Web-corpusen Ataria

elhuyar
Hizkuntza eta Teknologia

Igor Leturia, Iñaki San Vicente, Iker Manterola, Antton Gurrutxaga

IEB 2013

Miramon, 2013ko maiatzaren 8a

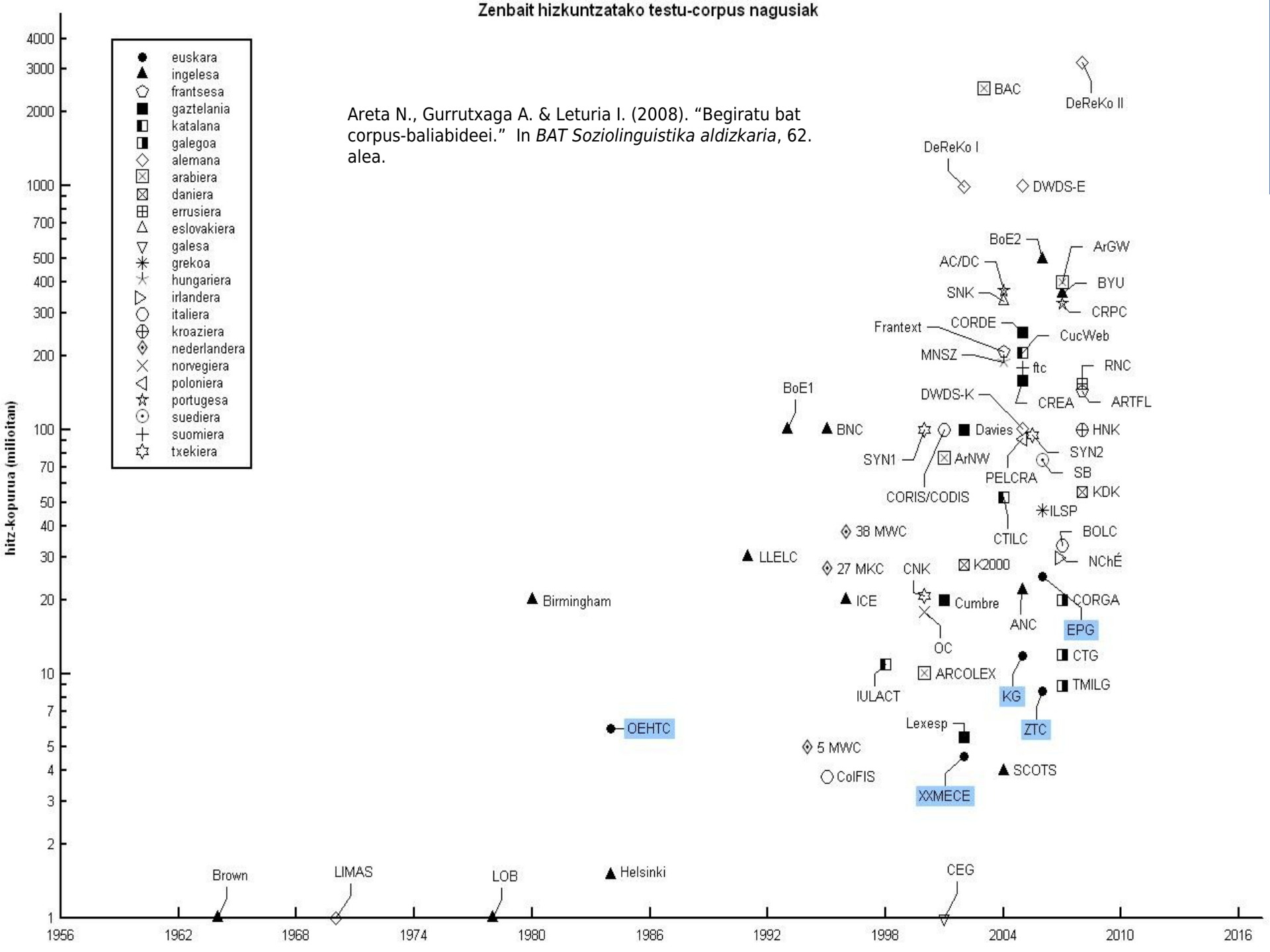
Edukia

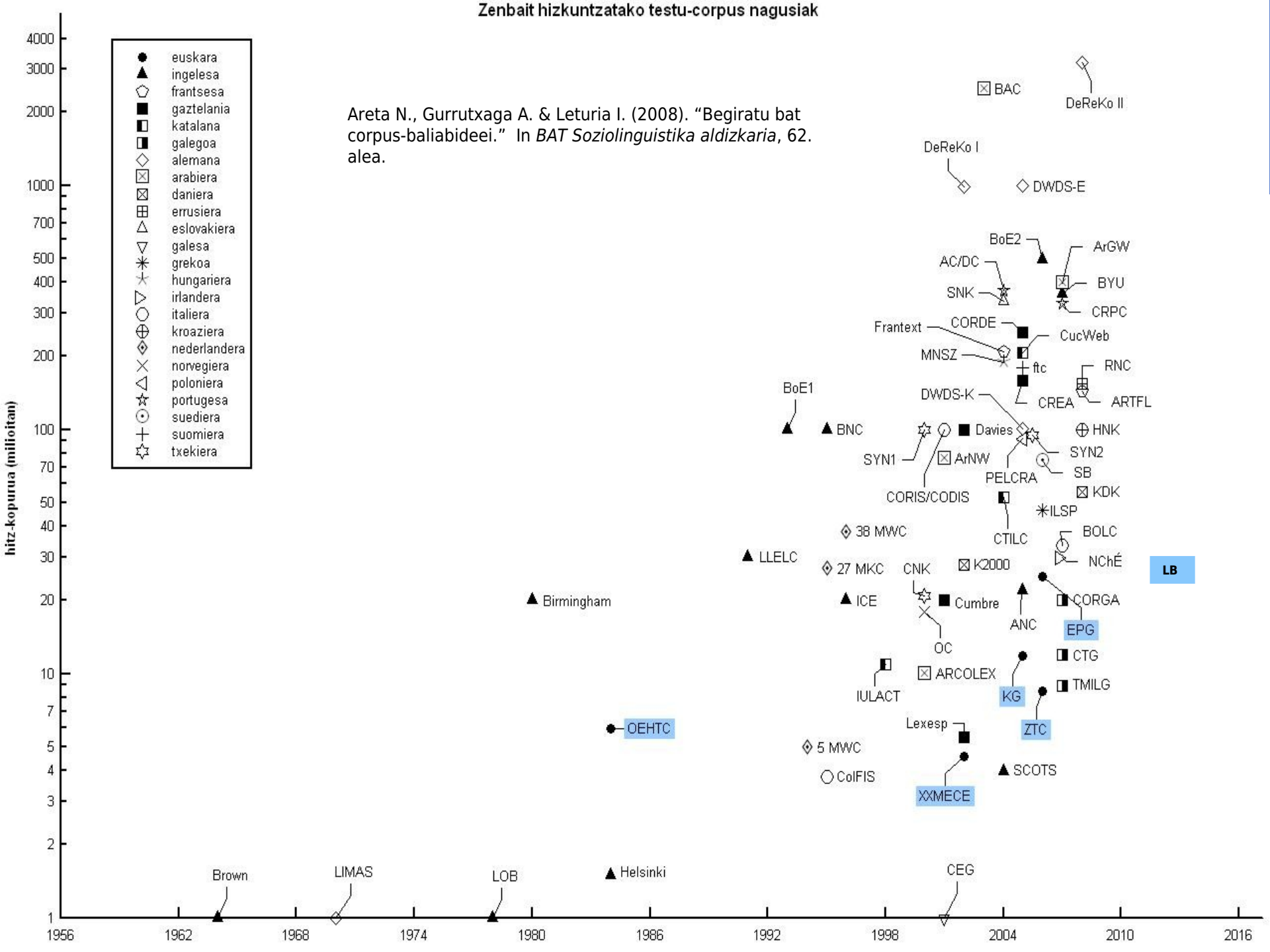
- ZERGATIK egin dugu?
- ZER da?
- ZENBAT du?
- NOLA egin dugu?
- ZERTARAKO balio dezake?
- Hurrengo urratsak

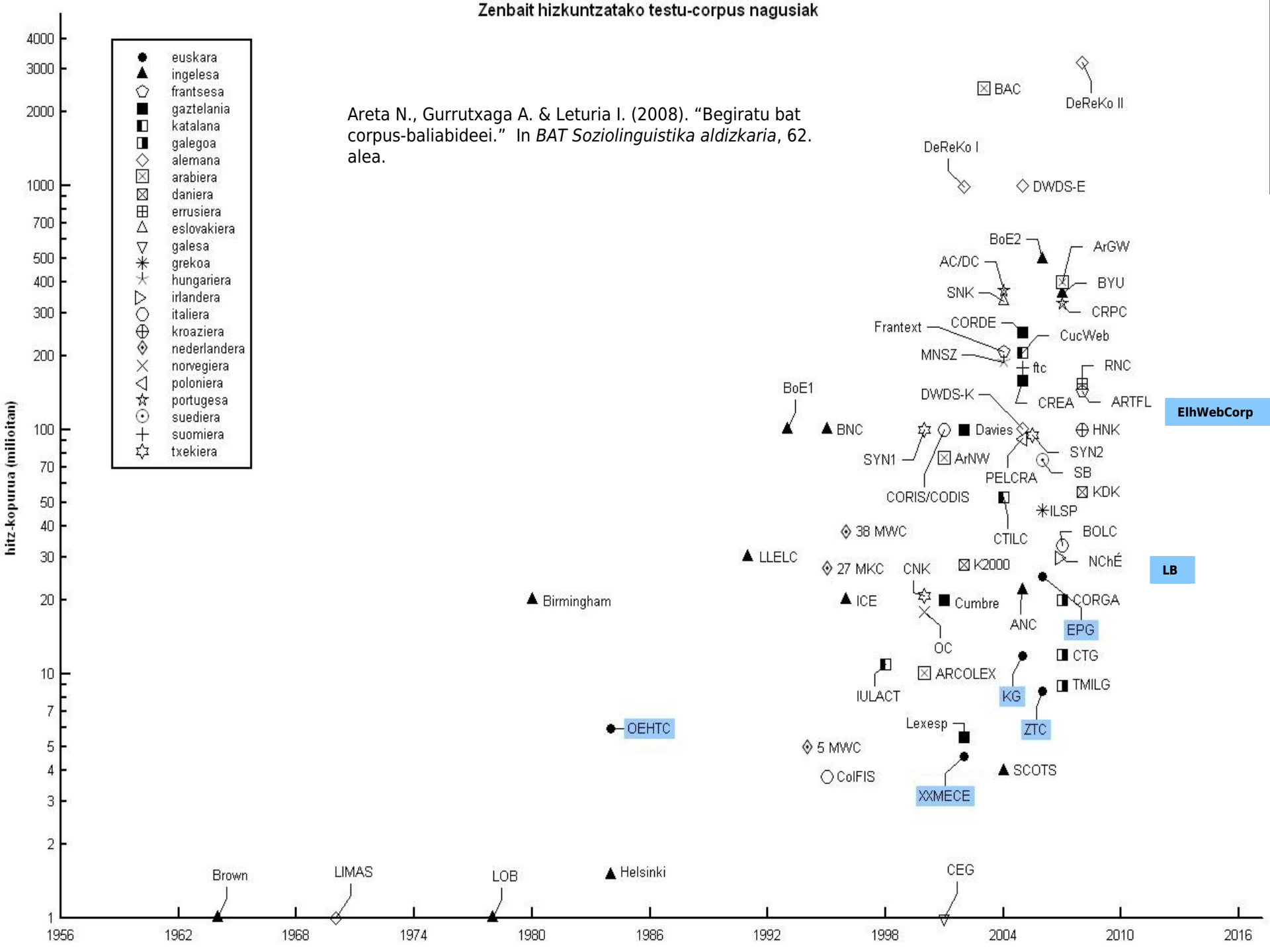
elhuyar
Hizkuntza eta Teknologia

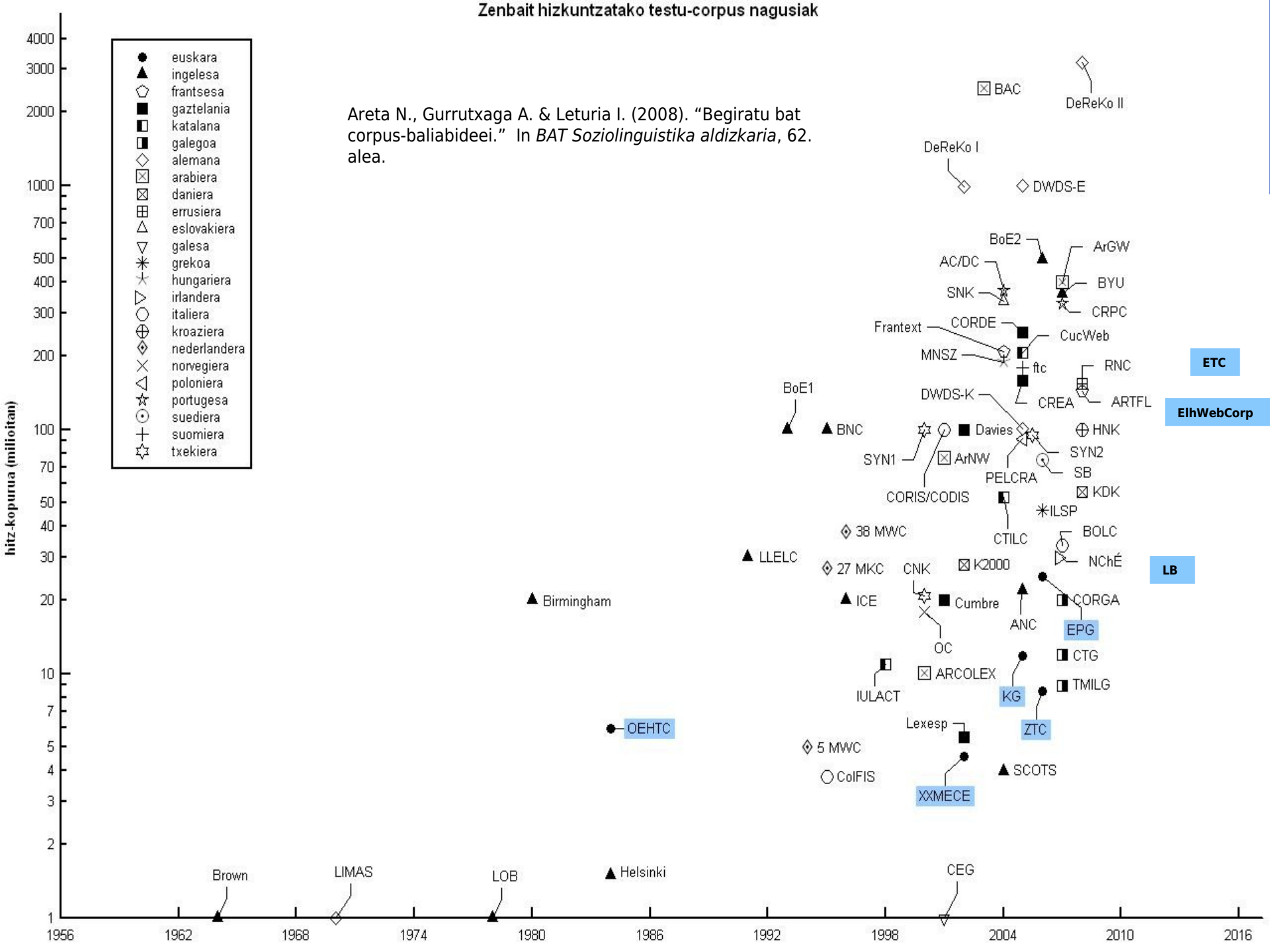
ZERGATIK egin dugu ?

elhuyar
Hizkuntza eta Teknologia


Zenbait hizkuntzatako testu-corpus nagusiak
Areta N., Gurrutxaga A. & Leturia I. (2008). “Begiratu bat corpus-baliabideei.” In *BAT Soziolinguistika aldizkaria*, 62. alea.
euskara
ingelesa
frantsesa
gaztelania
katalana
galegoa
alemana
arabiera
daniera
errusiera
eslovakiera
galesa
grekoa
hungariera
irlandera
italiera
kroaziera
nederlandera
norvegiera
poloniera
portugesa
suediera
suomiera
txekiera
hitz-kopurua (milioitan)
4000
3000
2000
1000
700
500
400
300
200
100
70
50
40
30
20
10
7
5
4
3
2
1
1956
1962
1968
1974
1980
1986
1992
1998
2004
2010
2016
BAC
DeReKo II
DeReKo I
DWDS-E
BoE2
ArGW
AC/DC
BYU
SNK
CRPC
Frantext
CORDE
CucWeb
MNSZ
ftc
RNC
BoE1
DWDS-K
CREA
ARTFL
BNC
Davies
HNK
SYN1
ArNW
SYN2
PELCRA
SB
CORIS/CODIS
KDK
ILSP
38 MWC
CTILC
BOLC
LLELC
27 MKC
CNK
K2000
NChÉ
Birmingham
ICE
Cumbre
CORGA
ANC
EPG
OC
CTG
ARCOLEX
IULACT
KG
TMILG
OEHTC
Lexesp
ZTC
5 MWC
SCOTS
ColFIS
XXMECE
Brown
LIMAS
LOB
Helsinki
CEG


Zenbait hizkuntzatako testu-corpus nagusiak
Areta N., Gurrutxaga A. & Leturia I. (2008). "Begiratu bat corpus-baliabideei." In *BAT Soziolinguistika aldizkaria*, 62. alea.
euskara
ingelesa
frantsesa
gaztelania
katalana
galegoa
alemana
arabiera
daniera
errusiera
eslovakiera
galesa
grekoa
hungariera
irlandera
italiera
kroaziera
nederlandera
norvegiera
poloniera
portugesa
suediera
suomiera
txekiera
hitz-kopurua (milioitan)
4000
3000
2000
1000
700
500
400
300
200
100
70
50
40
30
20
10
7
5
4
3
2
1
1956
1962
1968
1974
1980
1986
1992
1998
2004
2010
2016
BAC
DeReKo II
DeReKo I
DWDS-E
BoE2
ArGW
AC/DC
BYU
SNK
CRPC
Frantext
CORDE
CucWeb
MNSZ
ftc
RNC
BoE1
DWDS-K
CREA
ARTFL
BNC
Davies
HNK
SYN1
ArNW
SYN2
PELCRA
SB
CORIS/CODIS
KDK
ILSP
38 MWC
CTILC
BOLC
LLELC
27 MKC
CNK
K2000
NChÉ
LB
Birmingham
ICE
Cumbre
CORGA
ANC
EPG
OC
CTG
ARCOLEX
IULACT
KG
TMILG
OEHTC
Lexesp
ZTC
5 MWC
SCOTS
ColFIS
XXMECE
Brown
LIMAS
LOB
Helsinki
CEG


Zenbait hizkuntzatako testu-corpus nagusiak
Areta N., Gurrutxaga A. & Leturia I. (2008). "Begiratu bat corpus-baliabideei." In *BAT Soziolinguistika aldizkaria*, 62. alea.
euskara
ingelesa
frantsesa
gaztelania
katalana
galegoa
alemana
arabiera
daniera
errusiera
eslovakiera
galesa
grekoa
hungariera
irlandera
italiera
kroaziera
nederlandera
norvegiera
poloniera
portugesa
suediera
suomiera
txekiera
hitz-kopurua (milioitan)
4000
3000
2000
1000
700
500
400
300
200
100
70
50
40
30
20
10
7
5
4
3
2
1
1956
1962
1968
1974
1980
1986
1992
1998
2004
2010
2016
BAC
DeReKo II
DeReKo I
DWDS-E
BoE2
ArGW
AC/DC
BYU
SNK
CRPC
Frantext
CORDE
CucWeb
MNSZ
ftc
RNC
BoE1
DWDS-K
CREA
ARTFL
ElhWebCorp
BNC
Davies
HNK
SYN1
ArNW
SYN2
PELCRA
SB
CORIS/CODIS
KDK
ILSP
38 MWC
CTILC
BOLC
LLELC
27 MKC
CNK
K2000
NChÉ
LB
Birmingham
ICE
Cumbre
CORGA
ANC
EPG
OC
CTG
ARCOLEX
IULACT
KG
TMILG
ZTC
OEHTC
Lexesp
5 MWC
SCOTS
ColFIS
XXMECE
Brown
LIMAS
LOB
Helsinki
CEG


Zenbait hizkuntzatako testu-corpus nagusiak
Areta N., Gurrutxaga A. & Leturia I. (2008). "Begiratu bat corpus-baliabideei." In *BAT Soziolinguistika aldizkaria*, 62. alea.
euskara
ingelesa
frantsesa
gaztelania
katalana
galegoa
alemana
arabiera
daniera
errusiera
eslovakiera
galesa
grekoa
hungariera
irlandera
italiera
kroaziera
nederlandera
norvegiera
poloniera
portugesa
suediera
suomiera
txekiera
hitz-kopurua (milioitan)
4000
3000
2000
1000
700
500
400
300
200
100
70
50
40
30
20
10
7
5
4
3
2
1
1956
1962
1968
1974
1980
1986
1992
1998
2004
2010
2016
BAC
DeReKo II
DeReKo I
DWDS-E
BoE2
ArGW
AC/DC
BYU
SNK
CRPC
Frantext
CORDE
CucWeb
MNSZ
ftc
RNC
ETC
BoE1
DWDS-K
CREA
ARTFL
ElhWebCorp
BNC
Davies
HNK
SYN1
ArNW
SYN2
PELCRA
SB
CORIS/CODIS
KDK
ILSP
38 MWC
CTILC
BOLC
LLELC
27 MKC
CNK
K2000
NChÉ
LB
Birmingham
ICE
Cumbre
CORGA
ANC
EPG
OC
CTG
ARCOLEX
IULACT
KG
TMILG
ZTC
OEHTC
Lexesp
5 MWC
SCOTS
ColFIS
XXMECE
Brown
LIMAS
LOB
Helsinki
CEG

ZERGATIK egin dugu?

elhuyar Hizkuntza eta Teknologia

- Ikuspegi “praktikoaz” gain...
  - Ikuspegi linguistikoa
    - Internet ezin ukatuzko errealitate "linguistikoa" ere bada
    - Interneten BAKARRIK argitaratzen diren testuak gero eta ugariagoak dira, eta ezaugarri bereziak dituzte
    - Interneteko testu asko ez dira "diskurtso kontrolatuak", bat-batekoak baizik; hizkuntzaren erabilera erreala islatzen dute
- Interesgarria da webaren alderdi linguistiko bereziak aztertzea

ZER da?

elhuyar
Hizkuntza eta Teknologia

- Zer da Web-corpusen Ataria?
  - Webetik automatikoki osatutako euskarazko corpusen kontsulta aurreratua
  - Elhuyar Fundazioko hizkuntza-teknologien I+G taldeak egindako lanaren emaitza
  - Hiru atal:
    - Euskarazko web corpus elebakarra
    - Euskara-Gaztelania web corpus paraleloa
    - Hitz-konbinazioen kontsulta

Web-corpusen Ataria
Corpus elebakarra

Babeslea: EUSKO JAURLARITZA GOBIERNO VASCO

ELHUYAR fundazioa

Hasiera | Corpus elebakarra | Corpus paraleloa | Hitz-konbinazioak | Laguntza | Eranskina

Galdera

Zer	Aukerak	Bilatu	Kategoria
Lema	Da	xedapen	

Dist.	Non	Zer	Aukerak	Bilatu	Kategoria

Emaitza: Testuinguruak eta kopuruak

Ordenatu honen arabera: Dokumentua

Kopuruak: Forma, Lema, Kategoria, Lema eta kategoria

Gehienez: 10 %

Bilaketa arrunta — Bilatu — Garbitu

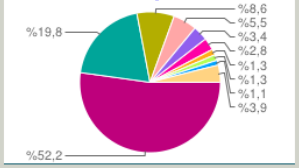
Emaitzak: 13486

Kopuruak

Forma

Forma	Kop
xedapen	7041
xedapenAK	2668
xedapenA	1158
xedapenen	736
xedapenetan	463
xedapenek	380
xedapenei	181
xedapenean	176
xedapenetatik	151
Beste guztiak	532
Guztira	13486

Guztien testuinguruak batera



Denak (13486)

https://www6.euskadi.net/u81-0003/eu/contenidos/autorizacion/2216/eu_4901/eu_17719.html (1)

...datzen ere prozedura administratiboak erregulatzen dituzten xedapen arautzaileak, eta, ondorioz, ezin ditu sortu berariazko Ara...

http://www.kutxa.net/wkn_corporativo/2009/informe_financiero/eu/memoria_consolidada_02.htm (1)

...kal, Administratibo eta Ordena Sozialari buruzkoaren, azken Xedapenetako hamaikagarrenak ezartzen duenaren arabera.

http://www.euskadi.net/cgi-bin_k54/ver_e?CMDVERDOC&BASEB03D&DOCN000038101&CONF/config/k54/bopv_e.cnf (2)

...3. artikuluetan eta artikulu horiekin erlazionatutako beste xedapenetan ezarritakoarekin bat etorriz.
AZKEN XEDAPENAK

https://ssl4.gipuzkoa.net/euskera/gao/2004/02/16/e0401200.htm (4)

XEDAPEN GEHIGARRIAK
XEDAPEN IRAGANKORRA
AZKEN XEDAPENAK
I. KAPITULUA. BABES-ARAUBIDEAREN XEDAPEN OROKORRAK

http://www.cgt-lkn.org/laescoba/convenio/Mungia.pdf (3)

...a jatorria edozein izanik ere. Aplikazio orokorreko legezko xedapenak, edo Hitzarmen Kolektiboak, edo banakako kontratuak direla ...
XEDAPENAK
40. artikulua.--Azken xedapenak Hitzarmen hau indarrean dagoen bitartean, bertan itundutako...

http://www.senado.es/brsweb/CALEX/textos/paisvasco/44/2003/03.pdf (7)

Bestelako Xedapenak
Xedapen Orokorrak
Lege honek hogeita bat artikulu ditu, lau xedapen iragankor eta azken xedapen bat. Legeak zehaztu egiten ditu...
... hogeita bat artikulu ditu, lau xedapen iragankor eta azken xedapen bat. Legeak zehaztu egiten ditu aurrezki-kutxen eremuan jar...
Aurrezki-kutxek lege honetako xedapenetara egokitu beharko dituzte beren estatutuak eta araudiak, eta ...
XEDAPEN
XEDAPENA

http://www.innobasque.com/Modulos/DocumentosVisor.aspx?docId233 (3)

Xedapen Orokorrak
...azaroaren 26ko 30/1992 Legean zehaztutakoa edo bere ordezko xedapenak ezartzen duena izango da ezargarri. Hirugarrena.Â Aurrekon...
...zarritako mugen barruan. Aipatutako Lege Organiko horren 5. xedapen gehigarrian ezarritakoaren arabera, Unibertsitatea Koordina...

http://www.donostia.org/ordenanzas.nsf/dbda625a50d6b4b8412566800044e4ed/b8970ea57f9ec357c12574ea002a49bc?OpenDocument (2)

...laren 5eko Foru Arauean ezarritako prezio publikoei buruzko xedapenekin bat, Donostiako Gazteriaren Aterpetxe eta Kanpalekuen Udal ...
10. art.- Azken xedapena

http://www.euskadi.net/bopv2/datos/1993/09/9302983a.pdf (1)

Bestelako Xedapenak

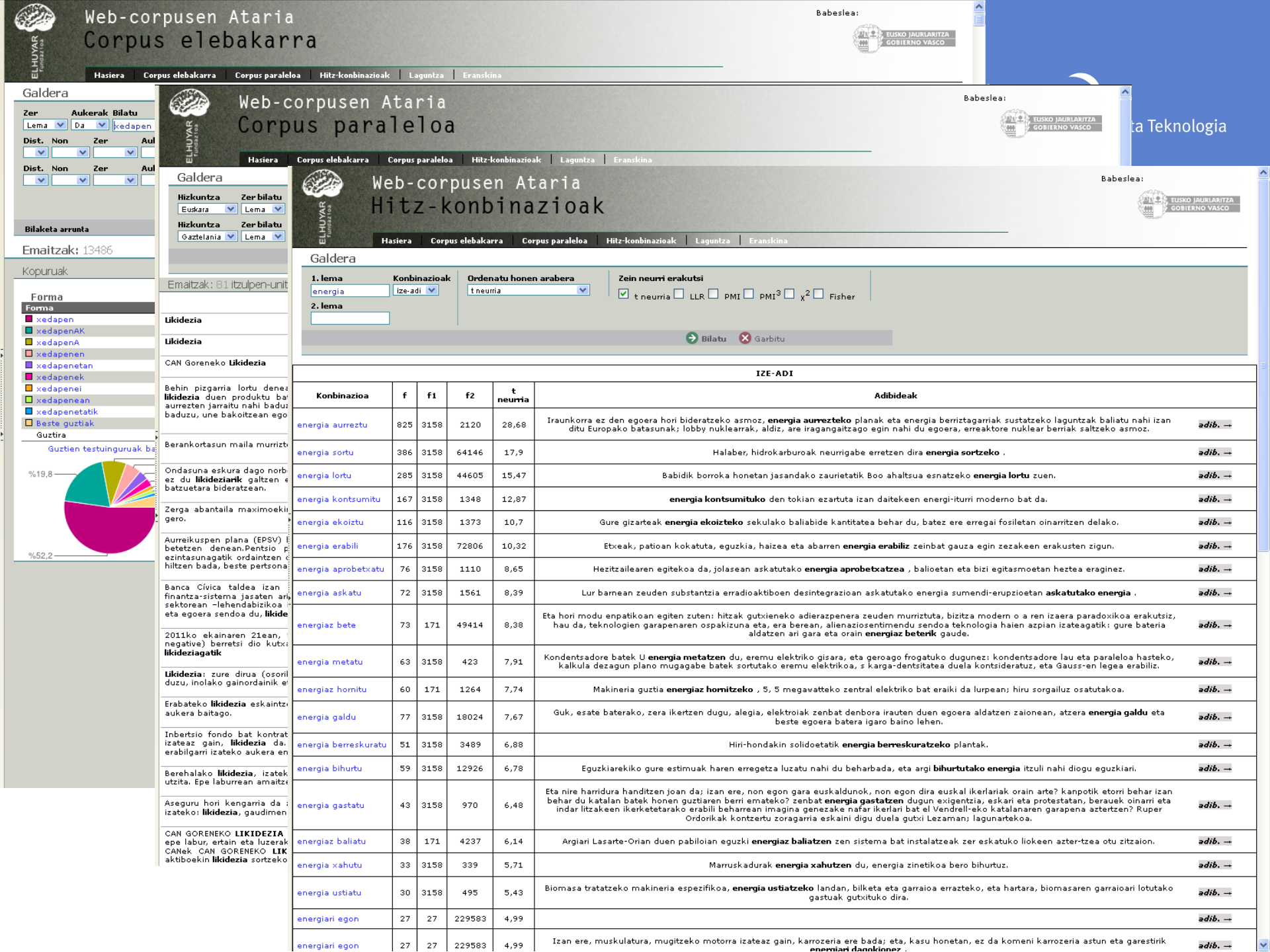
http://www.gipuzkoa.net/euskera/gao/2004/12/03/e0411069.htm (5)

XEDAPEN OROKORRAK
Xedapen Gehigarria.
Indargabetzeko Xedapena.
Azkenaurreko Xedapena.
Azken Xedapena.

http://www.euskara.euskadi.net/r59-euhadm2/eu/contenidos/informacion/ohiko_galderak/eu_irati/dekretuari_buruzkoak.html (2)

...tik kentzeko espedientea has daitekeela ondoriozta daiteke. Xedapen honen oinarria 6/1989 Legean dago, 50. artikuluan hain zuze...


Web-corpusen Ataria
Corpus elebakarra
Babeslea:
EUSKO JAURLARITZA
GOBIERNO VASCO
Hasiera | Corpus elebakarra | Corpus paraleloa | Hitz-konbinazioak | Laguntza | Eranskina
Galdera
Zer Aukerak Bilatu
Lema Da xedapen
Dist. Non Zer Aul
Bilaketa arrunta
Emaitzak: 13486
Kopuruak
Forma
Forma
xedapen
xedapenAK
xedapenA
xedapenen
xedapenetan
xedapenek
xedapenei
xedapenean
xedapenetatik
Beste guztiak
Guztira
Guztien testuinguruak ba
%19,8
%52,2
Web-corpusen Ataria
Corpus paraleloa
Babeslea:
EUSKO JAURLARITZA
GOBIERNO VASCO
Hasiera | Corpus elebakarra | Corpus paraleloa | Hitz-konbinazioak | Laguntza | Eranskina
Galdera
Hizkuntza Zer bilatu Aukerak Hitza Kategoria
Euskara Lema Da likidezia
Hizkuntza Zer bilatu Aukerak Hitza Kategoria
Gaztelania Lema Da
Ordenatu honen arabera
Maiztasuna
Bilatu
Garbitu
Emaitzak: 81 itzulpen-unitate, 10 domeinu
Euskara
www.tesoro.es (75)
Gaztelera
www.tesoro.es (75)
Likidezia
www.cajanavarra.es (27)
iturriak
Liquidez
www.cajanavarra.es (27)
fuentes
Likidezia
www.cajanavarra.es (26)
iturriak
Liquidez
www.cajanavarra.es (26)
fuentes
CAN Goreneko Likidezia
www.cajanavarra.es (14)
iturriak
CAN Máxima Liquidez
www.cajanavarra.es (14)
fuentes
Behin pizgarria lortu denean, aurrezki plana egiten jarraitzea komeni da. Lehen unetik likidezia duen produktu bat izanik, ahalik eta luzeena den epea ezartzea komeni zaizu, aurrezten jarraitu nahi baduzu edo zure aurrezkia eskura izateko unerik onena aukeratu nahi baduzu, une bakoitzean egon daitezkeen zerga abantailei onura atera baitiezaiekezu.
www.cajanavarra.es (11)
iturriak
Una vez conseguido el incentivo es recomendable seguir con el plan de ahorro, ya que al tratarse de un producto con liquidez desde el primer momento, conviene establecer un plazo lo más largo posible en caso de que quieras seguir ahorrando o que quieras elegir el momento óptimo para disponer de tu ahorro, pudiéndote beneficiar de las posibles ventajas fiscales del momento.
www.cajanavarra.es (11)
fuentes
Berankortasun maila murrizten da eta likidezia, gaudimen eta diruzaintza ratioak hobetzen.
www.cajanavarra.es (9)
iturriak
Se rebaja el nivel de morosidad, las pérdidas y se mejoran los ratios de liquidez, solvencia y tesorería.
www.cajanavarra.es (9)
fuentes
Ondasuna eskura dago norberaren altxortegia erabili behar izan gabe. Hori dela eta, enpresak ez du likidezia rik galtzen eta errentagarritasun bat lortzen du, altxortegia beste inbertsio batzuetara bideratzean.
www.cajanavarra.es (9)
iturriak
Al disponer del bien sin necesidad de utilizar tesorería propia, la empresa no pierde liquidez y obtiene una rentabilidad al destinarlo a otro tipo de inversiones.
www.cajanavarra.es (9)
fuentes
Zerga abantaila maximoekin, eta berehalako likidezia, lehen ekarpenetik 10 urte igaro eta gero.
www.cajanavarra.es (9)
iturriak
Con las máximas ventajas fiscales y de liquidez inmediata transcurridos 10 años desde la primera aportación.
www.cajanavarra.es (9)
fuentes
Aurreikuspen plana (EPSV) bada, gainera, berehalako likidezia du, 10 urteko antzinatasuna betetzen denean.Pentsio planek aparteko kapitala sortzen dute -erretiratzeagatik edo ezintasunagatik ordaintzen duten pentsioa osatzen duen kapitala, hain zuzen- eta, partaidea hiltzen bada, beste pertsona batzuei laguntzen diete.
www.cajanavarra.es (8)
iturriak
Estamos ante un producto financiero que, además de garantizar ingresos extra a la hora de la jubilación, supone una ayuda eficaz para ahorrar en la declaración de la renta. Si lo que se tiene es un plan de previsión (EPSV), además cuenta con una liquidez inmediata a los 10 años de antigüedad del mismo.
www.cajanavarra.es (8)
fuentes
Banca Cívica taldea izan zen estreinako erregio arteko integratze-prozesua Espainiako finantza-sistema jasaten ari den sendotze-prozesuaren barruan. Eragiketa aurreratuena da sektorean -lehendabizikoa izan zen banku propio operatiboa sortzen, 2010eko ekainetik-, eta egoera sendoa du, likidezian eta kaudimenari dagokienez, baita balantze saneatua ere.
www.kutxa.net (8)
iturriak
La creación del grupo Banca Cívica fue el primer proceso de integración interregional dentro del proceso de consolidación sectorial que vive el sistema financiero español y es la operación más avanzada del sector -fue el primero en crear un banco propio operativo desde junio de 2010- con una sólida posición en liquidez y solvencia y un balance saneado.
www.kutxa.net (8)
fuentes
2011ko ekainaren 21ean, Standard & Poor's rating agentziak A/A-1 kalifikazioa (outlook negative) berretsi dio kutxari, berriro ere entitatearen finantza-sendotasun, kaudimen eta likidezia gatik
www.seguroslagunaro.com (8)
iturriak
El 21 de junio de 2011 la agencia de rating Standard & Poor's ha ratificado a kutxa la calificación A/A-1, outlook negative, basada, una vez más, en la fortaleza financiera, la solvencia y la liquidez de la Entidad.
www.seguroslagunaro.com (8)
fuentes
Likidezia: zure dirua (osorik edo haren zati bat) noiznahi berreskuratzeko aukera izango duzu, inolako gainordainik eta komisio gehigarririk GABE.
www.seguroslagunaro.com (8)
iturriak
Liquidez: en cualquier momento podrás disponer de tu ahorro, total o parcialmente, SIN ningún recargo ni comisión adicional.
www.seguroslagunaro.com (8)
fuentes
Erabateko likidezia eskaintzen du, aurreztutako zenbatekoa edozein unetan berreskuratzeko aukera baitago.
www.cajanavarra.es (7)
iturriak
Ofrece plena liquidez, ya que el ahorro constituido está disponible en cualquier momento.
www.cajanavarra.es (7)
fuentes
Inbertsio fondo bat kontratatzeak dituen alde on nagusietako bat, fiskalitate erakargarria izateaz gain, likidezia da. Izan ere, kasu gehienetan, inbertitutako kopurua berehala erabilgarri izateko aukera ematen baitu.
www.cajanavarra.es (7)
iturriak
Una de las principales ventajas de contratar un fondo de inversión, además de una atractiva fiscalidad, es su liquidez, ya que permite disponer, en la mayor parte de los casos, de la cantidad invertida de forma prácticamente inmediata.
www.cajanavarra.es (7)
fuentes
Berehalako likidezia, izateko bidea ematen du, aipatutako dokumentuak endosatuta edo utzita. Epe laburrean amaitzen da, eskuarki, 90 egun igaro baino lehenago.
www.cajanavarra.es (7)
iturriak
Permite liquidez inmediata, mediante el endoso o cesión de los documentos citados, con vencimiento a corto plazo; habitualmente inferior a 90 días.
www.cajanavarra.es (7)
fuentes
Aseguru hori kengarria da zergei begira eta aukera ematen du banku finantzaketa hobea izateko: likidezia, gaudimen eta diruzaintza ratioak hobetzen ditu.
www.cajanavarra.es (5)
iturriak
Este seguro es deducible fiscalmente y permite una financiación bancaria más favorable: mejores ratios de liquidez, solvencia y tesorería.
www.cajanavarra.es (5)
fuentes
CAN GORENEKO LIKIDEZIA CENek azaldu du finantzatze lerro bat izatea interesatzen zaiola, epe labur, ertain eta luzerako finantza beharrak konpontzea ahalbidetzen duena. Horregatik, CANek CAN GORENEKO LIKIDEZIA finantzatze lerro berria sortu du. Gainera, enpresaren aktiboekin likidezia sortzeko aukera ematen du lerro horrek.
iturriak
CAN MAXIMA LIQUIDEZ: en base al interés manifestado por CEN en la disposición de una línea de financiación que posibilite resolver tanto las necesidades financieras a corto, medio y largo plazo, CAN crea la línea de financiación CAN MAXIMA LIQUIDEZ que además permite generar liquidez con los activos de la empresa.
fuentes


Web-corpusen Ataria
Corpus elebakarra
Babeslea:
EUSKO JAURLARITZA GOBIERNO VASCO
Hasiera | Corpus elebakarra | Corpus paraleloa | Hitz-konbinazioak | Laguntza | Eranskina
Galdera
Zer Aukerak Bilatu
Lema Da xedapen
Dist. Non Zer Aul
Bilaketa arrunta
Emaitzak: 13486
Kopuruak
Forma
Forma
xedapen
xedapenAK
xedapenA
xedapenen
xedapenetan
xedapenek
xedapenei
xedapenean
xedapenetatik
Beste guztiak
Guztira
Guztien testuinguruak ba
%19,8
%52,2
Web-corpusen Ataria
Corpus paraleloa
Babeslea:
EUSKO JAURLARITZA GOBIERNO VASCO
Hasiera | Corpus elebakarra | Corpus paraleloa | Hitz-konbinazioak | Laguntza | Eranskina
Galdera
Hizkuntza Zer bilatu
Euskara Lema
Hizkuntza Zer bilatu
Gaztelania Lema
Emaitzak: 81 itzulpen-unit
Likidezia
Likidezia
CAN Goreneko Likidezia
Behin pizgarria lortu denez likidezia duen produktu ba aurrezten jarraitu nahi badu baduzu, une bakoitzean ego
Berankortasun maila murrizt
Ondasuna eskura dago norb ez du likideziarik galtzen batzuetara bideratzean.
Zerga abantaila maximoeki gero.
Aurreikuspen plana (EPSV) betetzen denean.Pentsio ezintasunagatik ordaintzen hiltzen bada, beste pertsona
Banca Cívica taldea izan finantza-sistema jasaten ari sektorean –lehendabizikoa eta egoera sendoa du, likide
2011ko ekainaren 21ean, negatiboa) berretsi dio kutx likideziagatik
Likidezia: zure dirua (osoril duzu, inolako gainordainik e
Erabateko likidezia eskaintz aukera baitago.
Inbertsio fondo bat kontrat izateaz gain, likidezia da. erabilgarri izateko aukera en
Berehalako likidezia, izatek utzita. Epe laburrean amaitz
Aseguru hori kengarria da izateko: likidezia, gaudimen
CAN GORENEKO LIKIDEZIA epe labur, ertain eta luzerak CANek CAN GORENEKO LIK aktiboekin likidezia sortzeko
Web-corpusen Ataria
Hitz-konbinazioak
Babeslea:
EUSKO JAURLARITZA GOBIERNO VASCO
Hasiera | Corpus elebakarra | Corpus paraleloa | Hitz-konbinazioak | Laguntza | Eranskina
Galdera
1. lema
energia
Konbinazioak
ize-adi
2. lema
Ordenatu honen arabera
t neurria
Zein neurri erakutsi
t neurria LLR PMI PMI3 x2 Fisher
Bilatu Garbitu
IZE-ADI
Konbinazioa f f1 f2 t neurria Adibideak
energia aurreztu 825 3158 2120 28,68 Iraunkorra ez den egoera hori bideratzeko asmoz, energia aurrezteko planak eta energia berriztagarriak sustatzeko laguntzak baliatu nahi izan ditu Europako batasunak; lobby nuklearrak, aldiz, are iragangaitzago egin nahi du egoera, erreaktore nuklear berriak saltzeko asmoz. adib.
energia sortu 386 3158 64146 17,9 Halaber, hidrokarburoak neurrigabe erretzen dira energia sortzeko . adib.
energia lortu 285 3158 44605 15,47 Babidik borroka honetan jasandako zaurietatik Boo ahaltsua esnatzeko energia lortu zuen. adib.
energia kontsumitu 167 3158 1348 12,87 energia kontsumituko den tokian ezartuta izan daitekeen energi-iturri moderno bat da. adib.
energia ekoiztu 116 3158 1373 10,7 Gure gizarteak energia ekoizteko sekulako baliabide kantitatea behar du, batez ere erregai fosiletan oinarritzen delako. adib.
energia erabili 176 3158 72806 10,32 Etxeak, patioan kokatuta, eguzkia, haizea eta abarren energia erabiliz zeinbat gauza egin zezakeen erakusten zigun. adib.
energia aprobetxatu 76 3158 1110 8,65 Hezitzaileareen egitekoa da, jolasean askatutako energia aprobetxatzea , balioetan eta bizi egitasmoetan heztea eraginez. adib.
energia askatu 72 3158 1561 8,39 Lur barnean zeuden substantzia erradioaktiboen desintegrazioan askatutako energia sumendi-erupzioetan askatutako energia . adib.
energiaz bete 73 171 49414 8,38 Eta hori modu enpatikoan egiten zuten: hitzak gutxieneko adierazpenera zeuden murriztuta, bizitza modern o a ren izaera paradoxikoa erakutsiz, hau da, teknologien garapenaren ospakizuna eta, era berean, alienaziosentimendu sendoa teknologia haien azpian izateagatik: gure bateria aldatzen ari gara eta orain energiaz beterik gaude. adib.
energia metatu 63 3158 423 7,91 Kondentsadore batek U energia metatzen du, eremu elektriko gisara, eta geroago frogatuko dugunez: kondentsadore lau eta paraleloa hasteko, kalkula dezagun plano mugagabe batek sortutako eremu elektrikoa, s karga-dentsitatea duela kontsideratuz, eta Gauss-en legea erabiliz. adib.
energiaz hornitu 60 171 1264 7,74 Makineria guztia energiaz hornitzeko , 5, 5 megawatteko zentral elektriko bat eraiki da lurpean; hiru sorgailuz osatutakoa. adib.
energia galdu 77 3158 18024 7,67 Guk, esate baterako, zera ikertzen dugu, alegia, elektroiak zenbat denbora irauten duen egoera aldatzen zaionean, atzera energia galdu eta beste egoera batera igaro baino lehen. adib.
energia berreskuratu 51 3158 3489 6,88 Hiri-hondakin solidoetatik energia berreskuratzeko plantak. adib.
energia bihurtu 59 3158 12926 6,78 Eguzkiarekiko gure estimuak haren erregetza luzatu nahi du beharbada, eta argi bihurtutako energia itzuli nahi diogu eguzkiari. adib.
energia gastatu 43 3158 970 6,48 Eta nire harridura handitzen joan da; izan ere, non egon gara euskaldunok, non egon dira euskal ikerlariak orain arte? kanpotik etorri behar izan behar du katalan batek honen guztiaren berri emateko? zenbat energia gastatzen dugun exigentzia, eskari eta protestatan, berauek oinarri eta indar litzakeen ikerketetarako erabili beharrean imagina genezake nafar ikerlari bat el Vendrell-eko katalanaren garapena aztertzen? Ruper Ordorikak kontzertu zoragarria eskaini digu duela gutxi Lezaman; lagunartekoa. adib.
energiaz baliatu 38 171 4237 6,14 Argiari Lasarte-Orian duen pabiloian eguzki energiaz baliatzen zen sistema bat instalatzeak zer eskatuko liokeen azter-tzea otu zitzaion. adib.
energia xahutu 33 3158 339 5,71 Marruskadurak energia xahutzen du, energia zinetikoa bero bihurtuz. adib.
energia ustiatu 30 3158 495 5,43 Biomasa tratatzeko makineria espezifikoa, energia ustiatzeko landan, bilketa eta garraioa errazteko, eta hartara, biomasaren garraioari lotutako gastuak gutxituko dira. adib.
energiari egon 27 27 229583 4,99 adib.
energiari egon 27 27 229583 4,99 Izan ere, muskulatura, mugitzeko motorra izateaz gain, karrozeria ere bada; eta, kasu honetan, ez da komeni karrozeria astun eta garestirik energiari dagokionez . adib.

ZENBAT du?

elhuyar Hizkuntza eta Teknologia

- Zenbat du euskarazko web-corpus elebakarrak?
  - Hitzak: 125 milioi inguru
  - Dokumentuak: 82 mila inguru
  - Domeinu ezberdinak: 6.206
- Testu-motak: periodistikoa, literatura, akademikoa, hezkuntza, webeko berezkoak (blogak, foroak...), administratiboa...

Domeinua	Dokumentuak	Hitzak
www.argia.com	2.886	4.026.006
www.zientzia.net	2.353	3.093.083
www.euskomedia.org	757	2.638.731
eibar.org	1.299	2.538.105
eu.wikipedia.org	2.422	2.281.205
www.soziolinguistika.org	690	2.102.094
www.ikasbil.net	1.220	1.932.899
klasikoak.armiarma.com	905	1.930.030
www.uztaro.com	407	1.916.843
www.euskonews.com	1.354	1.743.869
www.euskadi.net	775	1.683.893
www.armiarma.com	600	1.601.705
www.jakingunea.com	581	1.583.140
revista.consumer.es	1.518	1.459.116
paperekoa.berria.info	1.769	1.448.311
www.euskaltzaindia.net	563	1.388.255

ZENBAT du?

elhuyar Hizkuntza eta Teknologia

- Zenbat du euskara-gaztelania web-corpus paraleloak?
  - Hitzak: 18 milioi inguru
  - Segmentuak: 1,2 milioi inguru
  - Dokumentuak: 174 mila inguru
  - Domeinuak: 84
- Testu-motak: administratiboak, erakundeen webguneak, turismoa, kultura, albisteak, enpresak, alderdi politikoak...

Domeinua	Dokumentuak	Segmentuak	Hitzak
www.euskonews.com	3.444	165.736	3.014.432
www.euskalkultura.com	9.888	98.042	1.679.440
www.hiru.com	20.116	104.738	1.418.600
www.donostia.org	8.594	84.502	1.379.709
www.euskomedia.org	1.356	41.998	998.369
www.ermua.es	12.930	66.886	971.038
www.3digitala.com	16.888	66.028	829.583
www.turismoa.euskadi.net	4.528	41.530	513.617
web.bizkaia.net	3.946	31.126	398.359
www.euskadi.net	2.884	28.604	355.664
www.eusko-ikaskuntza.org	4.426	16.544	306.171
www.ibaizabal.com	5.252	27.820	305.847
www.pasaia.net	1.716	22.664	291.672
www.mondragon.edu	2.772	20.812	286.619
www.lezo.net	892	15.598	279.035
www.euskara.euskadi.net	898	16.584	256.107
www.catedralvitoria.com	2.548	14.040	251.208
www.aralar.net	1.458	12.124	247.887
www.guggenheim-bilbao.es	3.036	17.304	235.906
www.basqueresearch.com	3.836	15.980	234.674

elhuyar
Hizkuntza eta Teknologia

ZENBAT du?

elhuyar Hizkuntza eta Teknologia

- Zenbat du konbinazioen atalak?
  - Konbinazioak: 133 mila inguru
  - Agerpenak: 10 milioi inguru

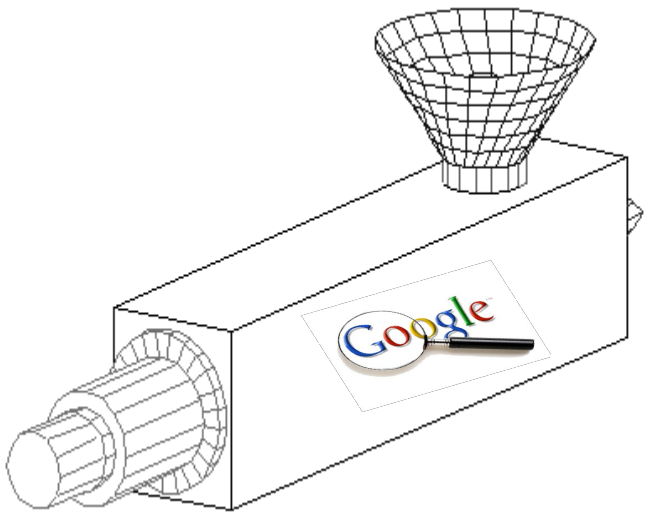
Mota	Kopurua	1. osagaiak	2. osagaiak
IZE-ADI	71.550	8.199	1.528
IZE-ADJ	27.931	4.797	2.782
IZE-IZE	33.759	7.067	7.002

NOLA egin dugu euskarazkoa?

elhuyar Hizkuntza eta Teknologia

XX. Mendeko Corpus Estatistikoko maiztasun handieneko 2.000 hitzak

Ausazko bikote-konbinazioak bidali web-bilatzaile baten APIari, sorkuntza morfologiko bidezko hedapena eginez, eta hizkuntza iragazteko hitzekin batera



Euskarazko web-corpus gordin erredundantea

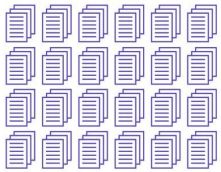
Leturia, I. (2012). "Evaluating different methods for automatically collecting large general corpora for Basque from the web". In *Proceedings of 24th International Conference on Computational Linguistics (COLING 2012)*. 2012/12, Mumbai, India

NOLA egin dugu euskarazkoa?

elhuyar Hizkuntza eta Teknologia

- **Luzera-iragazkia**: laburregiak eta luzeegiak iragazi, errore-orriak, testu gutxi dutenak eta spama saihesteko
- **Hizkuntza-iragazkia**: LangId hizkuntza-detektatzailea paragrafo mailan, testu elebidunetako zati ez euskarazkoak kentzeko
- **Orrien garbiketa**: nabigazio-menuak, oin-oharrak... kentzeko
- **Errepikatuen iragazkia**: ia berdinak diren dokumentuak iragazteko (agentzietako berriak, lizentzia libredun testuen kopiak...)
- **Barne-hartze iragazkia**: blog edo komunikabide bateko artikulu bat osorik orri nagusian agertzen denean bi bider ez sartzeko

Euskarazko web-corpus garbia



Eustagger

Euskarazko web-corpus etiketatua

Elhuyarren CAS → Web-corpusen Ataria

Leturia, I. (2012). "Evaluating different methods for automatically collecting large general corpora for Basque from the web". In *Proceedings of 24th International Conference on Computational Linguistics (COLING 2012)*. 2012/12, Mumbai, India

NOLA egin dugu paraleloa?

elhuyar
Hizkuntza eta Teknologia

Euskarazko hitz-konbinazioak

Euskarazko edukia duten domeinuak:

www.elhuyar.org

www.zientzia.net

Domeinuak eduki elebiduna du?
Alt testuak, Hizkuntza arteko estekak, Combo egiturak,...

Eduki elebiduna duten domeinuak

eu

es

San Vicente, I. & Manterola, I. (2012). "PaCo2: A Fully Automated tool for gathering Parallel Corpora from the Web".
In *Proceedings of the 8th international conference on Language Resources and Evaluation, LREC'12*. 23-25 May, Istanbul, Turkia.

NOLA egin dugu paraleloa?

elhuyar Hizkuntza eta Teknologia

Eduki elebiduna duten domeinuak

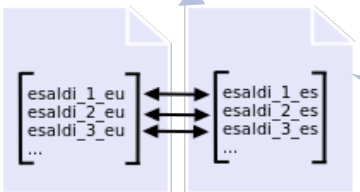
eu

es

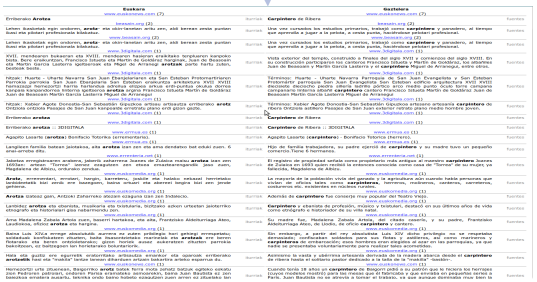
eu-es dokumentu paraleloen bilduma

Elkarren itzulpen diren dokumentuak aurkitu

- **Dokumentuen ezaugarriak bildu:** Hizkuntza (text_cat), URLa, luzera, hitz kopurua...
- **Elkarren itzulpen diren dokumentuak lotzeko:** luzera/hizkuntza heuristikoak, URLen arteko antzekotasuna, esteka-jarraitzailea, edukiaren azterketa (hizkuntza artekoa)



Elkarren itzulpen diren esaldiak parekatu: Hunalign+Postprozesua



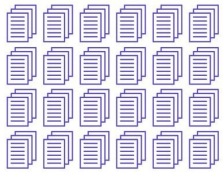
eu-es corpus paralelo garbia

San Vicente, I. & Manterola, I. (2012). "PaCo2: A Fully Automated tool for gathering Parallel Corpora from the Web". In *Proceedings of the 8th international conference on Language Resources and Evaluation, LREC'12*. 23-25 May, Istanbul, Turkia.

NOLA egin dugu hitz-konbinazioen erauzketa?

elhuyar Hizkuntza eta Teknologia

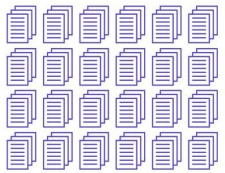
Euskarazko web-corpus etiketatua



- Corpus elebakar etiketatuaren irteera moldatu
- Interesatzen zaigun informazioa atxiki
- Analisi batzuk testuinguruaren arabera eraldatu

- egunean egun IZE ARR INE NUMS
- dokumentatutako_dokumentatu_**ADI_SINPART_GEL_0** grafien_grafia_IZE_ARR_GEN_NUMP
- indar_indar_IZE_0_0 armatuak_armatu_ADI_SINPART_ABS_NUMP → indar_indar_IZE_0_0 armatuak_armatu_**ADJ_ADIPART**_ABS_NUMP
- …

Bigrama-erauzketarako web-corpus etiketatua

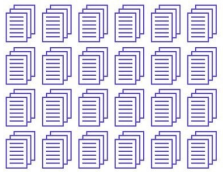


Gurrutxaga, A. & I. Alegria. “Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques."
In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011). ACL/HLT conference*. Portland, AEB.

NOLA egin dugu hitz-konbinazioen erauzketa?

elhuyar Hizkuntza eta Teknologia

Bigrama-erauzketarako web-corpus etiketatua



Bigrama-erauzketa

Bigrama-erauzketa + informazio estatistikoa

Patroi morfosintaktiko batzuk betetzen dituzten IZEADI, IZEADJ, IZEIZE agerkidetzetatik bigramak sortu (w=±1; *f* > 10)

- IZEADI: izenaren `forma_lema_kat_azpik_kasua_mugat`; aditzaren `lema` (+aspektua, flexioa...)
  - Izenaren formak normalizatu, *f* handienekora
- IZEIADJ: iz.Ø+izond. sintagmak edo sintagma-zatiak
- IZEIZE: iz. Ø +iz. sintagmak edo sintagma-zatiak (elkarteak)

Terminal — ssh — 178×57

id	l1	l2	f	f1	f2	N	am.t.score	am.log.likelihood	am.MI	am.MI3	am.chi.squared	am.z.score
10443	nahi_nahi_IZE_ARR_ABS_MG	ukan_ADI	115928	143684	426774	2817230	276.5541924	333326.1246	0.72640428	10.85478097	505872.384	638.237604
2328	behar_behar_IZE_ARR_ABS_MG	ukan_ADI	102522	214368	426774	2817230	218.7697593	142397.6740	0.49928146	10.52091560	192738.624	388.712066
13257	uste_uste_IZE_ARR_ABS_MG	ukan_ADI	56044	56371	426774	2817230	200.6644221	214195.3541	0.81709774	10.31415599	317803.094	514.066433
2315	behar_behar_IZE_ARR_ABS_MG	izan_ADI	99512	214368	629138	2817230	163.6992368	66382.3628	0.31779189	10.31354280	77627.050	236.016626
11169	parte_parte_IZE_ARR_ABS_MG	hartu_ADI	18292	22636	93411	2817230	129.6985361	106445.8006	1.38688579	9.91140818	427467.972	640.291844
8823	kontuan_kontu_IZE_ARR_INE_NUMS	hartu_ADI	14494	19298	93411	2817230	115.0761443	79664.0841	1.35510024	9.67747675	312392.106	547.690769
226	ahal_ahal_IZE_ARR_ABS_MG	izan_ADI	22077	27128	629138	2817230	107.8104326	43202.0911	0.56159867	9.24947878	55065.450	205.806991
5239	ezin_ezin_IZE_ARR_ABS_MG	izan_ADI	26592	48587	629138	2817230	96.5327738	24447.4904	0.38930730	9.23880931	29920.507	151.122259
9137	lan_lan_IZE_ARR_ABS_MG	egin_ADI	18584	55657	300597	2817230	92.7606719	21256.9122	0.49545707	9.03373546	30750.560	164.093893
2365	behera_behe_IZE_ARR_ALA_NUMS	utzi_ADI	8278	16263	28987	2817230	89.1443575	55953.1793	1.69434379	9.53019464	399519.065	626.997424
3178	deia_dei_IZE_ARR_ABS_NUMS	egin_ADI	9791	11233	300597	2817230	86.8366814	35812.4581	0.91216895	8.89382306	69232.961	248.191952
6904	hitz_hitz_IZE_ARR_ABS_MG	egin_ADI	11337	20466	300597	2817230	85.9662667	24982.5302	0.71530284	8.82429913	43264.120	195.875215
4103	erabakia_erabaki_IZE_ARR_ABS_NUMS	hartu_ADI	7961	10511	93411	2817230	85.3184028	43424.5220	1.35874787	9.16068312	172623.824	407.771297
5248	ezin_ezin_IZE_ARR_ABS_MG	ukan_ADI	19090	48587	426774	2817230	84.8953307	17026.5353	0.41391020	8.97552205	22416.856	136.722157
9036	lagun_lagun_IZE_ARR_ABS_MG	hil_ADI	6423	17523	21605	2817230	78.4668589	41778.0415	1.67939749	9.29487334	298416.712	542.481604
4822	espero_espero_IZE_ARR_ABS_MG	ukan_ADI	9651	13323	426774	2817230	77.6951952	22108.4695	0.67959464	8.64873927	34180.878	169.899258
5308	falta_falta_IZE_ARR_ABS_MG	izan_ADI	10739	12829	629138	2817230	75.9830069	21979.0205	0.57384742	8.63577511	27991.057	147.109320
10516	neurriak_neurri_IZE_ARR_ABS_NUMP	hartu_ADI	6364	9912	93411	2817230	75.6549226	31080.8866	1.28699316	8.89445350	115038.258	332.915313
1664	aurrera_aurre_IZE_ARR_ALA_NUMS	egin_ADI	8739	17482	300597	2817230	73.5289580	17027.2496	0.67070829	8.55363176	28531.913	159.152166
9927	martxan_martxa_IZE_ARR_INE_NUMS	jarri_ADI	5316	7367	47148	2817230	71.2199219	35447.6728	1.63465333	9.08582328	223009.393	467.657775
2422	berri_berri_IZE_ARR_ABS_MG	eman_ADI	5480	10005	110195	2817230	68.7405445	22355.8524	1.14622388	8.62378500	69107.205	257.231949
11221	partida_partida_IZE_ARR_ABS_MG	jokatu_ADI	5036	14837	36012	2817230	68.2922156	25847.8516	1.42411469	8.82828613	126106.546	351.907519
341	akordioa_akordio_IZE_ARR_ABS_NUMS	lortu_ADI	4921	11638	33380	2817230	68.1841466	28691.4956	1.55251098	8.93661771	168597.863	407.322970

Gurrutxaga, A. & I. Alegria. “Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques."
In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011). ACL/HLT conference*. Portland, AEB.

ZERTARAKO balio dezake?

elhuyar
Hizkuntza eta Teknologia

- http://webcorpusak.elhuyar.org

Hurrengo urratsak

elhuyar Hizkuntza eta Teknologia

- Euskarazko web-corpus elebakarrean
  - Handitu crawling metodoa erabiliz (bilatzaileetatik ezin da askoz gehiago lortu) eta online jarri
  - Corpuseko testuak karakterizatu (eremua, generoa...)
- Euskara-gaztelania web-corpus paraleloan
  - Esalditik beherako unitate-mailan parekatu (hitzak, terminoak, fraseologia...)
- Hitz-konbinazioak
  - Idiomatikotasuna karakterizatzeko teknika aurreratuak (konposizionaltasuna)
  - Konbinazio-mota gehiago: IZLG+IZE, ADB+ADI...
  - Patroi morfosintaktikoak (argumentuak, azpikategoria...)
  - Fraseologia elebiduna

Web-corpusen Ataria

elhuyar
Hizkuntza eta Teknologia

Eskerrik asko!

Web-corpusen Ataria

elhuyar
Hizkuntza eta Teknologia

Igor Leturia, Iñaki San Vicente, Iker Manterola, Antton Gurrutxaga

IEB 2013

Miramon, 2013ko maiatzaren 8a