilaketa-metodotzat. Guztira, 37 atributu hautatu ditu iragazki horrek: AM 3; DSim 16; MSFlex 7; LFlex 2; 9 aditz.

Ikasketa automatikoko esperimentuen emaitzak

Atrib.	Metod.	CCI	$F_{\texttt{id}}$	$F_{\texttt{col}}$	$F_{\texttt{free}}$	F_{mikro}	F_{makro}
Oinarri-lerroa		69,607	0,000	0,000	0,821	0,571	0,274
DSim	LR	74,061	0,270	0,468	0,842	0,714	0,527
	SMO	69,607	0,046	0,063	0,820	0,589	0,310
	RF	71,441	0,279	0,438	0,822	0,694	0,513
4 osag.	LR	73,362	0,355	0,487	0,837	0,722	0,560
	SMO	76,070	0,300	0,479	0,858	0,731	0,546
	RF	73,712	0,336	0,475	0,840	0,719	0,550
4 osag.+ad.	LR	62,795	0,274	0,490	0,751	0,656	0,505
	SMO	76,856	0,418	0,544	0,858	0,754	0,607
	RF	73,974	0,304	0,447	0,843	0,713	0,531
CS-BF	LR	75,721	0,339	0,487	0,854	0,732	0,560
	SMO	73,450	0,149	0,390	0,838	0,685	0,459
	RF	72,838	0,364	0,435	0,836	0,709	0,545

Ikasketa automatikoko esperimentuen emaitzak

Atrib.	Metod.	CCI	$F_{\texttt{id}}$	$F_{\texttt{col}}$	$F_{\texttt{free}}$	F_{mikro}	F_{makro}
Oinarri-lerroa		69,607	0,000	0,000	0,821	0,571	0,274
DSim	LR	74,061	0,270	0,468	0,842	0,714	0,527
	SMO	69,607	0,046	0,063	0,820	0,589	0,310
	RF	71,441	0,279	0,438	0,822	0,694	0,513
4 osag.	LR	73,362	0,355	0,487	0,837	0,722	0,560
	SMO	76,070	0,300	0,479	0,858	0,731	0,546
	RF	73,712	0,336	0,475	0,840	0,719	0,550
4 osag.+ad.	LR	62,795	0,274	0,490	0,751	0,656	0,505
	SMO	**76,856**	**0,418**	**0,544**	**0,858**	**0,754**	**0,607**
	RF	73,974	0,304	0,447	0,843	0,713	0,531
CS-BF	LR	75,721	0,339	0,487	0,854	0,732	0,560
	SMO	73,450	0,149	0,390	0,838	0,685	0,459
	RF	72,838	0,364	0,435	0,836	0,709	0,545

Ikasketa automatikoko esperimentuen emaitzak

Atrib.	Metod.	CCI	$F_{\texttt{id}}$	$F_{\texttt{col}}$	$F_{\texttt{free}}$	F_{mikro}	F_{makro}
Oinarri-lerroa		69,607	0,000	0,000	0,821	0,571	0,274
DSim	LR	74,061	0,270	0,468	0,842	0,714	0,527
	SMO	69,607	0,046	0,063	0,820	0,589	0,310
	RF	71,441	0,279	0,438	0,822	0,694	0,513
4 osag.	LR	73,362	0,355	0,487	0,837	0,722	0,560
	SMO	76,070	0,300	0,479	0,858	0,731	0,546
	RF	73,712	0,336	0,475	0,840	0,719	0,550
4 osag.+ad.	LR	62,795	0,274	0,490	0,751	0,656	0,505
	SMO	**76,856**	**0,418**	**0,544**	**0,858**	**0,754**	**0,607**
	RF	73,974	0,304	0,447	0,843	0,713	0,531
CS-BF	LR	**75,721**	**0,339**	**0,487**	**0,854**	**0,732**	**0,560**
	SMO	73,450	0,149	0,390	0,838	0,685	0,459
	RF	72,838	0,364	0,435	0,836	0,709	0,545

Sailkapen automatikoaren emaitzen analisia

Emaitzen alderdi nagusiak

- Ezagutza-iturri bakarreko lau datu-multzoetatik, `DSim` da emaitza onenak dituena, sailkatzaile guztietan (onena, LR)
- Oro har, algoritmo guztiek emaitza hobeak dituzte idiomatikotasunaren ezagutza-iturriak konbinatuz (`4 osag.` eta `4 osag.+ad.` datu-multzoak), iturri bakarra erabiliz baino (onena: SVM familiako SMO algoritmoa)
- `CS-BF` metasailkatzailearen LR esperimentuetako F_{mikro} eta F_{makro} (0,732 / 0,560) dira gehien hurbiltzen direnak SMO metodoak `4 osag.+ad.` datu-multzoarekin lortutako emaitza onenetara (0,754 / 0,607)

Predikatu konplexu batzuk birsailkatzearen eragina – Planteamendua

- Ebaluazio-erreferentzian badira konbinazio batzuk (esaterako, *jaramon egin, zor izan*), esapide erdiidiomatikotzat gabe, esapide idiomatikotzat jotzekoak liratekeenak. 17 dira. Horiek birsailkatuta: `id` 97; `col` 251; `free` 797
- Aurreikusitako ezaugarri nagusiak:
  - Maiztasun handikoak (ohiko `id` kategoriakoak baino handiagokoak)
  - Semantikoki erdikonposizionalak (izenak bere oinarrizko esanahia atxikitzen du)
  - Idiosinkrasia morfosintaktikoa (adib., forma kanonikoan, izena mugatzailerik gabe: *jaramon egin, zor izan*)
- Esperimentu bakun eta konbinatu guztiak egin ditugu sailkapen berriarekin

Predikatu konplexu batzuk birsailkatzearen eragina – Emaitzak

Esperimentu bakunak

- τ_B: igoera apalak (unitate gutxi birsailkatu dira)
- AM neurrien $AP_{\texttt{id}}$-ren igoera (batez ere, f-rekiko korrelazioa duten AMenak) $\rightarrow$ birsailkatzean, `id` kategoriakoen idiosinkrasia estatistikoa handitu egin da
- DSim: $AP_{\texttt{id}}$-ren balio hobexeagoak, aditzaren semantikarekiko neurketetan $\rightarrow$ birsailkatuek profil erdikonposizionala dute
- MSFlex neurketetan nabari da gehien $AP_{\texttt{id}}$-ren balioen igoera $\rightarrow$ birsailkatuak morfosintaktikoki idiosinkratikoak dira

Ikasketa automatikoa

- Emaitzak, oro har, zertxobait apalagoak. Baina F_{id} hobeak, instantzia gehiago baitaude ikasteko
- CS-BF iragazkiak, proportzionalki, MSFlex atributu gehiago aukeratu ditu aurreko sailkapenarekin baino

$\rightarrow$ Horrelako predikatu konplexuen profil berezia

Aurkezpenaren eskema

Ondorioak

- Idiomatikotasunaren izaera konplexua
- Idiosinkrasia semantikoaren nabarmentasuna, eta konposizionaltasunaren gradualtasuna
- Kolokazioen erdikonposizionaltasuna
- Kolokazioen malgutasun morfosintaktiko handia
- UF-kategoriaren eta konbinazioaren aditzaren arteko korrelazioa erabilgarria da ikasketa automatikoan
- Predikatu konplexu batzuen (*jaramon egin, zor izan* modukoen) profil berezia
- Teoria fraseologikoaren aurresanak eta lan esperimental batzuen emaitzak kontuan izanik, LFlex esperimentuen emaitzak ez dira espero genuen mailakoak

LFlex esperimentuen emaitzak esplikatzeko hipotesi batzuk

- Ordezkagarritasuna egiaztatzeko erabili diren euskarazko baliabideak estaldura gutxikoak izatea
- Euskara, batez ere kazetaritza-erregistroan, gutxiago finkatua izan liteke, eta, beraz, kolokatiboaren aukera lexikala ez litzateke hain argi finkatua
- LFlex emaitza konparatiboki onak lortu diren ikerlanetan kosinua erabili da DSim neurritzat, eta neurri horrek aski emaitza apalak izan ditu gure ikerketan ere

Ekarpenak

- Corpus handi bat prozesatu da (74 milioi hitz), eta bigrama-bilduma baliagarria erauzi
- Ebaluaziorako erreferentziak: hiztegi-erreferentzia eta eskuz landutako erreferentzia sailkatua
- `izena+aditza` osaerako konbinazioak automatikoki erauzteko prozedura
- Idiomatikotasunaren propietateak neurtzeko teknikak euskararako egokitu eta garatu ditugu
- Idiosinkrasia semantikoa neurtzeko, IR arloko teknikak erabili ditugu, eta horien bidez lortu ere emaitza onenak
- Sailkatzaile automatikoak trebatu eta ebaluatu dira; ekarpen handiena egiten duten propietateen informazioa eskuratu dugu
- Garatutako metodoa erraz egokitu daiteke beste egitura morfosintaktiko batzuetako bigramak erauzteko (`izena+izena`, `izena+izenondoa`, `izenlaguna+izena`...); Aplikazioa: Elhuyar Fundazioaren *Web-corpusen Atari*ko “Hitz-konbinazioak” atala[1]

[1] `http://webcorpusak.elhuyar.org/cgi-bin/kolokatuak.py`

Etorkizuneko lanak

- Informazio sintaktiko aberatsagoa erabiltzea: Maltixa mendekotasun-analizatzailea (Bengoetxea eta Gojenola, 2010)
- Interpretazio literala eta idiomatikoa izan ditzaketen konbinazioen agerpenak bereiztea
- Beste corpus-mota batzuk erabiltzea (literarioak, espezializate-arlokoak, web-corpusak), arloak eta erregistroak LFlex neurketetan duten eragina aztertzeko
- Corpus paraleloak erabiltzea, bi hizkuntzatako UF bikote baliokideak erauzteko eta karakterizatzeko
- UFak erauzi eta karakterizatzeko garatuko diren tresnetan aplikatzea. Tresna horiek hiztegigintzan, itzulpengintzan eta hizkuntzaren prozesamendu automatikoko hainbat atazatan izango dute aplikazioa

Idiomatikotasunaren karakterizazio automatikoa: `izena+aditza` konbinazioak

Doktoregaia: Antton Gurrutxaga Hernaiz

Zuzendariak: Iñaki Alegria Loinaz, Xabier Artola Zubillaga

Lengoaia eta Sistema Informatikoak Saila
Euskal Herriko Unibertsitatea / Universidad del País Vasco

2014ko uztailaren 21a

Argitalpenak

- Gurrutxaga, A. eta Alegria, I. (2011). Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques. *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, 2–7 or. Portland, Oregon: Association for Computational Linguistics.
- Gurrutxaga, A. eta Alegria, I. (2012). Measuring the compositionality of NV expressions in Basque by means of distributional similarity techniques. *Proceedings of the8th International Conference on Language Resources and Evaluation (LREC 2012)*, 2389–2394 or. Istanbul: ELRA.
- Gurrutxaga, A. eta Alegria, I. (2013). Combining different features of idiomaticity for the automatic classification of noun+ verb expressions in Basque. *Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013) NAACL–HLT 2013*, 116–125 or. Atlanta, Georgia: Association for Computational Linguistics.

Bibliografia

- Altzibar, X. (2005). Kolokazioak euskaraz. Zer axola duten kazetaritzan. *Euskarazko kazetaritzaren I. kongresua. Kazetaritza euskaraz: oraina eta geroa*, 383–395. UPV/EHU.
- Amosova,N.Ñ. (1963). *Osnovui anglijskoy frazeologii*. Leningrad: University Press.
- Baldwin, T. eta Kim, S. (2010). Multiword expressions. Indurkhya, N. eta Damerau, F. J.(ed.), *Handbook of Natural Language Processing, second edition*, 267–292. CRC Press, Taylor and Francis Group, Boca Raton, AEB.
- Bannard, C., Baldwin, T. eta Lascarides, A. (2003). A statistical approach to the semantics of verb-particles. *Proceedings of the ACL 2003 Workshop on Multiword Expressions: analysis, acquisition and treatment* 18. lib., 65–72. Association for Computational Linguistics.
- Bannard, C. (2007). A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, 1–8. Association for Computational Linguistics.
- Barkema, H. (1994a). Determining the syntactic flexibility of idioms. *Realising and Using English Language Corpora*, 39–52.
- Bengoetxea, K. eta Gojenola, K. (2010). Application of different techniques to dependency parsing of Basque. *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Language*s, 31–39. Association for Computational Linguistics.
- Berry-Rogghe, G. (1974). Automatic identification of phrasal verbs. *Computers in the Humanities*, 16–26.

Bibliografia

- Burger, H. (1998). *Phraseologie: eine Einführung am Beispiel des Deutschen*. Berlin: Erich Schmidt.
- Corpas Pastor, G. (1996). *Manual de Fraseología Espaňola*. Gredos, Madrid.
- Cowie, A. P. (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics*, 2(3):223–235.
- Cowie, A. P. (1998). Phraseological dictionaries: some east-west comparisons. Cowie, A., (ed.), *Phraseology: Theory, Analysis, and Applications*, 209–228. Oxford University Press, USA.
- Evert, S. (2005). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Doktorego-tesia, University of Stuttgart.
- Fazly, A. eta Stevenson, S. (2007). Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, 9–16. Association for Computational Linguistics.
- Fernando, C. eta Flavell, R. (1981). *On Idiom: Critical Views and Perspectives*. University of Exeter.
- Fontenelle, T. (1998). Discovering significant lexical functions in dictionary entries. Cowie, A., (ed.), *Phraseology: Theory, Analysis, and Applications*, 189–207. Oxford University Press, USA.
- Gläser, R. (1988). The grading of idiomaticity as a presupposition for a taxonomy of idioms. *Understanding the lexicon: Meaning, sense and world knowledge in lexical semantics*, 264–279.

Bibliografia

- Gläser, R. (1998). The stylistic potential of phraseological units in the light of genre analysis. Cowie, A., (ed.), *Phraseology: Theory, Analysis, and Applications*, 125–143. Oxford University Press, USA.
- Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.
- Hausmann, F. J. (1989). Le dictionnaire de collocations. *Wörterbücher, Dictionaries*, 1:1010–1019.
- Heid, U. (1994). On ways words work together - Research topics in lexical combinatorics. *Proceedings of the 6th International Congress of Lexicography - EURALEX '94*, 226–257.
- Howarth, P. A. (1996). *Phraseology in English academic writing: Some implications for language learning and dictionary making* 75. lib. Walter de Gruyter.
- Katz, G. eta Giesbrecht, E. (2006). Automatic identification of non-compositional multi-word expressions using Latent Semantic Analysis. *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, 12–19. Association for Computational Linguistics.
- Korkontzelos, I. eta Manandhar, S. (2009). Detecting compositionality in multi-word expressions. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 65–68. Association for Computational Linguistics.
- Krcmár, L., Jezek, K. eta Pecina, P. (2013). Determining compositionality of word expressions using word space models. *Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013) NAACL HLT 2013*, 13:42–50.
- Krenn, B. (2004). Manual zur Identifikation von Funktionsverbgefügen und figurativen Ausdrücken in PP-Verb-Listen. *Austrian Research Institute for Artificial Intelligence*.

Bibliografia

- Lin, D. (1998). Automatic retrieval and clustering of similar words. *Proceedings of the 17th International Conference on Computational Linguistics* 2. lib., 768–774. Association for Computational Linguistics.
- Lin, J., Li, S. eta Cai, Y. (2008). A new collocation extraction method combining multiple association measures. *Proceedings of International Conference on Machine Learning and Cybernetics* 1. lib., 12–17. IEEE.
- Mel'čuk, I. (1988). Semantic description of lexical units in an explanatory combinatorial dictionary: Basic principles and heuristic criteria. *International Journal of Lexicography*, 1(3):165–188.
- Mel'čuk, I. eta Wanner, L. (1994). Towards an efficient representation of restricted lexical cooccurrence. *Proceedings of the 6th International Congress of Lexicography - EURALEX '94* 94. lib.
- Mitchell, J. eta Lapata, M. (2008). Vector-based models of semantic composition. *Proceedings of ACL-08: HLT*, 236–244, Columbus, Ohio, AEB. Association for Computational Linguistics.
- Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus-based Approach*. Clarendon Press Oxford.
- Pawley, A. eta Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. *Language and Communication*, 191:225.
- Pecina, P. eta Schlesinger, P. (2006). Combining association measures for collocation extraction. *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, 651–658. Association for Computational Linguistics.

Bibliografia

- Pedersen, T., Banerjee, S., McInnes, B. T., Kohli, S., Joshi, M. eta Liu, Y. (2011). The Ngram Statistics Package (text::NSP): A flexible tool for identifying Ngrams, collocations, and word associations. *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, 131–133. Association for Computational Linguistics.
- Reddy, S., McCarthy, D., Manandhar, S., eta Gella, S. (2011). Exemplar-based word-space model for compositionality detection: Shared task system description. *Proceedings of the Workshop on Distributional Semantics and Compositionality*, 54–60. Association for Computational Linguistics.
- Ruiz Gurillo, L. (1998). Una clasificación no discreta de las unidades fraseológicas del español. Wotjak, G., (ed.), *Estudios de fraseología y fraseografía del español actual*, 13–37. Lingüística Iberoamericana.
- Salvador, V. (2000). Idiomaticitat i discurs prefabricat. Salvador, V. eta Piquer, A.(ed.), *El discurs prefabricat*, 19–31. Universitat Jaume I.
- Sinclair, J. (1996). The search for units of meaning. *Textus*, 9(1):75–106.
- Urizar, R. (2012). *Euskal lokuzioen tratamendu konputazionala*. Doktorego-tesia, Informatika Fakultatea, UPV/EHU, Donostia.
- Van de Cruys, T. eta Moirón, B. (2007). Semantics-based multiword expression extraction. *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, 25–32. Association for Computational Linguistics.
- Venkatapathy, S. eta Joshi, A. K. (2005). Measuring the relative compositionality of verb-noun (VN) collocations by integrating features. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 899–906. Association for Computational Linguistics.

Bibliografia

- Vinogradov, V. (1947). Ob osnovnuikh tipakh frazeologicheskikh edinits v russkom yazuike. A .A. Shakhmatov, 1864–1920. *Sbornik statey i materialov*, 339–364. Mosku: Nauka.
- Warren, B. (2005). A model of idiomaticity. *Nordic Journal of English Studies*, 4(1):35–54.
- Wulff, S. (2008). *Rethinking Idiomaticity*. Corpus and Discourse. Continuum International Publishing Group Ltd, New York.
- Zabala, I. (2004). Los predicados complejos en vasco. Zabala, I., Pérez Gaztelu, E. eta García, L.(ed.), *Las fronteras de la composición en lenguas románicas y en vasco*, 445–567. Universidad de Deusto, Servicio de Publicaciones, Donostia.