Hizkuntzaren Industria antolatzeko urratsak

Kepa Sarasola

IXA taldea http://ixa.si.ehu.es

UPV-EHU

Universidad del País Vasco-Euskal Herriko Unibertsitatea

Motibazioa

- ***Hizkuntza-Teknologiak*** **funtsezkoak dira *Informazio eta Komunikazioaren Gizartea* esaten dugun horretan**
- **Epe ertainean pertsona eta makinen arteko komunikazioa geure hizkuntzan egin ahal izango dugu, ez makinaren hizkuntzan**
- **Tresna mugatuak izango dira, eta beti errore-maila batekin, baina, hala ere, laguntza ederra emango digute.**

Motibazioa

- **Gaur egun badira zenbait hizkuntza-aplikazio eskuragarri**
  - **Ortografia-zuzentzaileak eta estilo-zuzentzaileak**
  - **Hiztegi-kontsultak on-line**
  - **Itzulpen-laguntzak**
  - **Interneterako bilatzaileak**
  - **Hizketa testua bihurtzen duten sistemak**
  - **Testuak irakurtzen dutenak**
  - **Bigarren hizkuntza ikasteko sistemak**
  - **...**

Motibazioa

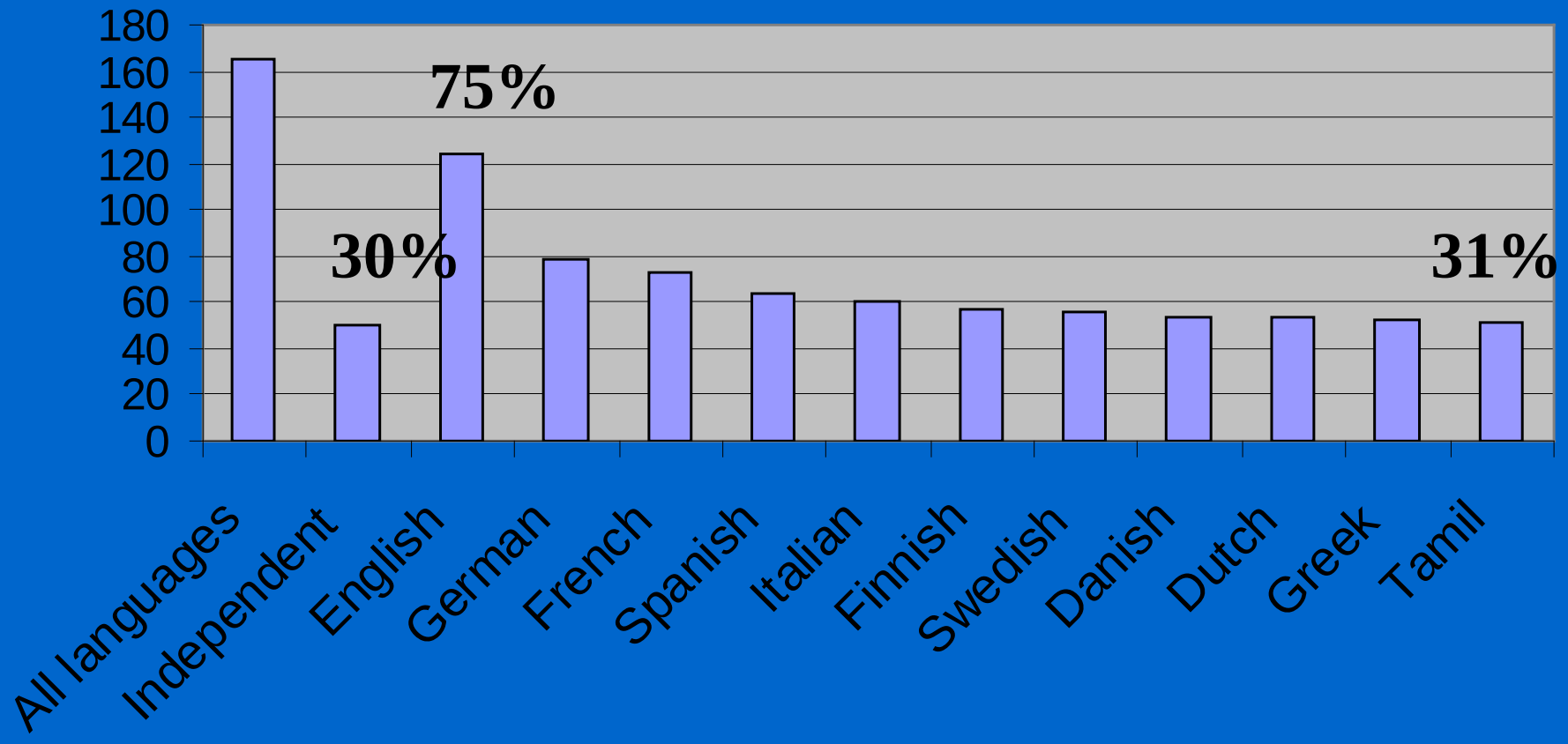
- **Baina horrelako sistema gehienak ingeleserako balio dute, ez beste hizkuntzetarako**

⇒ **Beste hizkuntzek ahalegin handia egin behar dute atzean ez gelditzeko**

⇒ **Are gehiago hizkuntza txikiek**

Motibazioa

NLSR:Hizkuntza tratatzeko programen katalogoa



Hizkuntzaren Industria antolatzeko urratsak

- Euskararen egoera orain
- Helburuak
- Estrategia
- Eragileak eta bezero posibleak
- Ondorioak

Euskararen softwarearen katalogoa
www.ueu.org/softkat

- **(9) BULEGO APLIKAZIOAK**
  - **Testu prozesatzaileak, kontabilitatea...**
- **(18) AISIALDIA**
  - **Musika,Jokuak...**
- **(33) HIZKUNTZA**
  - **Itzultzaileak, zuzentzaileak, hiztegiak...**
- **(12) INTERNETEN ARITZEKO**
  - **Nabigatzaileak, posta elektronikoko programak...**
- **(12) TRESNA OROKORRAK**
  - **Sistema eragileak, Interneteko datu-baseak eta bilatzaileak...**
- **(37) IRAKASKUNTZA ETA JOKU PEDAGOGIKOAK**
  - **Matematika, zientziak...**

Euskararako hizkuntza-aplikazioak

- **Ediziorako laguntzak:**
  - **Xuxen: zuzentzaile ortografikoa**
  - **Elhuyar hiztegia. Officeko plug-ina.**
- **Hizketaren tratamendua**
  - **BIZKAIFON (Bizkaieraren Fonoteka)**
  - **AhoTTS Testu-Ahots Bihurgailua**
- **Euskara ikasteko metodoak:**
  - **Bai & Bye / BOGA / HEZINET**
- **Lematizatzailea, informazioa bilatzeko tresna**
  - **Euslem**
- **Datu-base dokumentala**
  - **Kapsula**
- **Corpus**
  - **XX. mendeko euskararen corpus estatistikoa**
- **Baliabide lexikalak: hiztegiak, esamoldeak, ...**
  - **16 produktu**

Zer egin daiteke atzean ez geratzeko? Nola ekin erronka horri?

- **Proposamena:**
  - **Aurkezten dugu estrategia bat, urrats-kate bat hizkuntzaren teknologiari metodo batekin ekiteko.**
- **IXA taldearen 15 urteko ibilbidean oinarritua**
- **Nazioarteko foroetan aurkeztua eta kontrastatua**
- **Ideia nagusia:**
  - **Hasieran sortu oinarrizko baliabideak eta tresnak**
  - **Geroago sortu merkatu-aplikazioak**

  Alderantziz ez !

Hizkuntzaren Industria antolatzeko urratsak

- Euskararen egoera orain
- Helburuak
- Estrategia
- Eragileak eta bezero posibleak
- Ondorioak

Helburuak (epe erdian)

- **Euskararako baliabide linguistikoak sortzea**
  - **Corpus (100 Megahitz)**
  - **Lexikalak: Hiztegiak eta hitzen sailkapena adieraren arabera.**
- **Aplikazioak:**
  - **Informazio-bilatzaileak (eguraldi-parteak, burtsa, kirolak, berriak, bideo-eskaerak, irudi–eskaerak, ...)**
  - **Domotika**
  - **Itzulpen-automatikoa**
  - **Irakaskuntza-sistemak (e-learning)**
  - **Elbarrientzako laguntzak**
  - **Telebista digitala**
  - **Elkarrizketa-sistemak**
  - **Multimedia-sistemak**

Helburuak (2)

- **Ingeleserako produktuen merkatua oso handia da. Baina produktu horiek ez dira zabaldu modu egokian beste hizkuntzetarako.**

- **Guk euskaldunok, europarrok ohituta gaude eleaniztasunean bizi izaten. Gaitasun handiagoa dugu eleaniztasuna lantzeko.**

- **Euskara oso ezaugarri desberdinak ditu. Probaleku ezinhobea da produktuen moldagarritasuna frobatzeko.**

⇒ **Teknologia esportagarria eta nazioartekora ateratzeko modukoa**

Helburuak (3)

- **Ingeniaritza linguistikoan ikerketan eta Garapenean arituko den komunitatea sortu**
  - **Personal**
    - **2003: 120-150 2006: 400-500**
  - **Empresas/Agentes**
    - **2002: 35 2006: 50**

Konpartitzen dituena :

–Algoritmo eta programak
–Metodologiak
–Teknologia

Hizkuntzaren Industria antolatzeko urratsak

- Euskararen egoera orain
- Helburuak
- Estrategia
- Eragileak eta bezero posibleak
- Ondorioak

Lehentasunak (I)

- **Konpartitu eta berrerabili:**
  - **Teoriak, formalismoak, eta metodologiak**
  - **Teknikak eta eskarmentua**
  - **Teknologia**
- **Geure hizkuntzarako baliabide linguistikoak sortu**
- **Eta orduan:**
  - **Tresna orokor eta espezifikoak**
  - **Aplikazioak**

Lehentasunak(II)

Adibidea:

- OCR zenbait programatan euskarazko testuak tratatzeko balio dutela esaten da

Baina

- batek ere ez dauka informazio linguistikorik (hiztegirik, bigrama edo trigramen frekuentziak,…)
- Baten batean ŕ letra (r azentuduna) onatrzen dutelako esaten dute hori

Estrategia:
Hasieran baliabideak eta tresnak
Geroago merkatu-aplikazioak

Oinarri eta baliabide linguistikoak Tresnak eta aplikazioak

Produktu posibleen artean bereizten ditugu

- **Aplikazioak:**

  Azken erabiltzaile arruntentzako produktuak

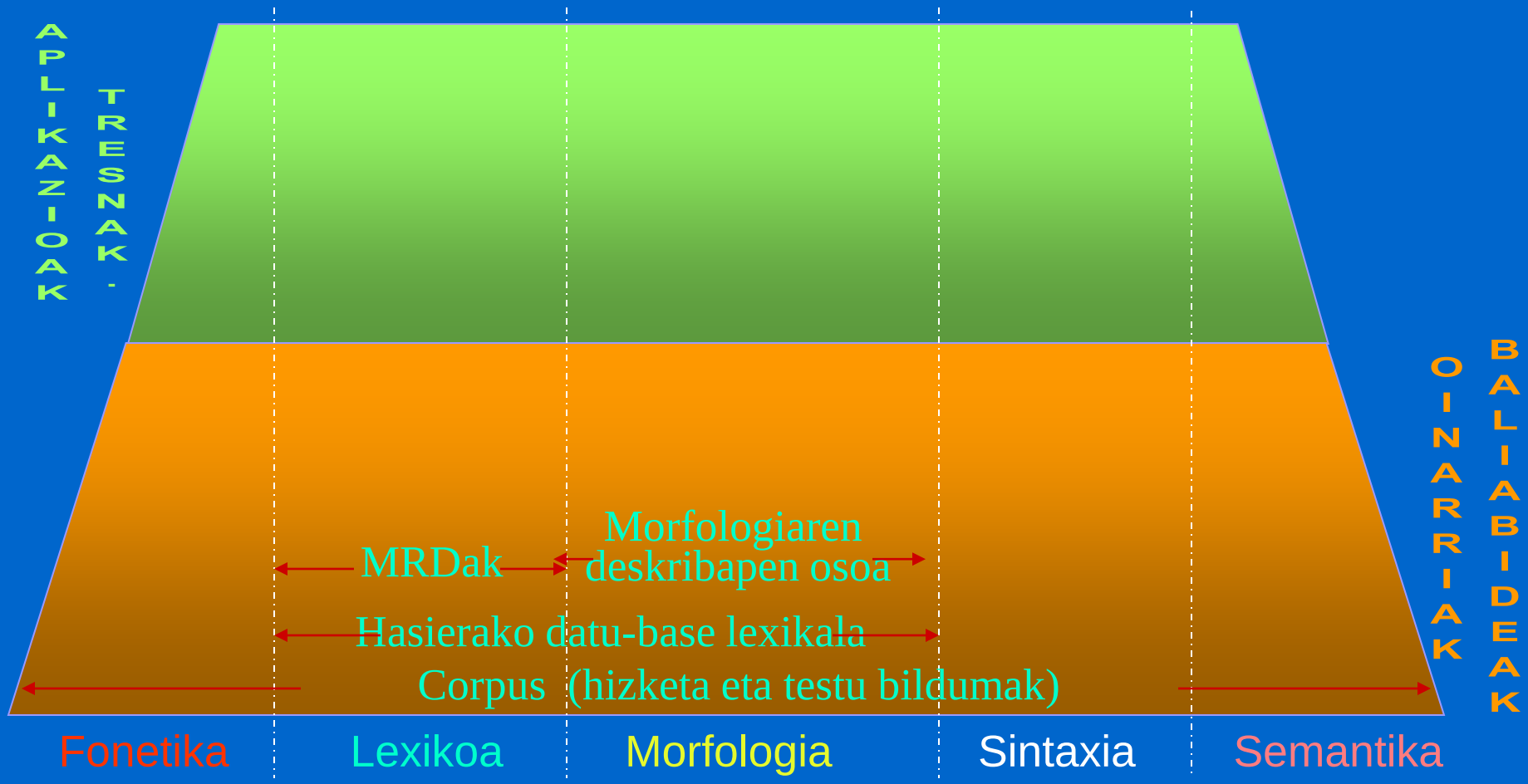
- **Tresna linguistikoak:**

  Adituentzat eta programa-garatzaileentzat

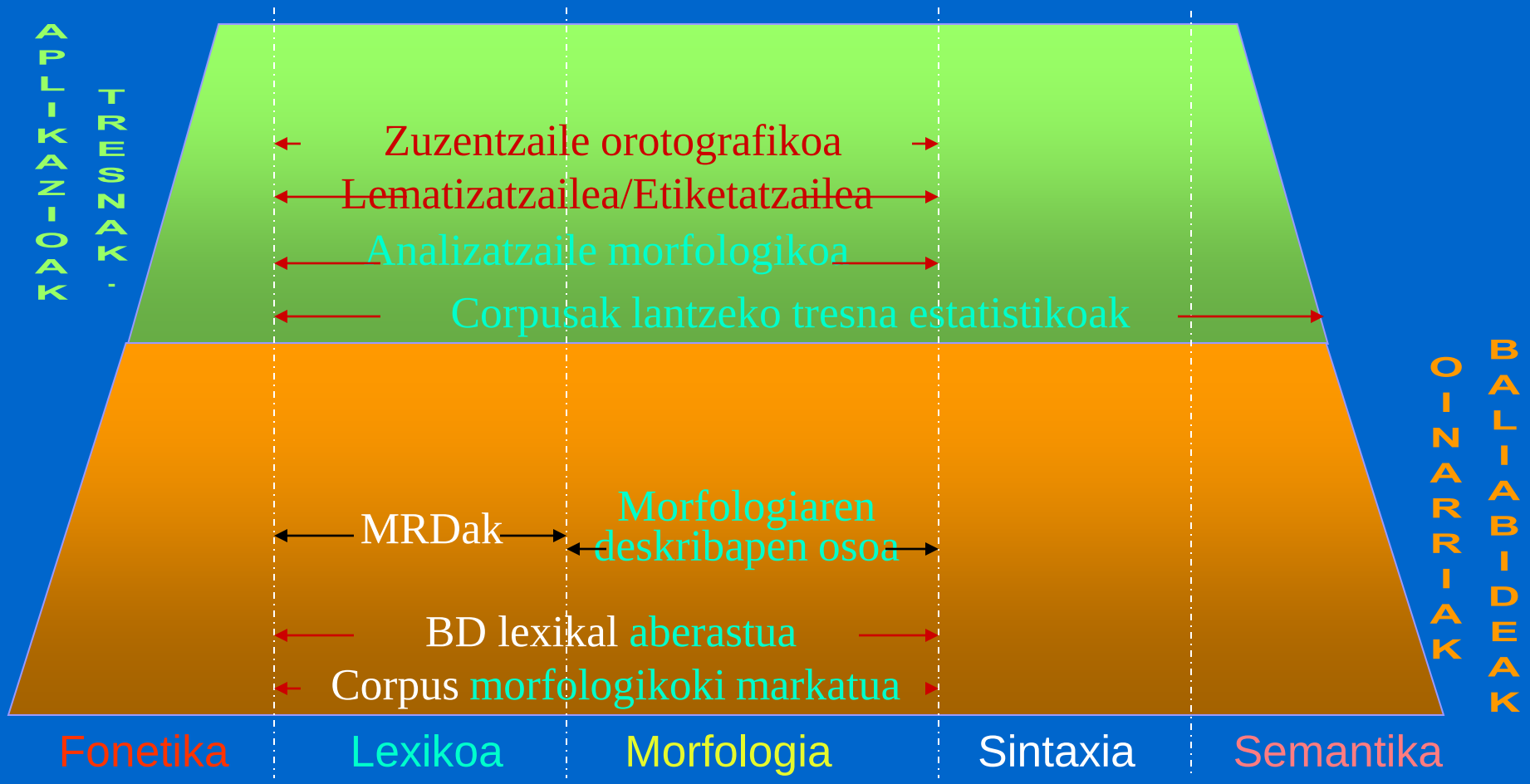
- **Oinarri eta baliabide linguistikoak**

  Hizkuntzaren tratamendu automatikorako ezinbestekoa den azpiegitura

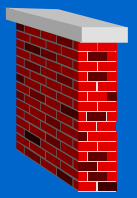
I fasea: Oinarri linguistikoak sortu



II fasea : Lehenengo tresnak eta aplikazioak

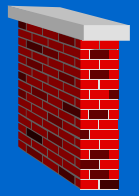


EDBL (Euskararen Datu-Base Lexikala)



- **Euskararen tratamendu automatikorako oinarri lexikala**
- **80.000 sarrera hiru ataletan:**
  - **hiztegi sarrera arruntak**
  - **aditz-formak**
  - **morfema ez independenteak**
- **Atsegina, eguneratua eta kontsistentea**
- **ORACLE V7 eta UNIXen pean**

Morfologiaren deskribapen osoa



- ***TWO LEVEL* formalismoan**
- **Morfema bakoitzaren atzetik zein etor daitekeen**
- **Aldaketa morfofonologikoak**

  Ad.: txakur + a → txakurra

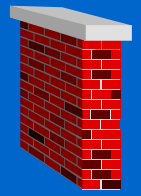
Analizatzaile morfologikoa

- **Hitz bakoitzaren analisi posible guztiak**
  Batezbeste 2.6 interpretazio (batzutan >5)
- **Hitz osoaren informazio morfosintaktikoa**
  etxekoek → elipsia, **ergatibo, plural, mugatua**
- **Zenbait errore tipiko eta aldaera dialektal**
  Ad.: *eritzi -> iritzi**)**
- **Batzutan analisi bitxiak: Amona → amon +a**
  → ama + on +a !?

Lematizatzailea. EUSLEM Adibidea

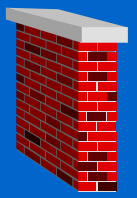
batzu	**- okerra-**	**batzuk(batzu)\DET_DZG**
irteten	**irten**	**\ADI_SIN**
,	**,**	**\KOMA**
beste	**beste**	**DET_DZG**
batzu	**- okerra-**	**batzuk(batzu)\DET_DZG**
sartzen	**sar**	**\ADI_SIN**
,	**,**	**\KOMA**
zerbaiten	**zerbait**	**IOR_IZG**
zain	**zain**	**\ADB_ADO**
zirudienez	**irudi**	**\ADT**
.	.	**\PUNTU**

Gramatikak (sintaxia)



- **2 formalismotan:**
  - **Baterakuntza Gramatikak (PATR-II)**
  - **Murriztapen Gramatika**
- **aplikazioak:**
  - **Informazio-bilaketa**
  - **gramatika-zuzentzailea**
  - **sintaxirako sistema tutorea**

PATR-II gramatika



- **Adibidea: “Zure aita sofa berdean dago .”**

```
Perpausa
    dago
            aditz-mota  nor
            kategoria aditz trinkoa (aditz laguntzailea)
    aita
            kasua absolutiboa
            pertsona 3
            numero singular
            determinazioa  mugatua
    sofa
            kasua inesiboa
            pertsona 3
            numero singular
            determinazioa  mugatua
```

HAIN (Hizkuntz Aplikazioetarako Ingurunea)

- **SGML eta XML markatze-lengoaian oinarrituta**
- **Tresnen erabilgarritasun modularra ziurtatzeko diseinua**
- **Tresna multzo integratua**
  - **EDBL: Datu-Base Lexikala**
  - **MORFEUS: Analizatzaile morfologikoa**
  - **EUSLEM: lematizatzaile etiketatzailea**
  - **XUXEN: zuzentzaile/egiaztatzaile ortografikoa**
  - **...**

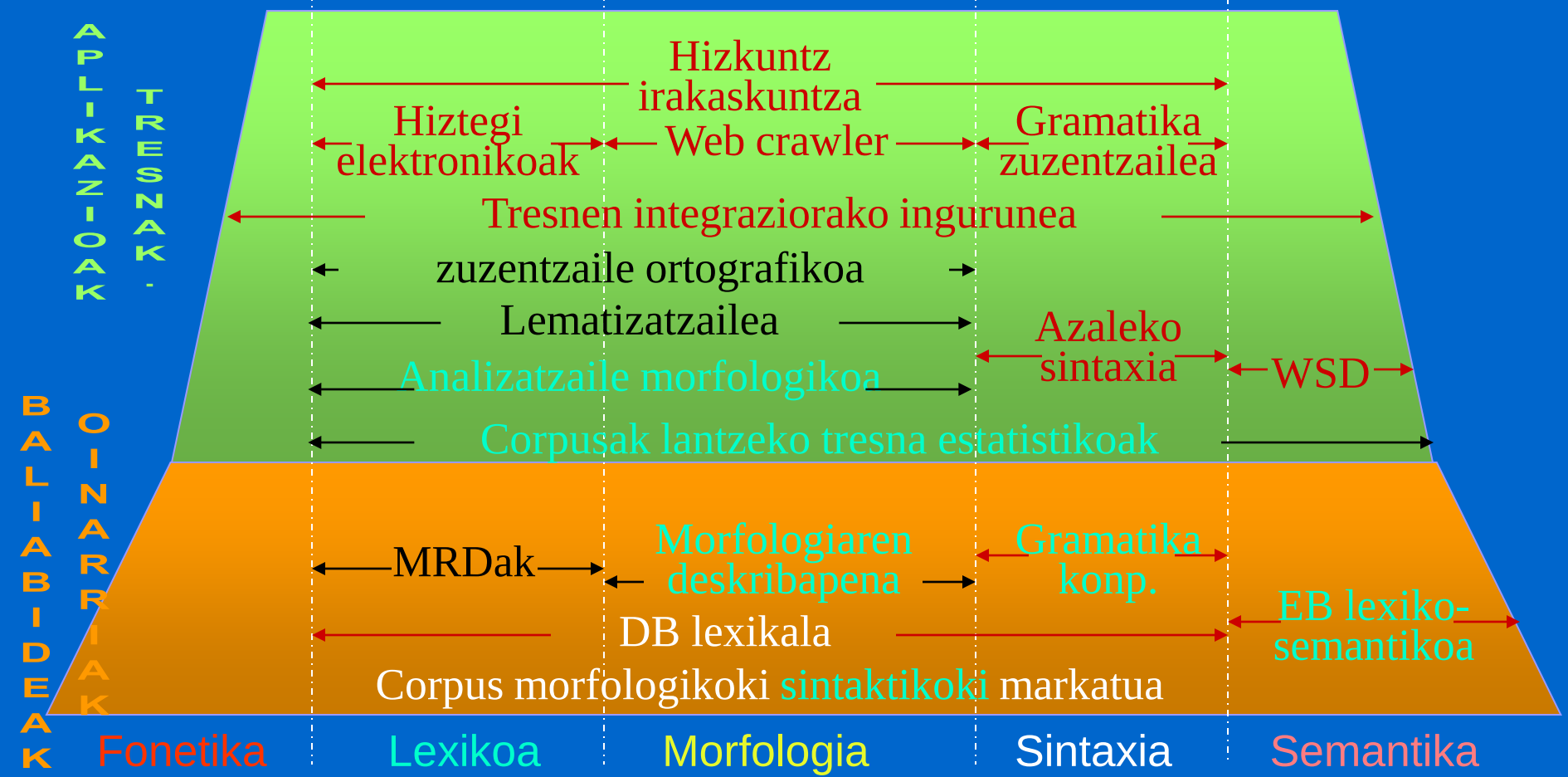
Zuzentzaile ortografikoa XUXEN

- **Zuzentzaile/egiaztatzaile ortografikoa**
  - **UNIX**
  - **Macintosh**
  - **PC**
- **Erabiltzailearen hiztegia**
- **Erabiltzailearen hiztegiko hitzen forma flexionatuak ere onartzen ditu**

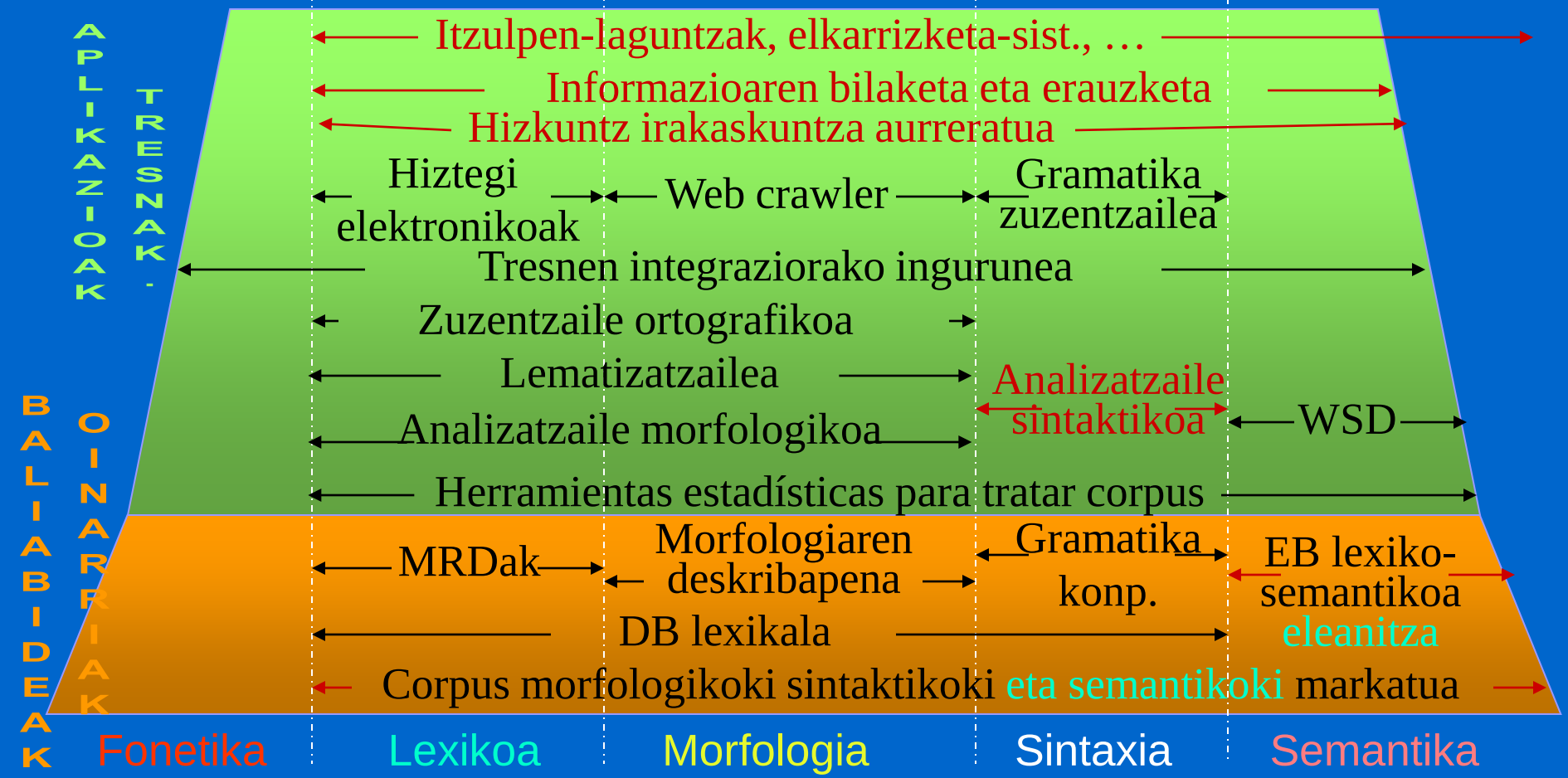
  inplementatu

  ↘ **inplementatuko, inplementatua, inplementatuarekin**

III fasea : Tresna eta aplikazio aurreratuagoak



IV fasea : Eleaniztasuna eta aplikazio orokorrak

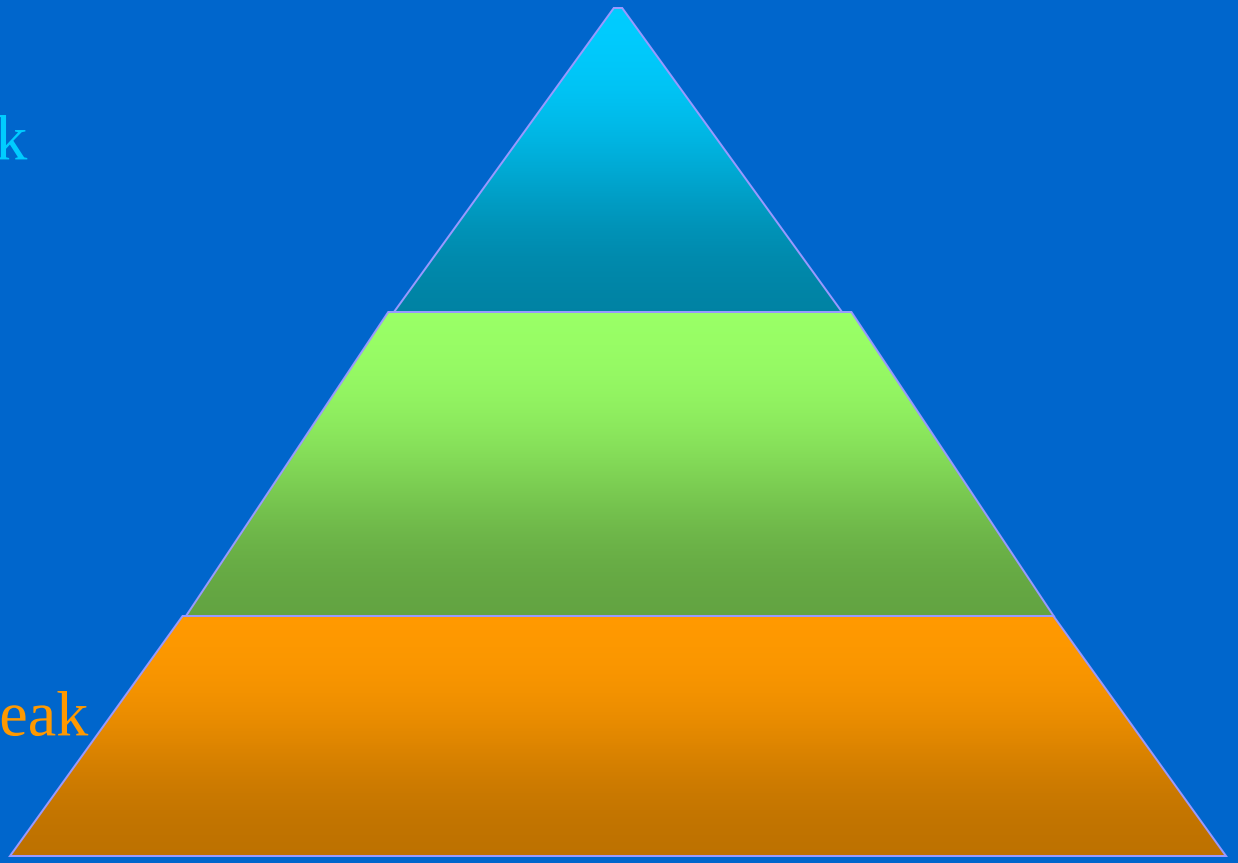


V. fasea: Industria-esplotazioa

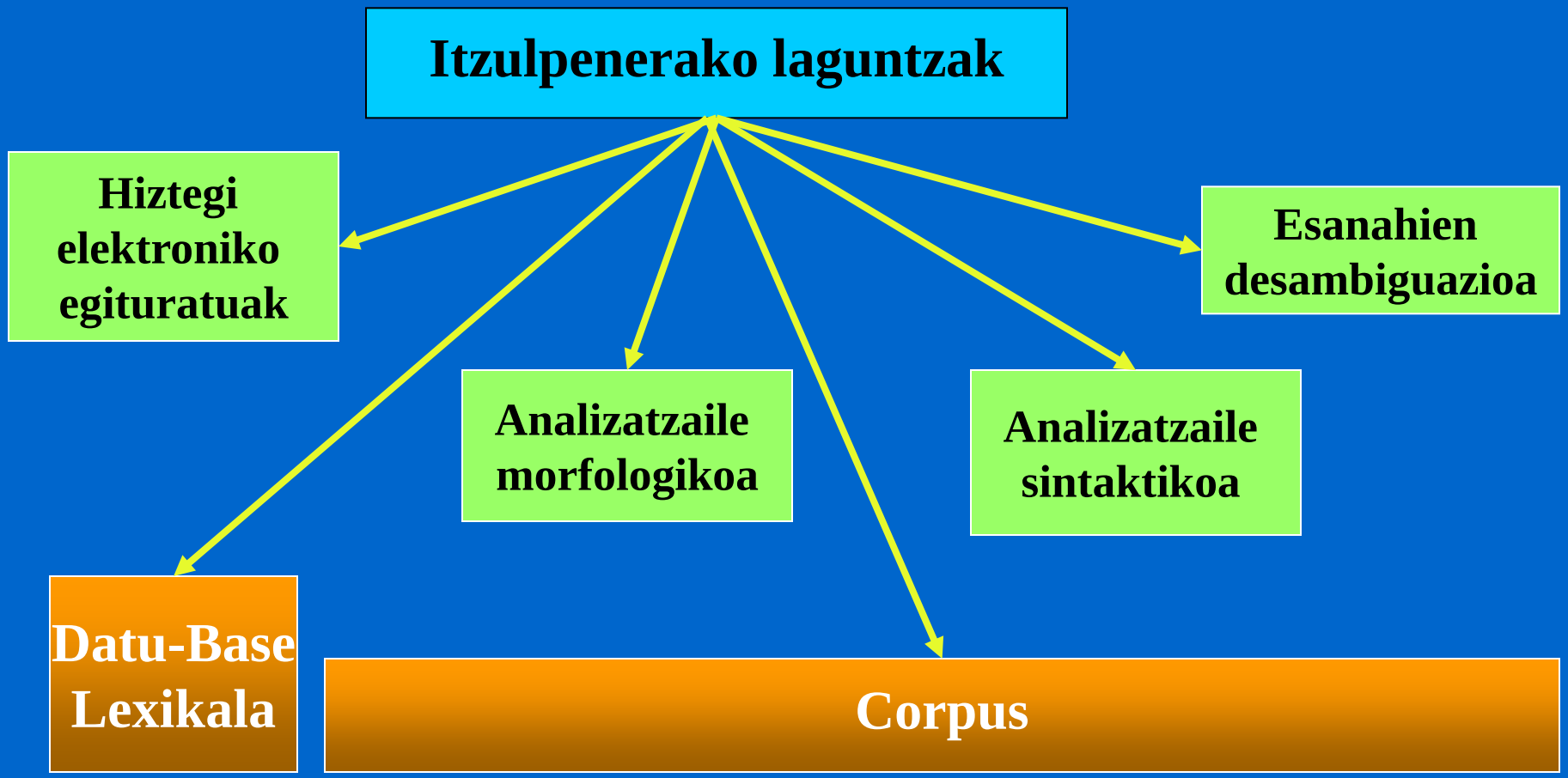
Merkatuko produktuak

Tresnak

Oinarriak eta baliabideak



Berrerabiltzea beharrezkoa da: Adibidea (II)



Zer ez egin?

- **Produktu bat lortzekotan, ez ezkutatu zeuretzat.**
  - **Ikertzaile asko dago ingeleserako ikerketan.**
  - **Baina gutxi hizkuntza txikientzat.**

- **Talde desberdinetan lan bera errepikatzea dirua eta lana xahutzea da.**
  - **Ezinbestekoa lanen koordinazioa,**
  - **Dauden produktuen katalogo integratua behar da. Geure hizkuntzarentzat eta antzeko hizkuntzentzat.**

➔ ***Emaitzak publikoak izan beharko lirateke eta beste ikertzaileek erabiltzeko moduan utzi beharko lirateke.***

Hizkuntzaren Industria antolatzeko urratsak

- Euskararen egoera orain
- Helburuak
- Estrategia
- Eragileak eta bezero posibleak
- Ondorioak

Eragileak
Ingeniaritza Linguistikoa ekinbideerako txostena (Eusko Jaurlaritza, 2000)

- ***ASP, Ametzagaiña AIE, Aurten Bai Fundazioa, BAI & BY, ELHUYAR, EHUko Ahots Taldea, EHUko Ixa Taldea, EHUko Zientzien Fakultatea, Euskaltzaindia, Eusko Ikaskuntza, Eusko Jaurlaritzako Kultura Sailaren Hizkuntza Politikarako Sailordetza, Eusko Jaurlaritzako Hezkuntza, Unibertsitate eta Ikerketa Saila, Eusko Jaurlaritzako Industria, Merkataritza eta Turismo Saila, GEINSA, HABE, Ihardun Multimedia, Interlinea 2000, KAIXO, LKS S. Coop., Telefonica, UZEI eta Zabaltzen.***

+ Enpresa berriak:

VICOMTech, Diana, CodeSyntax, Eleka

Eragileak

hizking21 HIZKUNTZ INGENIARITZA XXI. MENDEAREN ATEAN

- Aholab: EHUko Bilboko Ingeniaritza.
  Hizketaren ezagutza eta sorkuntza
- IXA: EHUko informatika Fakultatea.
  Testuidatzien tratamendua (morfologia, sintaxia, semantika, corpus, itzulpen automatikoa, IE-IR, ...)
- Vicomtech: Ikerketa aplikatuen zentrua (EiTB + Franhaufer) Ordenadore-irudi interaktiboak eta multimedia digital
- Elhuyar Fundazioa:Bitarteko ikerketa-zentrua.
  Lexikografia, terminologia, hiztegiak, hizkuntz planak, zientzia eta teknologiaren zabalkuntza, multimedia-produktu eta zerbitzuak.
- Robotiker: Zentru teknologikoa.
  Informazioaren telekomunikazioaren teknologiak

Eragileak HIZKING21. Helburuak

hizking21 HIZKUNTZ INGENIARITZA XXI. MENDEAREN ATEAN

- **ETORTEK**

  EJ-ko ikerketa lerro estrategikoen deialdiko proiektua

  - **I+G+b (I+D+i)**
    - **Hizkuntza-baliabideen sorkuntza**
    - **Garapen-tresnak**
    - **Teknikak: teknologia eguneroko bizimoduan txertatu ahal izateko**
  - **Formazioa**
  - **Nazioarteko lankidetza**
  - **Zabalkundea**
  - **Behatoki teknologikoa**

Erabiltzaileak: Argitaletxeak

- ***Editorial Desclee de Brouwer S.A., Grupo Delta, Zabaltzen banatzailea, Auñamendi argitaldaria, Editorial Donostiarra, Sendoa, Ostoa S.A., Erein S.A., Lur argitaletxea, Editorial Planeta S.A., Euskal Kulturgintza S.A., Sendoki S.A., Ediciones Saldaña S.A., Aralar liburuak S.A., Alberdania S.L., Donostiako Komunikabideak E.M., Ediciones Txingudi S.L., Euskalgaiak Abarka S.L., Basandere argitaletxea S.L., Udako Euskal Unibertsitatea U.E.U., Miatzen S.A. R.L., Elhuyar Kultur Elkartea, Harlouxet, Susa, Ttarttalo, Elkarlan .***

Erabiltzaileak: *Erakundeak*

- ***EIZIE, HAEE/IVAP, Aldundiak, Eusko Jaurlaritza, Udalak,***
- ***Euskara taldeak: Ikastolen Elkartea, Goiena, Oarso Komunikabideak Fundazioa, Ttipi Ttapa, Topagunea,, Bertsozaleen Elkartea., ...***

Erabiltzaileak: *Komunikabideak*

- ***Egunkaria S.A., Grupo Correo, Gara, Deia, Diario El País, El Mundo, Diario As, Diario Marca, herri-aldizkariak***
- ***Telebistak: EiTB, TVE, A3,T5, Canal+, ...***
- ***Irratiak ...***
- ***...***

Erabiltzaileak:
Bankuak eta aurrezki-kutxak

- ***Kutxa, BBK, Vital Kutxa, Euskadiko Kutxa, Caixa, ...***
- ***BBVA, Banco Guipuzcoano, ...***

Hizkuntzaren Industria antolatzeko urratsak

- Euskararen egoera orain
- Helburuak
- Estrategia
- Eragileak eta bezero posibleak
- Ondorioak

Ondorioak

- **Badira hainbat produktu euskara eta sotwarea uztartzen dituztenak (105 Software-aren Katalogoan)**
- **Horietarik 33 lotuta daude Hizkuntzaren Industriarekin**
- **Hori ez da zeroren hurrengoa baina bai oso gutxi**

⇒ **ahalegin handia egin behar dugu atzean ez gelditzeko**

Ondorioak (2)

- ***Aurkeztu dugu epe erdirako estrategia Ingeniaritza linguistikoan ikerketan eta garapenean lan egiteko***
- ***IXA taldearen 15 urteko eskarmentuan oinarritua***
- ***Oinarri linguistiko bakoitza, tresna eta aplikazio bakoitza ondo diseinatu behar da ondorengo produktuetan erabilgarria izan dadin.***
- ***Nazioartekoan puntako mailan mugituko den industria sendoa sortu dezakegu***
- ***Ikerketa-taldeek, industriak eta erakunde ofizialek koordinatu egin behar dira helburu hori lortzeko***

Hizkuntzaren Industria antolatzeko urratsak

Kepa Sarasola

IXA taldea http://ixa.si.ehu.es

UPV-EHU

Universidad del País Vasco-Euskal Herriko Unibertsitatea