Analysis and performance of morphological query expansion and language-filtering words on Basque web searching

I. Leturia, A. Gurrutxaga, N. Areta, E. Pociello

Elhuyar R&D, Usurbil, Basque Country

LREC 2008 – May 29, 2008 – Marrakech

Contents

- Introduction
- Current study
- Morphological query expansion
- Language-filtering words
- Conclusions

Contents

- **Introduction**
- Current study
- Morphological query expansion
- Language-filtering words
- Conclusions

Basque IR problems

- Looking for conjugations and inflections
  - Basque is an agglutinative language
  - A given lemma makes many different surface forms: *lan* ("work"), *lana* ("the work"), *lanak* ("the works"), *lanari* ("to the work"), *lanei* ("to the works"), *lanaren* ("of the work")...
  - Looking only for the exact given word, or the word plus an "s" for the plural, is not enough
  - Wildcards are not an appropriate solution: looking for lan* would also return forms of the words *lanabes* ("tool"), *lanbro* ("fog")...

Basque IR problems

- Language discrimination
  - No search engine offers the possibility of returning only pages in Basque
  - Big problem when looking for technical words that exist also in other languages (*anorexia*, *sulfuroso*, *byte*, *allegro*, *sistema*, *energia*...), short words (*katu*, *ur*...) or proper nouns (*Egipto*, *Newton*, *Pluton*...)
  - Many non-Basque results are returned, often no Basque results at all

Our approach

- API based
  - We use APIs of major search engines
  - Cost-effective solution
  - NLP techniques applied to obtain better results

Our approach

- Morphological query expansion or MQE (I)
  - We use a morphological generator for Basque created by the IXA Group of the University of the Basque Country
  - We obtain all the forms of a given lemma
  - We ask the search engine for all of them using an OR operator
  - etxe => etxe OR etxea OR etxeak OR etxeari OR etxeei OR etxeek OR…

Our approach

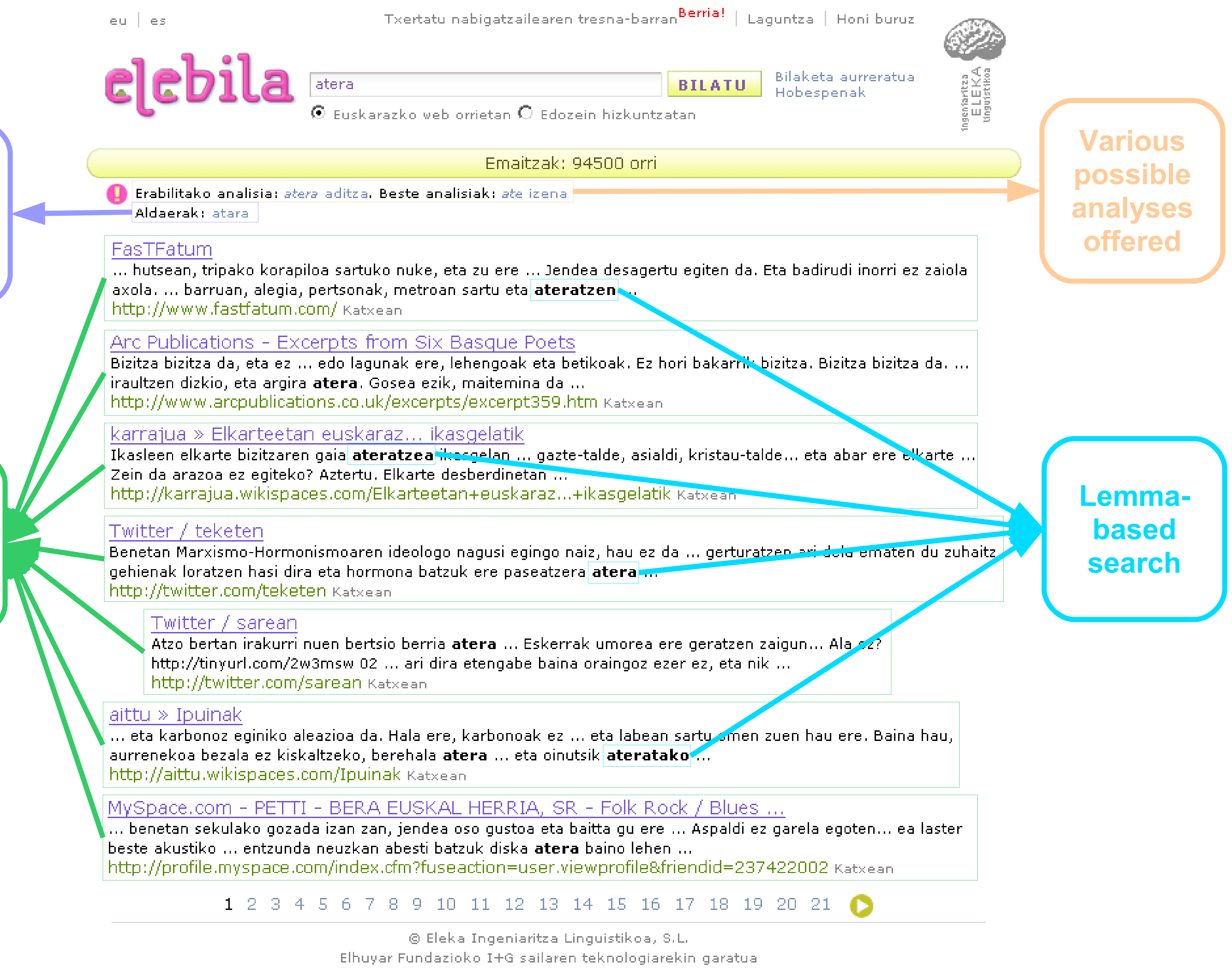
- Morphological query expansion or MQE (II)
  - The APIs of the search engines have each a limit in number of words of the queries
  - This makes real lemma-based search impossible
  - But good results can be obtained if the forms sent in the query are the most frequent ones

Our approach

- Language-filtering words or LFW
  - Some of the most frequent Basque words are added to the query using an AND operator
  - Several LFWs have to be used, since the most frequent words in Basque exist in other languages too
  - The more LFWs used, the better language-precision we obtain, but with loss in recall

Tools built

- Elebila
  - Search service for Basque
  - API based
  - Lemma-based search (MQE)
  - Returns pages in Basque alone (LFWs)
  - Optional search for variants of words
  - Optional lemma-based search for whole noun phrases or terms (including them in double quotes)
  - http://www.elebila.eu



eu | es

Txertatu nabigatzailearen tresna-barran Berria! | Laguntza | Honi buruz

elebila

atera BILATU

Bilaketa aurreratua
Hobespenak

● Euskarazko web orrietan ○ Edozein hizkuntzatan

Emaitzak: 94500 orri

Erabilitako analisia: atera aditza. Beste analisiak: ate izena
Aldaerak: atara

FasTFatum
... hutsean, tripako korapiloa sartuko nuke, eta zu ere ... Jendea desagertu egiten da. Eta badirudi inorri ez zaiola axola. ... barruan, alegia, pertsonak, metroan sartu eta **ateratzen** ...
http://www.fastfatum.com/ Katxean

Arc Publications - Excerpts from Six Basque Poets
Bizitza bizitza da, eta ez ... edo lagunak ere, lehengoak eta betikoak. Ez hori bakarrik bizitza. Bizitza bizitza da. ... iraultzen dizkio, eta argira **atera**. Gosea ezik, maitemina da ...
http://www.arcpublications.co.uk/excerpts/excerpt359.htm Katxean

karrajua » Elkarteetan euskaraz... ikasgelatik
Ikasleen elkarte bizitzaren gaia **ateratzea** ikasgelan ... gazte-talde, asialdi, kristau-talde... eta abar ere elkarte ... Zein da arazoa ez egiteko? Aztertu. Elkarte desberdinetan ...
http://karrajua.wikispaces.com/Elkarteetan+euskaraz...+ikasgelatik Katxean

Twitter / teketen
Benetan Marxismo-Hormonismoaren ideologo nagusi egingo naiz, hau ez da ... gerturatzen ari dela ematen du zuhaitz gehienak loratzen hasi dira eta hormona batzuk ere paseatzera **atera** ...
http://twitter.com/teketen Katxean

Twitter / sarean
Atzo bertan irakurri nuen bertsio berria **atera** ... Eskerrak umorea ere geratzen zaigun... Ala ez? http://tinyurl.com/2w3msw 02 ... ari dira etengabe baina oraingoz ezer ez, eta nik ...
http://twitter.com/sarean Katxean

aittu » Ipuinak
... eta karbonoz eginiko aleazioa da. Hala ere, karbonoak ez ... eta labean sartu omen zuen hau ere. Baina hau, aurrenekoa bezala ez kiskaltzeko, berehala **atera** ... eta oinutsik **ateratako** ...
http://aittu.wikispaces.com/Ipuinak Katxean

MySpace.com - PETTI - BERA EUSKAL HERRIA, SR - Folk Rock / Blues ...
... benetan sekulako gozada izan zan, jendea oso gustoa eta baitta gu ere ... Aspaldi ez garela egoten... ea laster beste akustiko ... entzunda neuzkan abesti batzuk diska **atera** baino lehen ...
http://profile.myspace.com/index.cfm?fuseaction=user.viewprofile&friendid=237422002 Katxean

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

© Eleka Ingeniaritza Linguistikoa, S.L.
Elhuyar Fundazioko I+G sailaren teknologiarekin garatua

Variant suggestion

Various possible analyses offered

All results in Basque

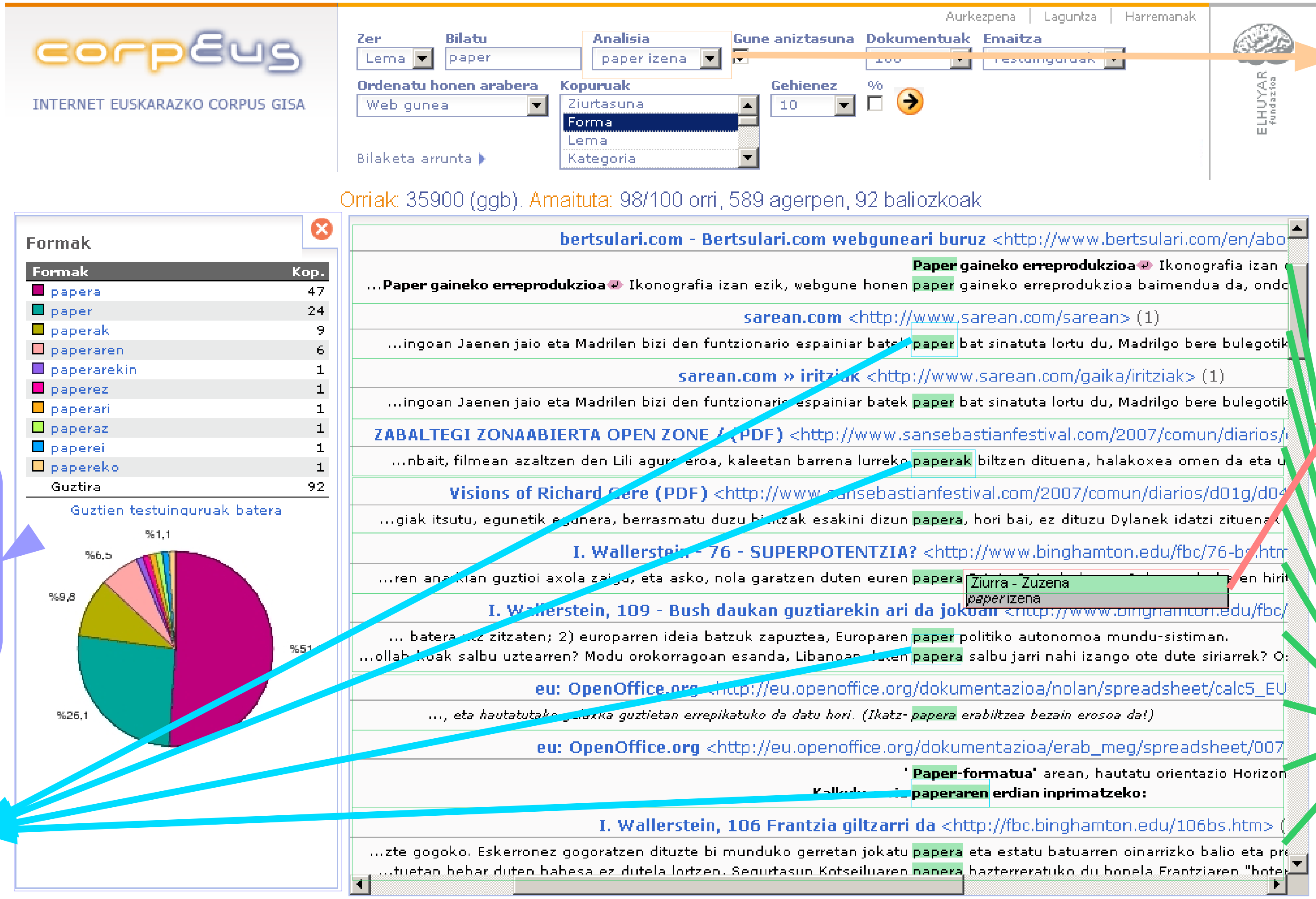
Lemma-based search

Tools built

- CorpEus (I)
  - Web-as-corpus tool for Basque
  - API based
  - Lemma-based search (MQE)
  - Returns occurrences in Basque alone (LFWs)
  - Optional search for variants of words
  - Optional lemma-based search for whole noun phrases or terms (including them in double quotes)

Tools built

- CorpEus (II)
  - Parallel downloading of pages
  - Analyses of the results
  - Different ordering criteria
  - Occurrence counts and charts
  - http://www.corpeus.org



corpEus

INTERNET EUSKARAZKO CORPUS GISA

Aurkezpena | Laguntza | Harremanak

Orriak: 35900 (ggb). Amaituta: 98/100 orri, 589 agerpen, 92 baliozkoak

Formak

Formak	Kop.
papera	47
paper	24
paperak	9
paperaren	6
paperarekin	1
paperez	1
paperari	1
paperaz	1
paperei	1
papereko	1
Guztira	92

Guztien testuinguruak batera

bertsulari.com - Bertsulari.com webguneari buruz <http://www.bertsulari.com/en/abo

Paper gaineko erreprodukzioa Ikonografia izan

...Paper gaineko erreprodukzioa Ikonografia izan ezik, webgune honen paper gaineko erreprodukzioa baimendua da, ondo

sarean.com <http://www.sarean.com/sarean> (1)

...ingoan Jaenen jaio eta Madrilen bizi den funtzionario espainiar batek paper bat sinatuta lortu du, Madrilgo bere bulegotik

sarean.com » iritziak <http://www.sarean.com/gaika/iritziak> (1)

...ingoan Jaenen jaio eta Madrilen bizi den funtzionario espainiar batek paper bat sinatuta lortu du, Madrilgo bere bulegotik

ZABALTEGI ZONAABIERTA OPEN ZONE / (PDF) <http://www.sansebastianfestival.com/2007/comun/diarios/

...nbait, filmean azaltzen den Lili agure eroa, kaleetan barrena lurreko paperak biltzen dituena, halakoxea omen da eta u

Visions of Richard Gere (PDF) <http://www.sansebastianfestival.com/2007/comun/diarios/d01g/d04

...giak itsutu, egunetik egunera, berrasmatu duzu hitzak esakini dizun papera, hori bai, ez dituzu Dylanek idatzi zituenak

I. Wallerstein - 76 - SUPERPOTENTZIA? <http://www.binghamton.edu/fbc/76-bs.htm

...ren anaiakian guztioi axola zaigu, eta asko, nola garatzen duten euren papera en hirit

Ziurra - Zuzena
paper izena

I. Wallerstein, 109 - Bush daukan guztiarekin ari da jokoan <http://www.binghamton.edu/fbc/

... batera ezz zitzaten; 2) europarren ideia batzuk zapuztea, Europaren paper politiko autonomoa mundu-sistiman.

...ollabakoak salbu uztearren? Modu orokorragoan esanda, Libanoan duten papera salbu jarri nahi izango ote dute siriarrek? O:

eu: OpenOffice.org <http://eu.openoffice.org/dokumentazioa/nolan/spreadsheet/calc5_EU

..., *eta hautatutako gelaxka guztietan errepikatuko da datu hori. (Ikatz- papera erabiltzea bezain erosoa da!)*

eu: OpenOffice.org <http://eu.openoffice.org/dokumentazioa/erab_meg/spreadsheet/007

'**Paper-formatua**' arean, hautatu orientazio Horizon

Kalkulu-orria paperaren erdian inprimatzeko:

I. Wallerstein, 106 Frantzia giltzarri da <http://fbc.binghamton.edu/106bs.htm> (

...zte gogoko. Eskerronez gogoratzen dituzte bi munduko gerretan jokatu papera eta estatu batuarren oinarrizko balio eta pre

...tuetan behar duten babesa ez dutela lortzen. Segurtasun Kotseiluaren papera hazterreratuko du honela Frantziaren "hoter

Various possible analyses offered

Analysis of the results

All results in Basque

Occurrence counts and charts

Lemma-based search

Contents

- Introduction
- Current study
- Morphological query expansion
- Language-filtering words
- Conclusions

Contents

- Introduction
- **Current study**
- Morphological query expansion
- Language-filtering words
- Conclusions

Current study

- Analysis and performance measurement of MQE and LFWs
- Corpora based

Current study

- Implementation details of the methodology very important in its performance
    - Cases for MQE
    - Which and how many LFWs
- Previously
    - LFWs chosen based on a classic corpus
    - Cases for MQE quite intuitively
    - Improvement not measured quantitatively

Corpora used

- ZT Corpusa
  - Corpus of Science and Technology
  - 7.6 million words
- A web corpus
  - Downloaded all the pages of the Basque branch of Google Directory (+3,000) and recursively followed links of pages in Basque
  - 44,000 documents
  - 20 million words

Words used

- Some words needed to perform the various measurements
  - For observing the most frequent cases for MQE
  - For measuring the language-precision obtained by LFWs
- Most asked-for words of the Elebila logs
  - Four months, 400,000 queries, 800,000 words, 70,000 different words
  - Lemmatised and used the most frequent ones

Contents

- Introduction
- Current study
- Morphological query expansion
- Language-filtering words
- Conclusions

Contents

- Introduction
- Current study
- **Morphological query expansion**
- Language-filtering words
- Conclusions

Most frequent cases

- Observed which are the most frequent cases
  - For each POS
  - Using the most frequently asked-for words of Elebila
  - Using both corpora
  - We have opted for the web corpus lists

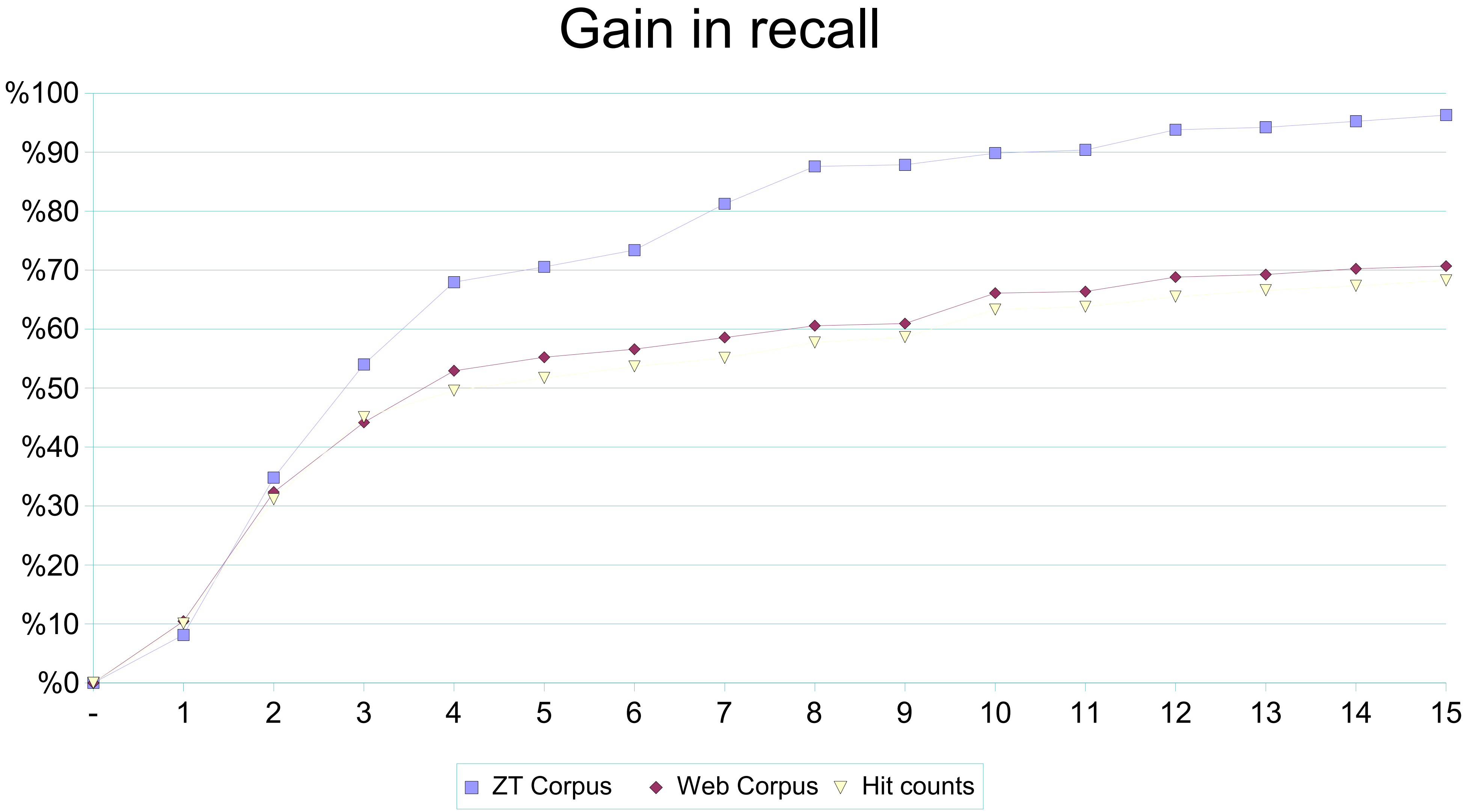
Most frequent cases

	Verb	Adjective	Noun	Proper noun	Place name
1	Participle / perfective aspect (*sortu*)	Nominative singular (*berria*)	Nominative indefinite (*hiztegi*)	Nominative (*Mikel*)	Nominative (*Egipto*)
2	Imperfective aspect (*sortzen*)	Nominative plural/Ergative singular (*berriak*)	Nominative singular (*hiztegia*)	Ergative (*Mikelek*)	Genitive locative (*Egiptoko*)
3	Verbal noun + *-ko* (*sortzeko*)	Nominative indefinite (*berri*)	Nominative plural/Ergative singular (*hiztegiak*)	Genitive (*Mikelen*)	Inessive (*Egipton*)
4	Unrealized aspect (*sortuko*)	Genitive plural (*berrien*)	Genitive locative singular (*hiztegiko*)	Dative (*Mikeli*)	Allative (*Egiptora*)
5	Short stem (*sor*)	Inessive singular (*berrian*)	Genitive singular (*hiztegiaren*)	Associative (*Mikelekin*)	Ablative (*Egiptotik*)
6	Verbal noun + Nominative singular (*sortzea*)	Genitive singular (*berriaren*)	Dative singular (*hiztegiari*)	Genitive + Nominative singular (*Mikelena*)	Genitive (*Egiptoren*)
7	Adjectival participle (*sortutako*)	Associative singular (*berriarekin*)	Inessive singular (*hiztegian*)	Partitive (*Mikelik*)	Dative (*Egiptori*)
8	Participle + Nominative singular (*sortua*)	Ergative indefinite (*berrik*)	Partitive (*hiztegirik*)	Genitive + Nominative Plural/Ergative singular (*Mikelenak*)	Genitive locative + Nominative singular (*Egiptokoa*)
9	Dynamic adverbial participle (*sortuz*)	Dative singular (*berriari*)	Instrumental indefinite (*hiztegiz*)	Instrumental (*Mikelez*)	Allative + Genitive locative (*Egiptorako*)
10	*-ta*/*-da* stative adverbial participle (*sortuta*)	Instrumental indefinite (*berriz*)	Instrumental singular (*hiztegiaz*)	Inessive (*Mikelengan*)	Associative (*Egiptorekin*)
11	Participle + Nominative plural/Ergative singular (*sortuak*)	Inessive indefinite (*berritan*)	Genitive singular + Nominative singular (*hiztegiarena*)		Genitive locative + Nominative plural/Ergative singular (*Egiptokoak*)
12	Verbal noun + Inessive singular (*sortzean*)	Sociative plural (*berriekin*)	Genitive plural (*hiztegien*)		Destinative (*Egiptorentzat*)
13	*-(r)ik* stative adverbial participle (*sorturik*)	Inessive plural (*berrietan*)	Sociative singular (*hiztegiarekin*)		Instrumental (*Egiptoz*)
14	Verbal noun + Allative singular (*sortzera*)	Genitive locative singular (*berriko*)	Ablative singular (*hiztegitik*)		Terminal allative (*Egiptoraino*)
15	Adjectival participle + Nominative plural/Ergative singular (*sortutakoak*)	Partitive (*berririk*)	Allative singular (*hiztegira*)		Genitive locative + Inessive singular (*Egiptokoan*)

Gain in recall

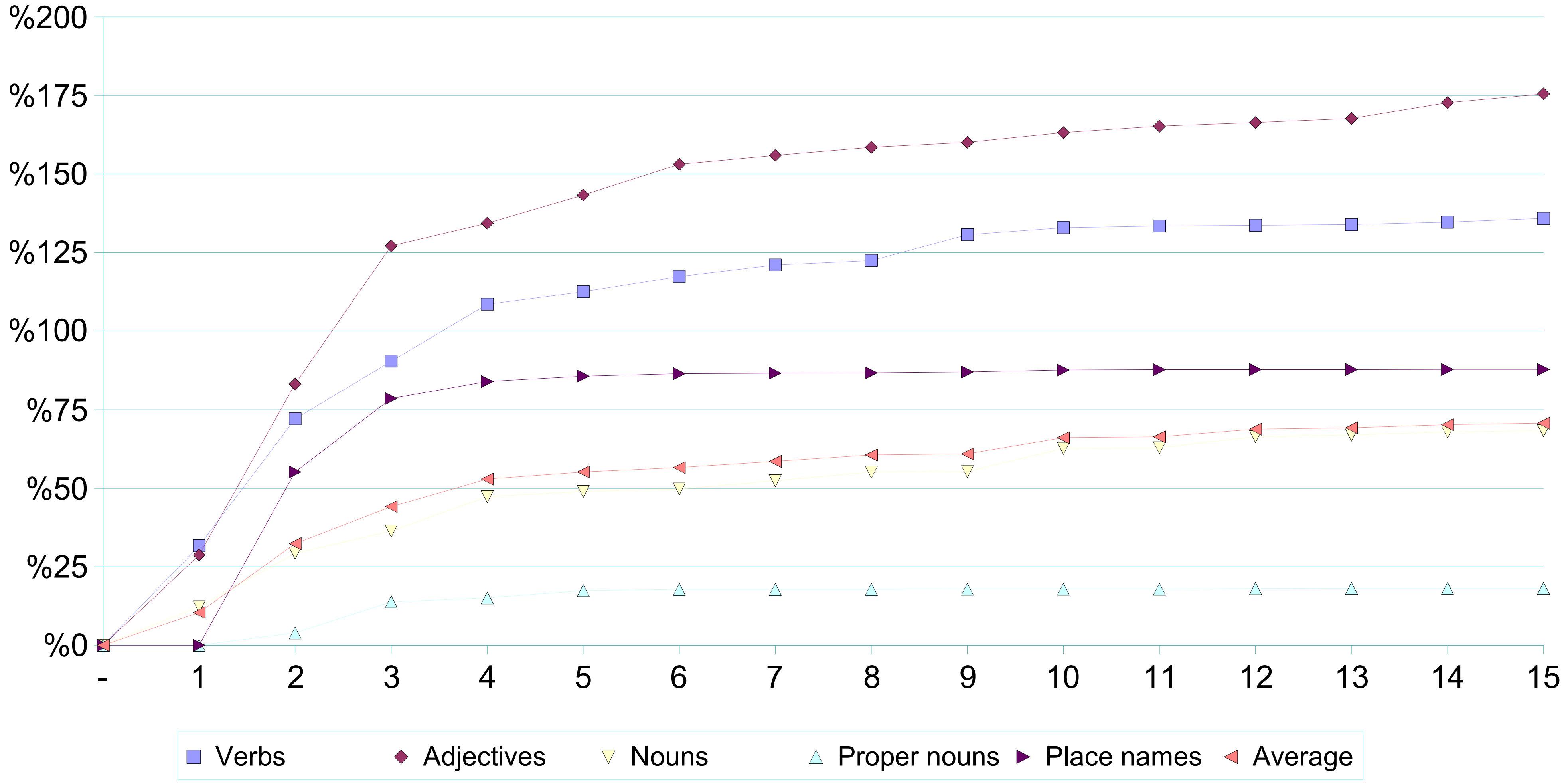
- Measured the gain in recall that would be obtained by including 1, 2, 3... of the most frequent cases in the queries within OR operators
- Using both corpora and also hit counts of Microsoft's Live Search API
- The remarkable similarity between the web corpus and hit counts series prove our supposition that it was better to base our study in a web corpus

Gain in recall



Gain in recall

Gain in recall for each POS in the web corpus



Contents

- Introduction
- Current study
- Morphological query expansion
- Language-filtering words
- Conclusions

Contents

- Introduction
- Current study
- Morphological query expansion
- **Language-filtering words**
- Conclusions

Choosing the words

- Language-filtering words need to be:
  - Very frequent, so that as many Basque pages as possible contain them
  - Specifically Basque, so that as few pages in other languages as possible contain them
- Observed which are the most frequent Basque words
  - Using both corpora

Choosing the words

Web corpus		**ZT Corpus**	
eta ("and")	91.94%	*eta* ("and")	98.44%
da ("is")	74.37%	*da* ("is")	92.67%
ez ("no")	64.51%	*ez* ("no")	79.05%
du ("has")	64.11%	*dira* ("are")	78.65%
bat ("a")	62.81%	*ere* ("too")	78.27%
ere ("too")	55.65%	*du* ("has")	75.49%
dira ("are")	55.45%	*izan* ("be")	73.45%
izan ("be")	54.24%	*dute* ("have")	72.14%
egin ("do")	52.77%	*bat* ("a")	67.66%
beste ("other")	47.74%	*baina* ("but")	64.41%
edo ("or")	42.94%	*den* ("that is")	64.04%
dute ("have")	41.72%	*egin* ("do")	62.56%
den ("that is")	39.19%	*beste* ("other")	57.21%
egiten ("doing")	38.98%	*baino* ("than")	56.77%
baina ("but")	36.94%	*egiten* ("doing")	55.78%
baino ("than")	27.29%	*edo* ("or")	55.59%

Choosing the words

- We have opted for the web corpus words
- *eta* and *da* are the clear first candidates because of the significant difference in frequency with the next ones
- Tried precision-recall on different combinations of the first six words: *eta*, *da*, *ez*, *du*, *bat* and *ere*

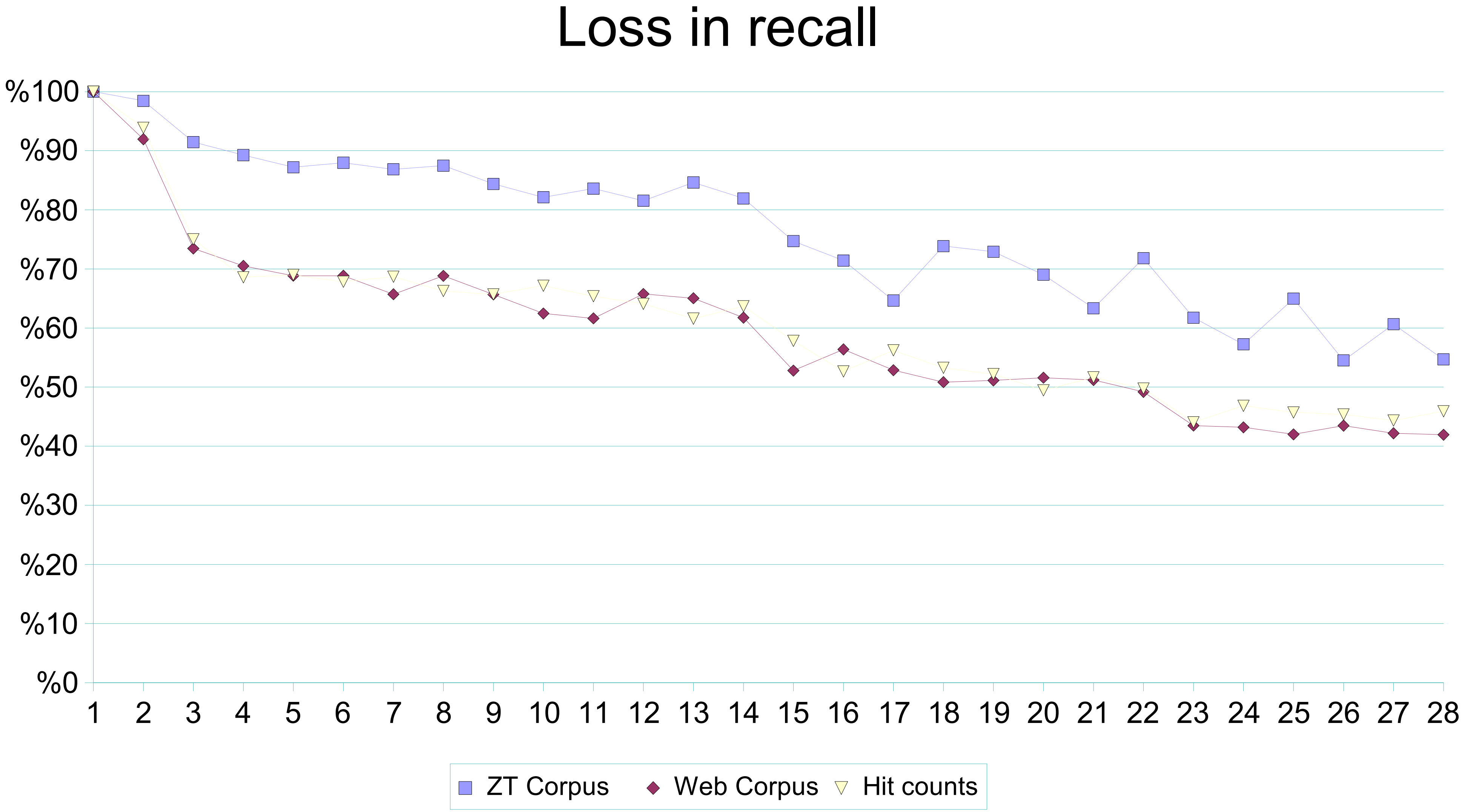
Choosing the words

Combinations	
0 words	-
1 word	*eta*
2 words	*eta* AND *da*
3 words	*eta* AND *da* AND (*ez* OR *du* OR *bat* OR *ere*)
	eta AND *da* AND (*ez* OR *du* OR *bat*)
	eta AND *da* AND (*ez* OR *du* OR *ere*)
	eta AND *da* AND (*ez* OR *bat* OR *ere*)
	eta AND *da* AND (*du* OR *bat* OR *ere*)
	eta AND *da* AND (*ez* OR *du*)
	eta AND *da* AND (*ez* OR *bat*)
	eta AND *da* AND (*ez* OR *ere*)
	eta AND *da* AND (*du* OR *bat*)
	eta AND *da* AND (*du* OR *ere*)
	eta AND *da* AND (*bat* OR *ere*)
	eta AND *da* AND *ez*
	eta AND *da* AND *du*
	eta AND *da* AND *bat*
	eta AND *da* AND *ere*
4 words	*eta* AND *da* AND *ez* AND (*du* OR *bat* OR *ere*)
	eta AND *da* AND *du* AND (*ez* OR *bat* OR *ere*)
	eta AND *da* AND *bat* AND (*ez* OR *du* OR *ere*)
	eta AND *da* AND *ere* AND (*ez* OR *du* OR *bat*)
	eta AND *da* AND *ez* AND *du*
	eta AND *da* AND *ez* AND *bat*
	eta AND *da* AND *ez* AND *ere*
	eta AND *da* AND *du* AND *bat*
	eta AND *da* AND *du* AND *ere*
	eta AND *da* AND *bat* AND *ere*

Loss in recall

- Measured the loss in recall
- Using both corpora and also hit counts of Microsoft's Live Search API
- Again the remarkable similarity between the web corpus and hit counts series confirm our supposition that it was better to base our study in a web corpus

Loss in recall

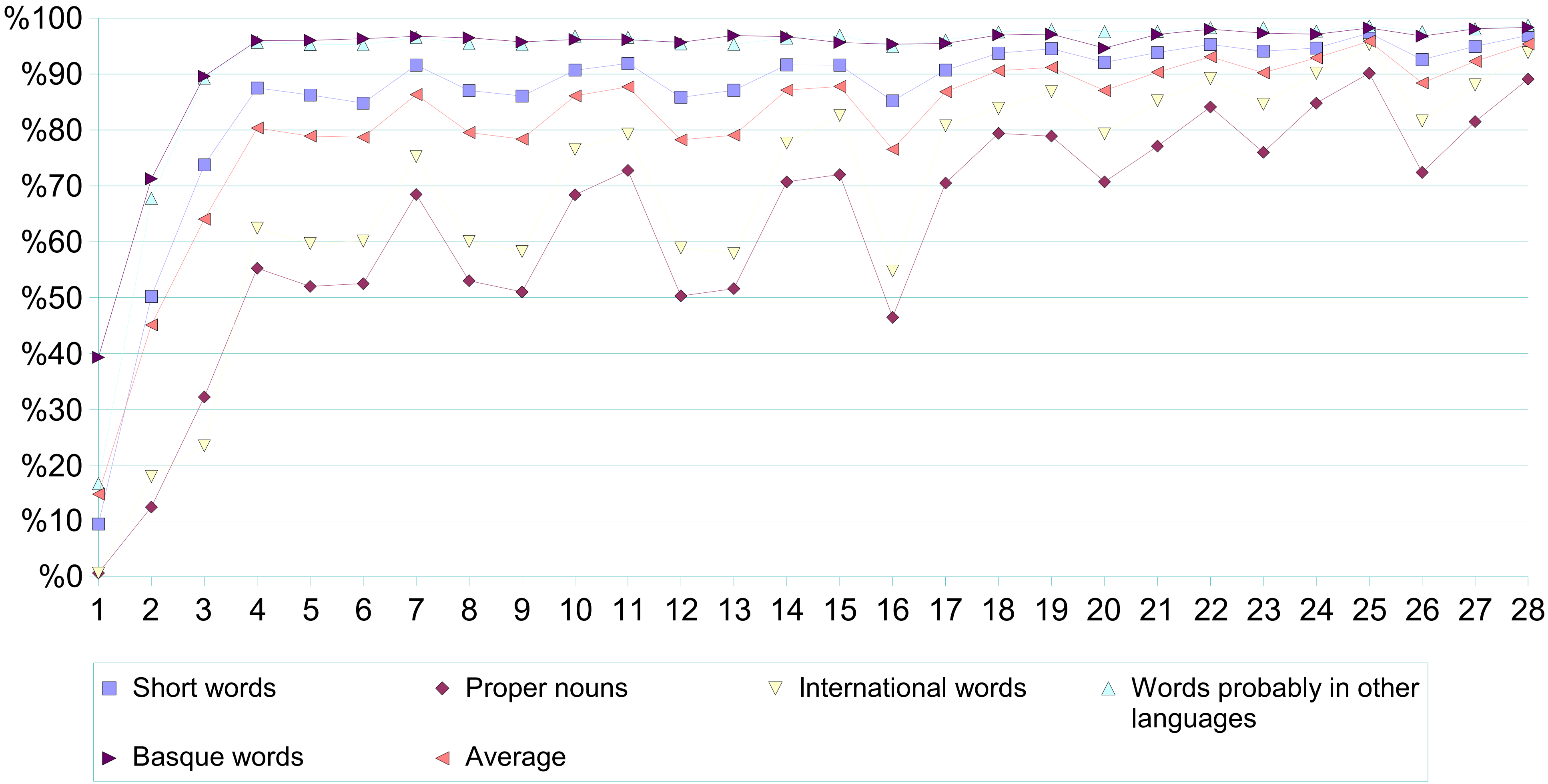


Gain in language-precision

- Impossible to measure the gain in precision over corpora: we would need a multilingual corpus with the same proportions of each language as in the web
- Instead used Microsoft Live Search's API
- Combined with LangId, an automatic language classifier specialized on Basque
- Applied LangId to the snippets returned by the API

Gain in language-precision

Gain in precision for each category of word



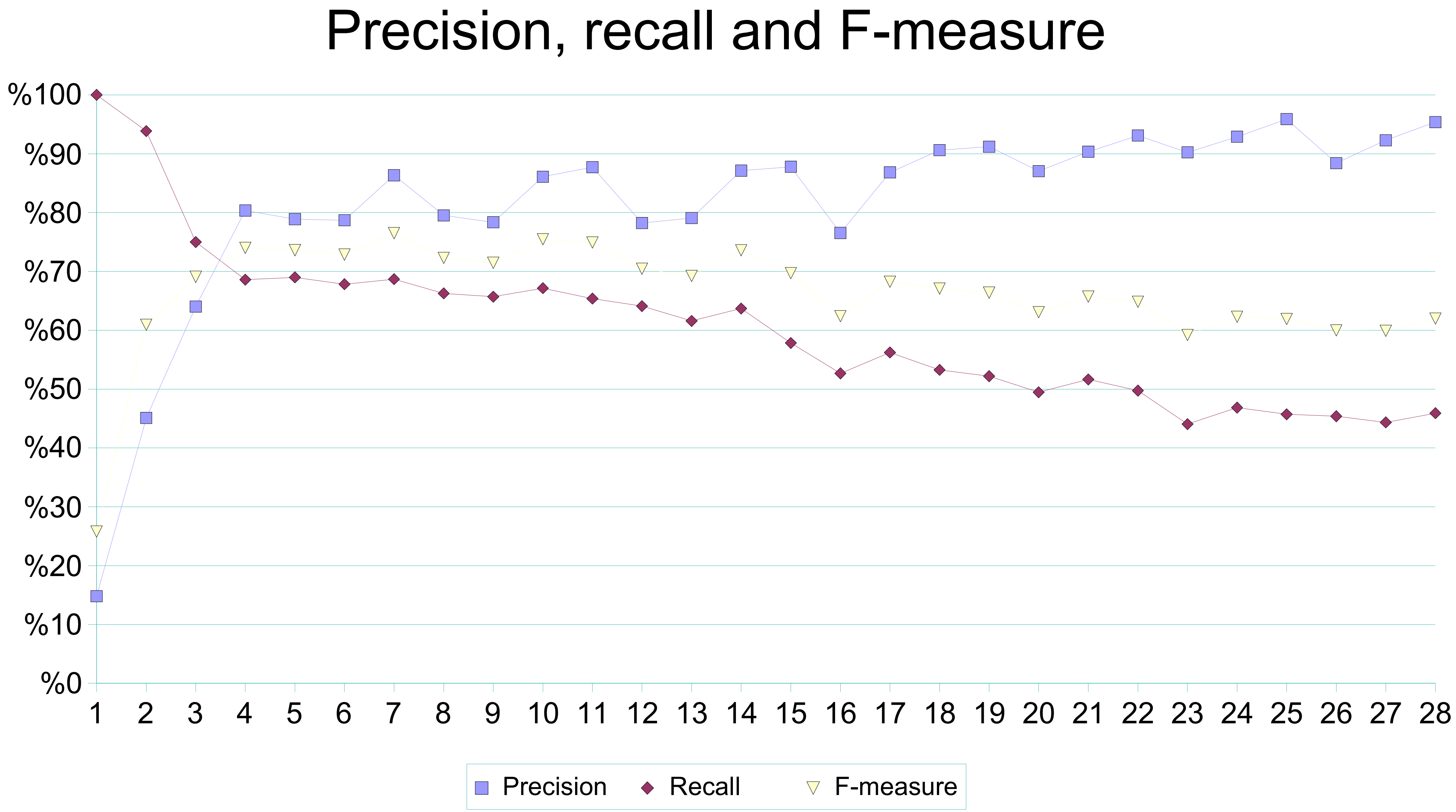
Gain in language-precision

- Observing the peaks and valleys gives indications as to which can be the best and worst words for being LFWs
  - Valleys contain *du* (a very common French word)
  - Peaks contain *ere*, but *bat* and *ez* also perform well

Best LFW combination

- For choosing the best LFW combination, we put together the precision and recall, and also the F-measure

Best LFW combination



Best LFW combination

- 4-word combinations can obtain a language-precision high above 90%, but with a recall near or below 50%
- 3-word combinations without *du* are the ones with the highest F-measure, as they achieve a precision of 86-87% and a recall of 68-65%; but for proper nouns or international words precision falls to 70%

Best LFW combination

- Two implementation options
  - Keep a database of the most searched-for proper nouns an international words, and use a 4-word combination for them and a 3-word combination otherwise
  - Use a 4-word combination by default to prioritise precision and, if the user does not find what he/she was looking for, offer the possibility of retry increasing recall (with a 3-word combination)

Contents

- Introduction
- Current study
- Morphological query expansion
- Language-filtering words
- Conclusions

Contents

- Introduction
- Current study
- Morphological query expansion
- Language-filtering words
- **Conclusions**

Conclusions

- This study has produced very valuable data for Basque IR projects (most frequent cases for MQE, best word combinations for LFWs, etc.)
- Specifically, they will soon be applied in the Basque web services Elebila and CorpEus

Conclusions

- The study has also produced quantitative precision-recall measurements, proving that MQE and LFWs clearly improve the performance of search engines for Basque
  - LFWs raise precision from 15% to even 90%, although with a non-negligible loss in recall
  - MQE can improve recall up to 70%

Conclusions

- MQE and LFWs can be valid for building web IR services for other agglutinative or under-resourced languages in a cost-effective way, and also the corpora-based methodology described here can be used to define the implementation details and measure the improvement obtained

Analysis and performance of morphological query expansion and language-filtering words on Basque web searching

I. Leturia, A. Gurrutxaga, N. Areta, E. Pociello

Elhuyar R&D, Usurbil, Basque Country

LREC 2008 – May 29, 2008 – Marrakech