Exploiting the Internet to build language resources for less-resourced languages

Antton Gurrutxaga Igor Leturia Eli Pociello
Iñaki San Vicente Xabier Saralegi

Elhuyar Foundation
R & D

Creation and use of basic lexical resources for less-resourced languages
7th SaLTMiL Workshop on
LREC 2010, Valetta, Malta
2010/05/23

Outline

1. Motivation
   - General motivation
   - Elhuyar's particular motivation
2. Using the web to build corpora
   - Monolingual specialized corpora
   - Multilingual domain-comparable corpora
   - Other kinds of corpora
3. Building other kinds of resources
   - A web-as-corpus tool
   - Terminology
   - Ontologies
4. Experiments
   - Experiments
   - Monolingual specialized web corpora
   - Multilingual domain-comparable web corpora
5. Conclusions
   - Conclusions
6. References

Need for corpora

- Any modern language needs language resources such as corpora or dictionaries
  - For use in everyday life (education, media. . .)
  - For development of language technologies
- Corpora are essential
  - For obtaining any kind of linguistic evidence
  - For lexicography and terminology
    - Manual
    - Via (semi)automatic extraction tools
- The bigger the corpora, the better

Turning to the Internet

- Less-resourced languages are not rich in corpora
  - Building them the classical way (out of printed texts) very costly
  - Less researchers, linguists, lexicographers...
- The Internet
  - Huge number of texts
  - In digital format
  - In an open standard format
- Turning to the Internet to build corpora (and then automatically also dictionaries) very logical for less-resourced languages

About the Elhuyar Foundation

- A non-profit organization aimed to the promotion of the Basque language, specially in the fields of Science and Technology, founded in 1972
- Popular Science
  - TV program
  - Radio program
  - Magazines
  - Websites
  - Museums
- Editions
  - Popular science books
  - Educational material
  - Multimedia
- Language services
  - Translations
  - Dictionary making (1990)
  - R & D on language technologies (2000)

Elhuyar R & D research areas

- Corpora building, lexicon and terminology extraction, ontologies
  - Dictionary making
  - R & D
- Machine translation
  - Translations
- Information retrieval
  - Popular science
  - Editions

Monolingual specialized corpora

- Corpora made out of texts belonging to a certain domain or topic
- Very valuable
  - Monolingual terminology tasks, either manual or automatic
  - NLP research
- Approaches to obtain monolingual specialized corpora from the web
  - Crawling websites dedicated to the domain, sometimes with machine-learning filters
  - Asking search engines for 3/4 word combinations of seed words on the topic, as in BootCaT (Baroni and Bernardini, 2004)

Problems of Basque language

- Crawling method:
  - Not many websites specialized on a topic with a significant amount of texts
  - No training data for building machine-learning filters
- Search engine method:
  - 66% domain-precision falls to 25% in the case of Basque (Leturia et al., 2008b)
    - No search engine returns pages only in Basque; looking for technical words yield many non-Basque results
    - A Basque lemma has many different word forms; looking only for the lemma brings fewer results

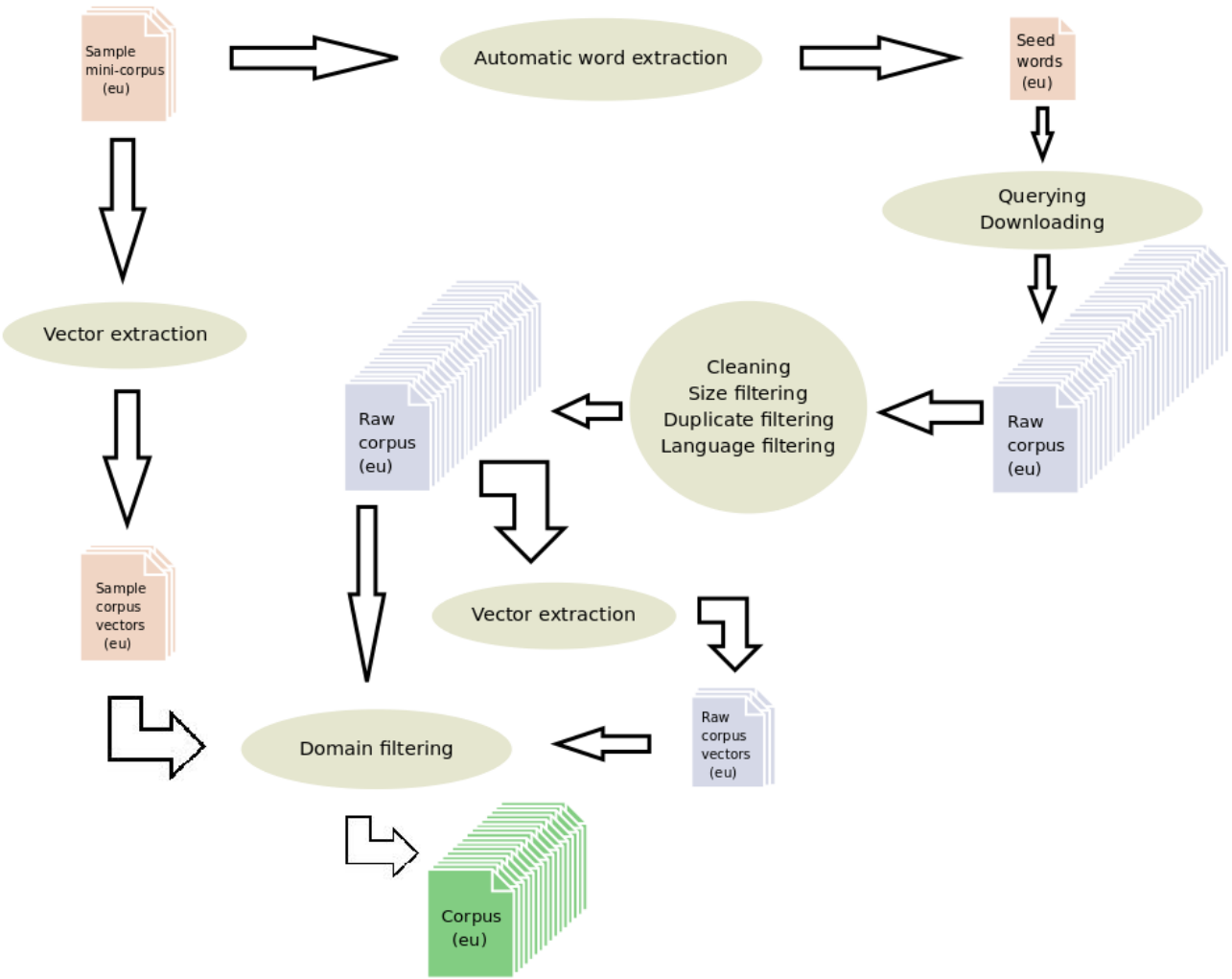
Our approach I

- AutoCorpEx, based on search engines (Leturia et al., 2008b)
- Start from a sample mini-corpus of documents on the topic, as varied as possible
- List of seed terms automatically extracted from it, (manually edited and improved if necessary)
- 3 word combinations sent to APIs of search engines, with techniques for optimization for Basque (Leturia et al., 2008a)
  - Add language-filtering words for obtaining only pages in Basque
  - Morphological query expansion for lemma-based search (inflections of the words with an OR operator)
  - The search string for *etxe* would be `(etxe OR etxea OR etxeak OR etxeek OR etxean OR etxeetan) eta da ez ere`
  - These two techniques raise topic-precision to the 66% baseline

Our approach II

- Download pages returned by search engines
- Various cleaning and filtering stages
  - Boilerplate stripping (Saralegi and Leturia, 2007)
  - Size filtering (Fletcher, 2004)
  - Paragraph-level language filtering
  - Near-duplicate filtering (Broder, 2000)
  - Containment filtering (Broder, 1997)
- Topic filtering stage
  - Initial sample mini-corpus as reference
  - Using document similarity techniques (Saralegi and Alegria, 2007)
- Topic precision of 90% can be achieved

Our approach III



Multilingual domain-comparable corpora

- Multilingual corpora where all the texts are in the same domain
- Very valuable
  - Bilingual terminology extraction
  - Machine translation training
- Compared to parallel corpora
  - More difficult to exploit
  - Easier to obtain in large sizes

Our approach I

- Co3 (Comparable Corpus Collector), domain-comparable corpora builder with the method explained above (Leturia et al., 2009)
  - Use a sample mini-corpus for each language, comparable enough between them (ideally parallel)
  - Launch the corpus-collecting processes independently
  - Both of the corpora obtained will be on the same domain so we get a domain-comparable corpus

Our approach II

- Alternative
  - Use a single mini-corpus
  - Translate the seed word to the other language using a dictionary
  - Download the corpora
  - Translate the keyword vectors for the topic filtering using a dictionary
- Advantages:
  - Sample mini-corpora as similar as possible (there is only one)
  - Less initial work for collecting the mini-corpora
- Drawbacks:
  - Out Of Vocabulary (OOV) words: keep them as they are (many named entities)
  - Ambiguity: first translation approach
- Similar results

Monolingual general corpora

- Very interesting
  - Language standardization
  - General lexicography
  - Discourse analysis
  - ...
- Approaches to obtain monolingual general corpora from the web
  - Crawling method
    - WaCky initiative (Baroni et al., 2009): gigaword-size corpora for German, Italian, English and French
  - Search engine method
    - (Sharoff, 2006): using combinations of 500 most frequent words
- Basque
  - Tried both for comparing
  - Crawling 250 Mw, search engine 140 Mw

Genre-specific corpora

- Very interesting
  - Discourse analysis
  - ...
- Approaches
  - By crawling
  - By genre filters or classifiers, using character n-grams, punctuation signs, sentence length, POS n-grams... as features (Sharoff, 2006) (Sharoff et al., 2010) (Maekawa et al., 2010)
- Basque
  - In the near future
  - Possible by crawling, at least for certain genres (journalism, blogs, administration...)
  - Which are the appropriate features to develop genre filters or classifiers for an agglutinative language like Basque to be tested

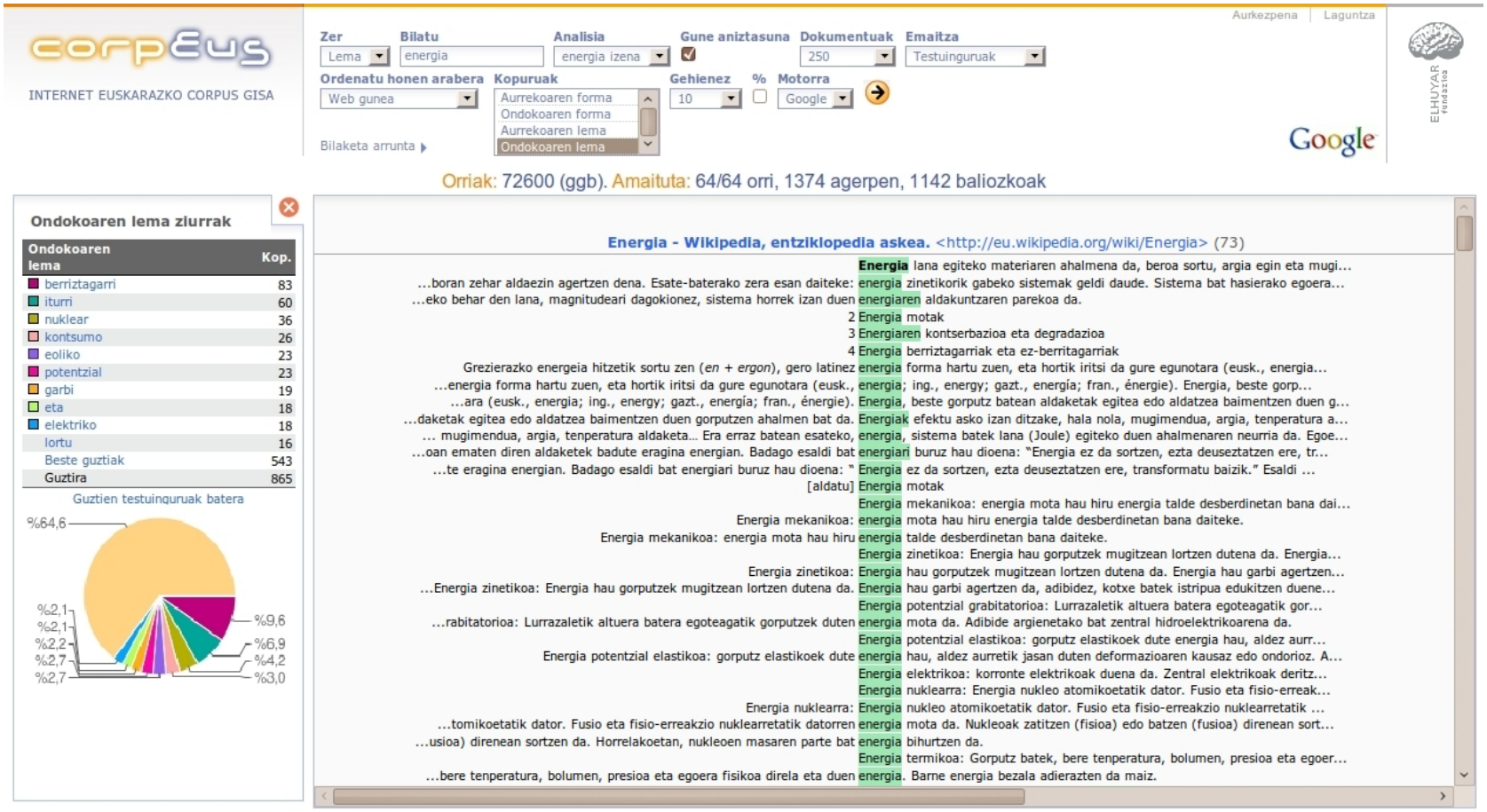
Parallel corpora

- Very interesting
  - Bilingual terminology extraction
  - Machine translation training
- Scarce resource, specially in non-official languages
- But most corporate or public websites with content in a less-resourced language have a version in one or more major languages
- Approaches for building parallel corpora out of multilingual sites (Resnik, 1998)
  - Download pages
  - Document-level alignment
  - Sentence-level alignment
- Basque
  - Just started PaCo2
  - Also interested in multilingual site detection

Web-as-corpus tools

- Web-as-corpus tools
  - Tools that ease the use of the web as a source of linguistic evidence
  - They query APIs of search engines for the words the user enters
  - Show results in a corpus querying tool style: KWiCs, counts, most frequent surrounding words...
- Existing tools
  - WebCorp (Renouf et al., 2006)
  - KWiCFinder (Fletcher, 2006)
- Not appropriate for Basque
- CorpEus, a web-as-corpus tool for Basque (Leturia et al., 2007)
  - Language-filtering words
  - Morphological query expansion
  - http://www.corpeus.org

CorpEus


corpEus
INTERNET EUSKARAZKO CORPUS GISA
Aurkezpena | Laguntza
Zer
Lema
Bilatu
energia
Analisia
energia izena
Gune aniztasuna
Dokumentuak
250
Emaitza
Testuinguruak
Ordenatu honen arabera
Web gunea
Kopuruak
Aurrekoaren forma
Ondokoaren forma
Aurrekoaren lema
Ondokoaren lema
Gehienez
10
%
Motorra
Google
Bilaketa arrunta
Google
ELHUYAR fundazioa
Orriak: 72600 (ggb). Amaituta: 64/64 orri, 1374 agerpen, 1142 baliozkoak
Ondokoaren lema ziurrak
Ondokoaren lema
Kop.
berriztagarri 83
iturri 60
nuklear 36
kontsumo 26
eoliko 23
potentzial 23
garbi 19
eta 18
elektriko 18
lortu 16
Beste guztiak 543
Guztira 865
Guztien testuinguruak batera
%64,6
%9,6
%6,9
%4,2
%3,0
%2,1
%2,1
%2,2
%2,7
%2,7
Energia - Wikipedia, entziklopedia askea. <http://eu.wikipedia.org/wiki/Energia> (73)
Energia lana egiteko materiaren ahalmena da, beroa sortu, argia egin eta mugi...
...boran zehar aldaezin agertzen dena. Esate-baterako zera esan daiteke: energia zinetikorik gabeko sistemak geldi daude. Sistema bat hasierako egoera...
...eko behar den lana, magnitudeari dagokionez, sistema horrek izan duen energiaren aldakuntzaren parekoa da.
2 Energia motak
3 Energiaren kontserbazioa eta degradazioa
4 Energia berriztagarriak eta ez-berritagarriak
Grezierazko energeia hitzetik sortu zen (en + ergon), gero latinez energia forma hartu zuen, eta hortik iritsi da gure egunotara (eusk., energia...
...energia forma hartu zuen, eta hortik iritsi da gure egunotara (eusk., energia; ing., energy; gazt., energía; fran., énergie). Energia, beste gorp...
...ara (eusk., energia; ing., energy; gazt., energía; fran., énergie). Energia, beste gorputz batean aldaketak egitea edo aldatzea baimentzen duen g...
...daketak egitea edo aldatzea baimentzen duen gorputzen ahalmen bat da. Energiak efektu asko izan ditzake, hala nola, mugimendua, argia, tenperatura a...
... mugimendua, argia, tenperatura aldaketa... Era erraz batean esateko, energia, sistema batek lana (Joule) egiteko duen ahalmenaren neurria da. Egoe...
...oan ematen diren aldaketek badute eragina energian. Badago esaldi bat energiari buruz hau dioena: "Energia ez da sortzen, ezta deuseztatzen ere, tr...
...te eragina energian. Badago esaldi bat energiari buruz hau dioena: " Energia ez da sortzen, ezta deuseztatzen ere, transformatu baizik." Esaldi ...
[aldatu] Energia motak
Energia mekanikoa: energia mota hau hiru energia talde desberdinetan bana dai...
Energia mekanikoa: energia mota hau hiru energia talde desberdinetan bana daiteke.
Energia mekanikoa: energia mota hau hiru energia talde desberdinetan bana daiteke.
Energia zinetikoa: Energia hau gorputzek mugitzean lortzen dutena da. Energia...
Energia zinetikoa: Energia hau gorputzek mugitzean lortzen dutena da. Energia hau garbi agertzen...
...Energia zinetikoa: Energia hau gorputzek mugitzean lortzen dutena da. Energia hau garbi agertzen da, adibidez, kotxe batek istripua edukitzen duene...
Energia potentzial grabitatorioa: Lurrazaletik altuera batera egoteagatik gor...
...rabitatorioa: Lurrazaletik altuera batera egoteagatik gorputzek duten energia mota da. Adibide argienetako bat zentral hidroelektrikoarena da.
Energia potentzial elastikoa: gorputz elastikoek dute energia hau, aldez aurr...
Energia potentzial elastikoa: gorputz elastikoek dute energia hau, aldez aurretik jasan duten deformazioaren kausaz edo ondorioz. A...
Energia elektrikoa: korronte elektrikoak duena da. Zentral elektrikoak deritz...
Energia nuklearra: Energia nukleo atomikoetatik dator. Fusio eta fisio-erreak...
Energia nuklearra: Energia nukleo atomikoetatik dator. Fusio eta fisio-erreakzio nuklearretatik ...
...tomikoetatik dator. Fusio eta fisio-erreakzio nuklearretatik datorren energia mota da. Nukleoak zatitzen (fisioa) edo batzen (fusioa) direnean sort...
...usioa) direnean sortzen da. Horrelakoetan, nukleoen masaren parte bat energia bihurtzen da.
Energia termikoa: Gorputz batek, bere tenperatura, bolumen, presioa eta egoer...
...bere tenperatura, bolumen, presioa eta egoera fisikoa direla eta duen energia. Barne energia bezala adierazten da maiz.
Publizitatea | Kontaktua
Copyright © 2007 Elhuyar Fundazioa | Lege-oharra
webmaster@elhuyar.com
W3C XHTML 1.1
W3C CSS 2.0

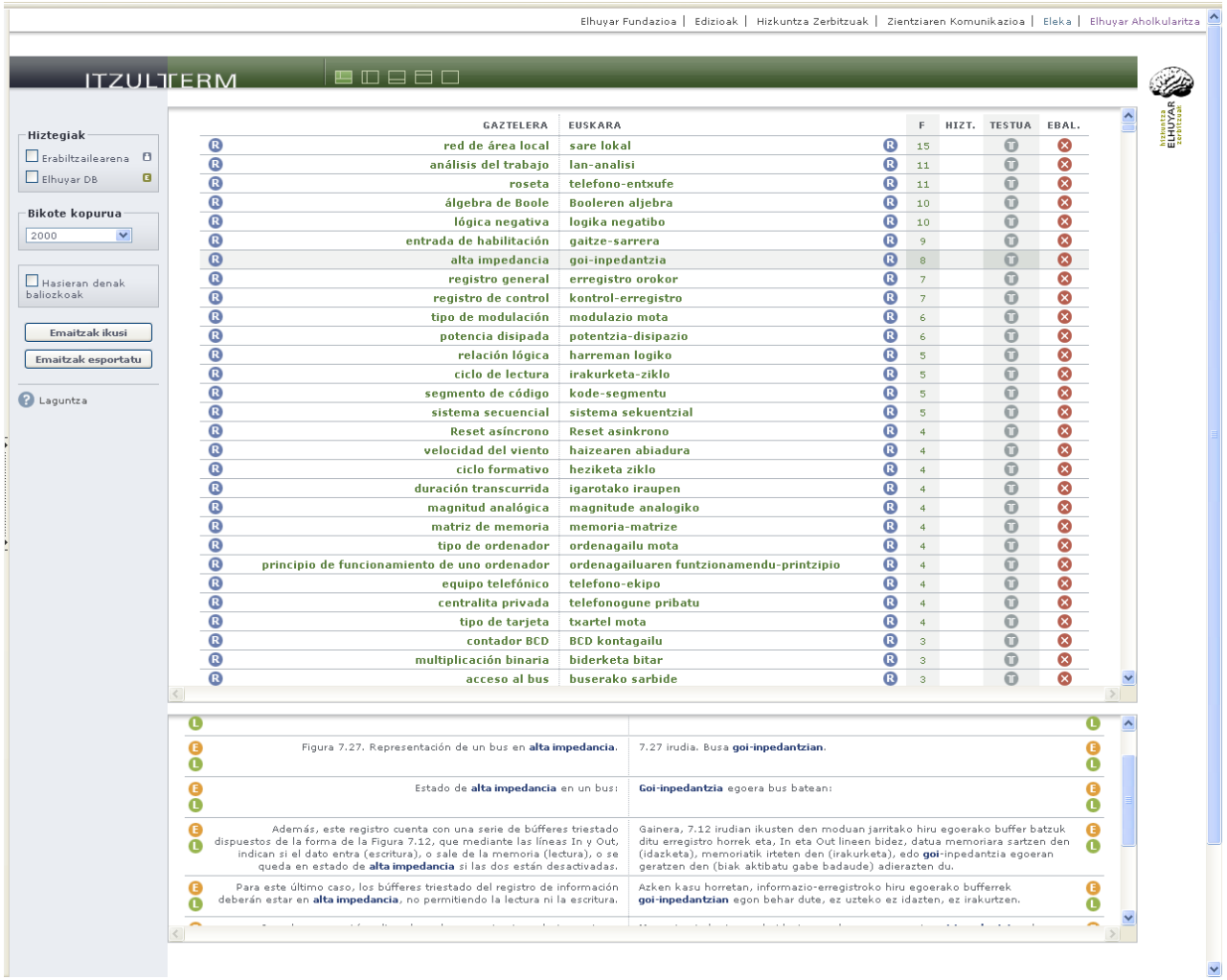
Monolingual terminology extraction

- Erauzterm, monolingual terminology extraction out of monolingual specialized corpora (Alegria et al., 2004)
- Combines linguistical and statistical methods
- Results for the first 2,000 candidates extracted from a corpus on electricity and electronics
  - F measure
    - Multi-word terms: 0.4229
    - Single word terms: 0.4693
  - Precision
    - Multi-word terms: 0.65
    - Single word terms: 0.75

Multilingual terminology extraction I

- ELexBI, equivalent term pairs extraction out of sentence-level aligned parallel corpora (Alegria et al., 2006)
- Based on monolingual term candidate extraction in Basque (Erauzterm) and Spanish (Freeling) and subsequent statistical alignment
- Results for the first 4,000 candidates extracted from a corpus of 10,900 sentences
  - Precision: 0.9

ItzulTerm web service



GAZTELERA	EUSKARA	F	HIZT.	TESTUA	EBAL.
red de área local	sare lokal	15			
análisis del trabajo	lan-analisi	11			
roseta	telefono-entxufe	11			
álgebra de Boole	Booleren aljebra	10			
lógica negativa	logika negatibo	10			
entrada de habilitación	gaitze-sarrera	9			
alta impedancia	goi-inpedantzia	8			
registro general	erregistro orokor	7			
registro de control	kontrol-erregistro	7			
tipo de modulación	modulazio mota	6			
potencia disipada	potentzia-disipazio	6			
relación lógica	harreman logiko	5			
ciclo de lectura	irakurketa-ziklo	5			
segmento de código	kode-segmentu	5			
sistema secuencial	sistema sekuentzial	5			
Reset asíncrono	Reset asinkrono	4			
velocidad del viento	haizearen abiadura	4			
ciclo formativo	heziketa ziklo	4			
duración transcurrida	igarotako iraupen	4			
magnitud analógica	magnitude analogiko	4			
matriz de memoria	memoria-matrize	4			
tipo de ordenador	ordenagailu mota	4			
principio de funcionamiento de uno ordenador	ordenagailuaren funtzionamendu-printzipio	4			
equipo telefónico	telefono-ekipo	4			
centralita privada	telefonogune pribatu	4			
tipo de tarjeta	txartel mota	4			
contador BCD	BCD kontagailu	3			
multiplicación binaria	biderketa bitar	3			
acceso al bus	buserako sarbide	3			

Figura 7.27. Representación de un bus en **alta impedancia**.	7.27 irudia. Busa **goi-inpedantzian**.
Estado de **alta impedancia** en un bus:	**Goi-inpedantzia** egoera bus batean:
Además, este registro cuenta con una serie de búfferes triestado dispuestos de la forma de la Figura 7.12, que mediante las líneas In y Out, indican si el dato entra (escritura), o sale de la memoria (lectura), o se queda en estado de **alta impedancia** si las dos están desactivadas.	Gainera, 7.12 irudian ikusten den moduan jarritako hiru egoerako buffer batzuk ditu erregistro horrek eta, In eta Out lineen bidez, datua memoriara sartzen den (idazketa), memoriatik irteten den (irakurketa), edo **goi**-inpedantzia egoeran geratzen den (biak aktibatu gabe badaude) adierazten du.
Para este último caso, los búfferes triestado del registro de información deberán estar en **alta impedancia**, no permitiendo la lectura ni la escritura.	Azken kasu horretan, informazio-erregistroko hiru egoerako bufferrek **goi-inpedantzian** egon behar dute, ez uzteko ez idazten, ez irakurtzen.

Multilingual terminology extraction II

- AzerHitz, tool to automatically extract pairs of equivalent terms from Basque-English or Basque-Spanish domain-comparable corpora (Saralegi et al., 2008a) (Saralegi et al., 2008b)
- Based on context similarity
- Results
  - Precision
    - Top 1: 0.58
    - Top 5: 0.79

Dictionaries out of the web

- Combination of terminology extraction tools with corpora collection tools provides some semi-automatic ways of building dictionaries out of the web

Web	⇒	AutoCorpEx	⇒	Basque specialized corpora	⇒	Erauzterm	⇒	Basque term lists
Web	⇒	Co3	⇒	English-Basque domain-comparable corpora	⇒	AzerHitz	⇒	Basque-English term lists
Web	⇒	PaCo2	⇒	Spanish-Basque parallel corpora	⇒	Elexbi	⇒	Spanish-English term lists

Ontologies

- WNTERM: a domain ontology of Science and Technology (Pociello et al., 2008)
  - EusWN (IXA) $\leftrightarrow$ *ZTH - Zientzia eta Teknologiaren Hiztegi Entziklopedikoa* or "Encyclopaedic Dictionary of Science and Technology" (Elhuyar)
- Future project
  - Goal
    - Automatically (or semi-automatically) enrich existing concept taxonomies such as WordNet
    - Build domain-specific ontologies
  - The specialized corpora to be used in this project can also be collected automatically out of the web

Experiments

- Monolingual specialized web corpora (Gurrutxaga et al., 2009)
- Multilingual domain-comparable web corpora (unpublished)

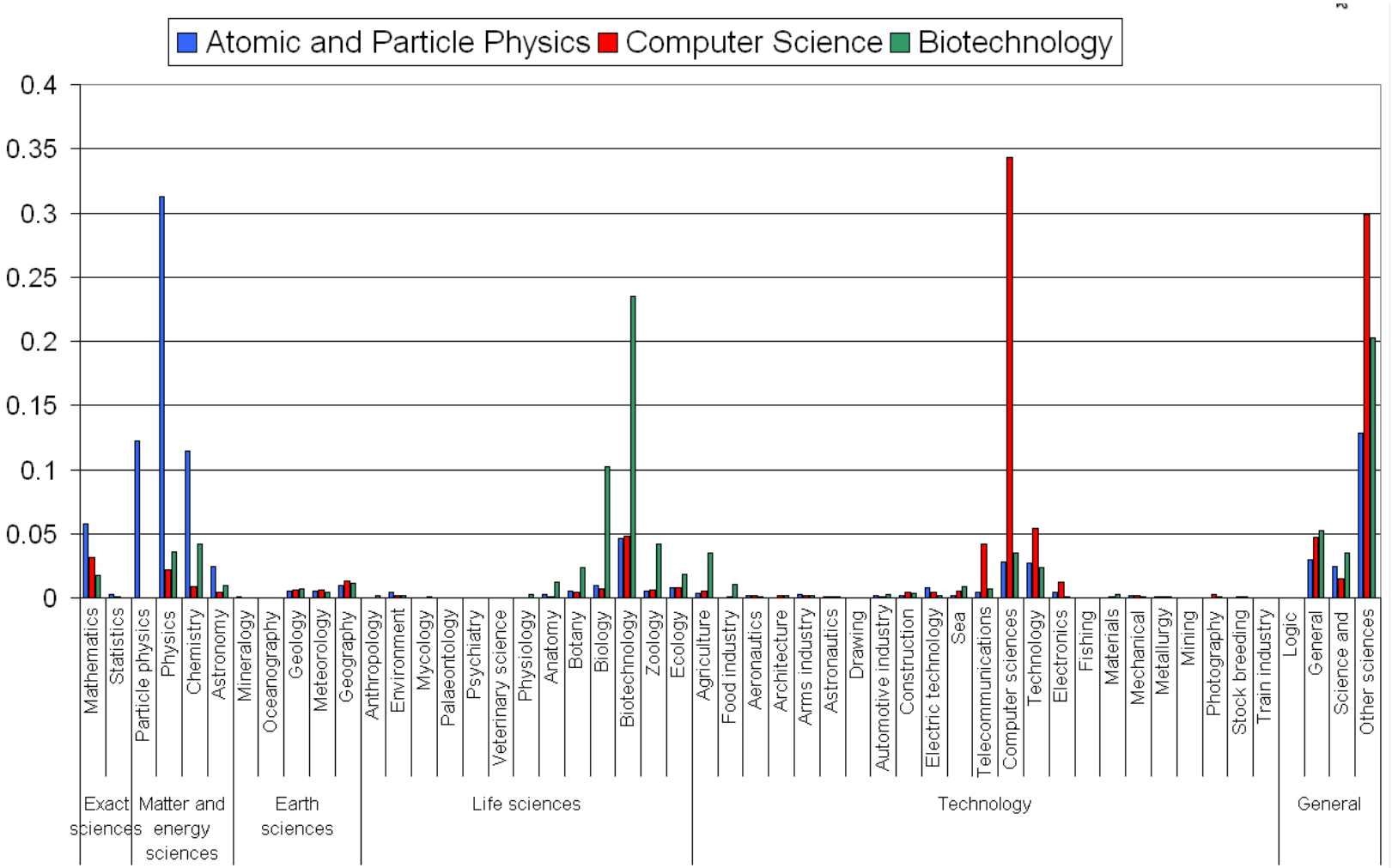
Goal and design

- Goal: to evaluate the domain precision of web corpora built with Co3 by means of automatic term extraction and dictionary based validation
- Design
  - Collected three specialized corpora with AutoCorpEx: Atomic & Particle Physics, Computer Science and Biotechnology
  - Applied Erauzterm's terminology extraction to them and obtained three term lists
  - Lists automatically evaluated against
    - A recently built specialized dictionary (2009), ZTH (http://zthiztegia.elhuyar.org/): 50 domains, 23.000 concepts
    - Online version of *Euskalterm*, the Basque Public Term Bank: 22 domains, 91.000 registers
  - Those terms from the first 1,500 candidates not found in those sources were manually evaluated by experts

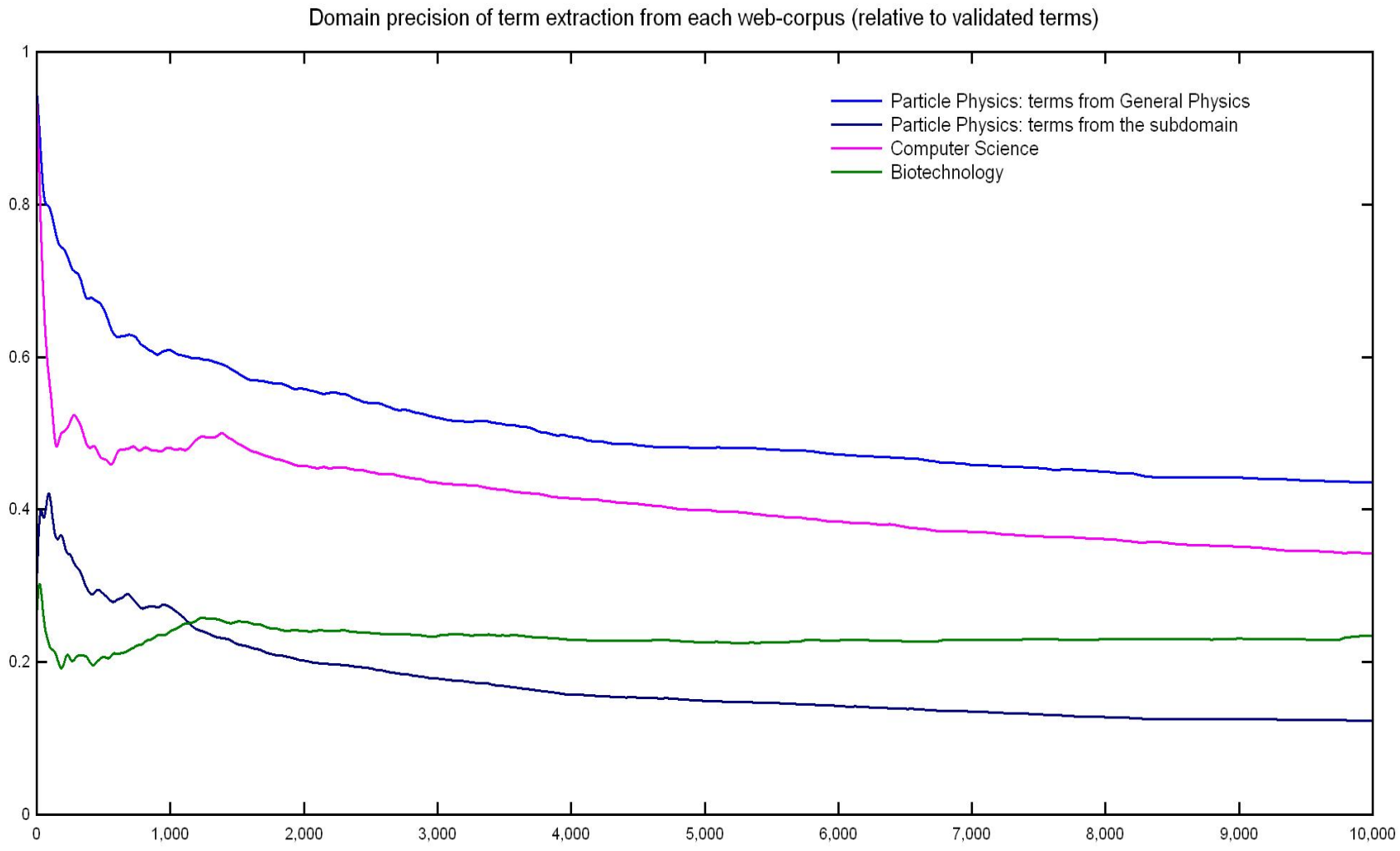
Corpus and term list sizes obtained

Corpus	**Atomic and Particle Physics**	**Computer Science**	**Biotechnology**
Sample corpus size	32 docs, 26,164 words	33 docs, 34,266 words	55 docs, 41,496 words
Seed words	63	60	105
Obtained corpus size	48 domains 310 pages 320,212 words	485 domains 1,810 pages 2,514,290 words	68 domains 358 pages 578,866 words
Extracted term list size	46,972	163,698	34,910
Dictionary validated	6,432	8,137	6,524
- First 10,000 candidates	2,827	2,755	2,403
Manually evaluated	869	904	628
- Terms	628	512	432
- Not terms	241	392	196

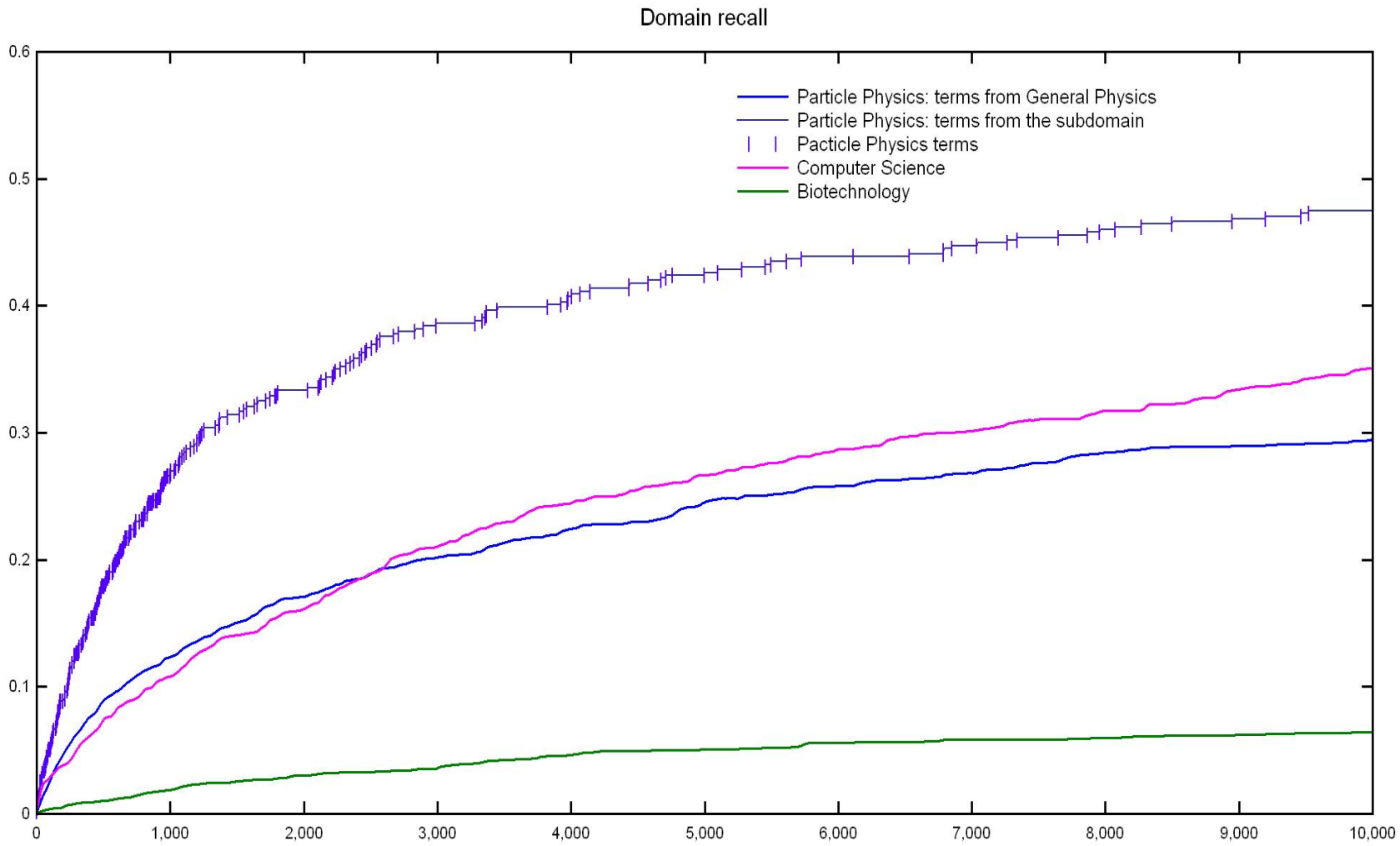
Distribution of the term lists



Domain precision of the term lists



Recall relative to the specialized dictionary



Terms not extracted from the web corpora and included in ZTH

- Atomic and Particle Physics - ZTH: 474
  - Not extracted in the experiment: 150 (31.64%)
    - Not found in the Internet (42)
    - Not in the web corpus (4)
    - Not extracted by Erauzterm (104) / $f = 1$ (101)

Distribution of manually validated terms

Atomic and Particle Physics		**Computer Science**		**Biotechnology**	
Physics	377	Computer Science	348	Biotechnology	146
Atomic and Particle Physics	109	**General**	**112**	Biology	99
Chemistry	56	Telecommunications	22	**General**	**92**
Others	86	Others	30	Others	95
Total	628	Total	512	Total	432

Some conclusions

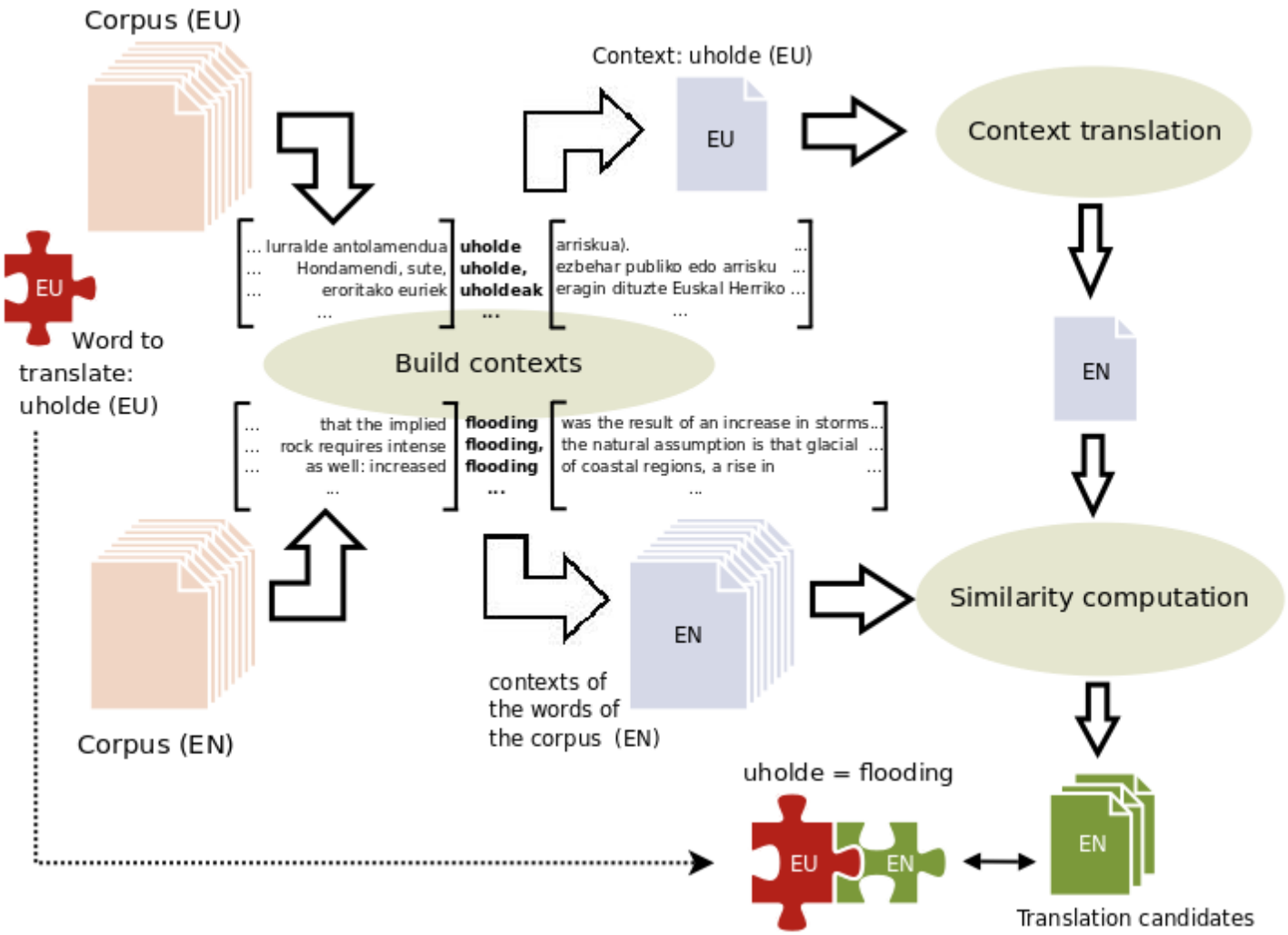
- Corpora built are specialized
- New opportunities for terminological work (lexical observatory, dictionary enrichement...)
- Recall problems
  - Text scarceness in some specialized domains
  - Hidden web?
  - Effect of seed documents selection
  - Term extractor

  $\rightarrow$ Web corpora can not be used as a the sole source of information
- Results differ depending on the corpus
  - Science better than technology?
  - Broadness and complexity of the domain
- Effect of corpus size?

Goal and design

- Goal: to evaluate the improvement in AzerHitz (bilingual terminology extraction out of comparable corpora based on context similarity) by enhancing contexts using the web
- Problem: most words have not enough context information in the corpus
- Design
  - Collected a Basque-English domain-comparable corpus of Computer Science using Co3 (2.6 Mw in each language)
  - Applied AzerHitz to extract translations of a 100 Basque words' set
  - Applied AzerHitz but enhancing the contexts of the source words and their translation candidates via web searches.

How AzerHitz works



Improvement

- Improvement: expand the contexts of the words using web concordancers
  - CorpEus for Basque (Leturia et al., 2007)
  - WebCorp for English (Renouf et al., 2006)
- Only interested in one sense of the word, including contexts of other senses of the word adds noisy data $\rightarrow$ Add words characteristic of the domain to the query

Results

Setup	top1	top5	top10	top15	top20
Baseline	0.32	0.54	0.60	0.62	0.66
WaC	0.36	0.56	0.68	0.72	0.72
Baseline + cognates	0.54	0.62	0.62	0.64	0.66
WaC + cognates	0.58	0.66	0.70	0.72	0.72

Conclusions I

- Common problem of less resourced languages: economic resources devoted to the development of NLP tools are also scarce
- The use of the Internet for building language resources such as corpora and, through them, other resources and NLP tools, is very attractive indeed
- The hypothesis of its valuability and profitability as a source for developing language resources for less resourced languages must be tested
- Any attempt to build web corpora in a given language is conditioned by the size of the web in the target domains or genres

Conclusions II

- The results of the experiments we have carried out for Basque are very encouraging
  - The size of the specialized web corpora compiled
  - The domain-precision achieved
  - The fact that the use of web-derived contexts improves the results of terminology extraction from domain-comparable corpora
- Not in any way comparable with the size of the webs of major languages
- For the time being, some domains and genres may not have enough representation in the web
- Elhuyar Foundation will go on working with the web as a source of corpora of many kinds and other types of language resources for Basque

THANK YOU FOR YOUR ATTENTION!!!

References I

I. Alegria, A. Gurrutxaga, P. Lizaso, X. Saralegi, S. Ugartetxea, and R. Urizar. 2004. A xml-based term extraction tool for basque. In *Proceedings of LREC 2004*, pages 1733–1736, Lisbon, Portugal. ELRA.

I. Alegria, A. Gurrutxaga, X. Saralegi, and S. Ugartetxea. 2006. Elexbi, a basic tool for bilingual term extraction from spanish-basque parallel corpora. In *Proceedings of Euralex 2006*, pages 159–165, Torino, Italy. Euralex.

M. Baroni and S. Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*, pages 1313–1316, Lisbon, Portugal. ELRA.

References II

M. Baroni, S. Bernardini, A. Ferraresi, and Zanchetta E. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation Journal*, (43):209–226.

A.Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings of Compression and Complexity of Sequences 1997*, pages 21–29, Los Alamitos, USA. IEEE Computer Society.

A.Z. Broder. 2000. Identifying and filtering near-duplicate documents. In *Proceedings of Combinatorial Pattern Matching*, pages 1–10, Montreal, Canada. Springer.

References III

W. H. Fletcher, 2004. *Corpus Linguistics in North America 2002*, chapter Making the web more useful as a source for linguistic corpora. Rodopi, Amsterdam, The Netherlands.

W. H. Fletcher, 2006. *Corpus Linguistics and the Web*, chapter Concordancing the Web: Promise and Problems, Tools and Techniques, pages 25–46. Rodopi, Amsterdam, The Netherlands.

A. Gurrutxaga, I. Leturia, E. Pociello, X. Saralegi, and I. San Vicente. 2009. Evaluation of an automatic process for specialized web corpora collection and term extraction for basque. In *Proceedings of eLexicography in the 21st century*, Louvain-la-Neuve, Belgium. EURALEX & SIGLEX.

References IV

I. Leturia, A. Gurrutxaga, I. Alegria, and A. Ezeiza. 2007. Corpeus, a 'web as corpus' tool designed for the agglutinative nature of basque. In *Proceedings of the 3rd Web as Corpus workshop*, pages 69–81, Louvain-la-Neuve, Belgium. ACL SIGWAC, Presses universitaires de Louvain.

I. Leturia, A. Gurrutxaga, N. Areta, and E. Pociello. 2008a. Analysis and performance of morphological query expansion and language-filtering words on basque web searching. In *Proceedings of LREC 2008*, Marrakech, Morocco. ELRA.

References V

I. Leturia, I. San Vicente, X. Saralegi, and M. Lopez de Lacalle. 2008b. Collecting basque specialized corpora from the web: language-specific performance tweaks and improving topic precision. In *Proceedings of the 4th Web as Corpus Workshop*, pages 40–46, Marrakech, Morocco. ACL SIGWAC.

I. Leturia, I. San Vicente, and X. Saralegi. 2009. Search engine based approaches for collecting domain-specific basque-english comparable corpora from the internet. In *Proceedings of 5th International Web as Corpus Workshop (WAC5)*, pages 53–61, Donostia, Spain. ACL SIGWAC.

References VI

K. Maekawa, M. Yamazaki, T. Maruyama, M. Yamaguchi, H. Ogura, W. Kashino, T. Ogiso, H. Koiso, and Y. Deu. 2010. Design, compilation and preliminary analyses of balanced corpus of contemporary written japanese. In *Proceedings of LREC 2010*, Valletta, Malta. ELRA.

E. Pociello, A. Gurrutxaga, E. Agirre, I. Aldezabal, and G. Rigau. 2008. Wnterm: Combining the basque wordnet and a terminological dictionary. In *Proceedings of LREC 2008*, Marrakech, Morocco. ELRA.

A. Renouf, A. Kehoe, and J. Banerjee, 2006. *Corpus Linguistics and the Web*, chapter WebCorp: an Integrated System for WebText Search, pages 47–67. Rodopi, Amsterdam, The Netherlands.

References VII

P. Resnik. 1998. Parallel strands: A preliminary investigation into mining the web for bilingual text. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, pages 72–82, Langhorne, USA. AMTA.

X. Saralegi and I. Alegria. 2007. Similitud entre documentos multilingües de carácter científico-técnico en un entorno web. *Procesamiento del Lenguaje Natural*, (39):71–78.

X. Saralegi and I. Leturia. 2007. Kimatu, a tool for cleaning non-content text parts from html docs. In *Proceedings of the 3rd Web as Corpus workshop*, pages 163–167, Louvain-la-Neuve, Belgium. ACL SIGWAC, Presses universitaires de Louvain.

References VIII

X. Saralegi, I. San Vicente, and A. Gurrutxaga. 2008a. Automatic extraction of bilingual terms from comparable corpora in a popular science domain. In *Proceedings of Building and using Comparable Corpora workshop*, Marrakech, Morocco. ELRA.

X. Saralegi, I. San Vicente, and M. López de Lacalle. 2008b. Mining term translations from domain restricted comparable corpora. *Procesamiento del Lenguaje Natural*, (41):273–280.

S. Sharoff, Z. Wu, and K. Markert. 2010. The web library of babel: evaluating genre collections. In *Proceedings of LREC 2010*, Valletta, Malta. ELRA.

S. Sharoff, 2006. *WaCky! Working Papers on the Web as Corpus*, chapter Creating General-Purpose Corpora Using Automated Search Engine Queries, pages 63–98. Gedit Edizioni, Bologna, Italy.

Exploiting the Internet to build language resources for less-resourced languages

Antton Gurrutxaga Igor Leturia Eli Pociello
Iñaki San Vicente Xabier Saralegi

Elhuyar Foundation
R & D

Creation and use of basic lexical resources for less-resourced languages
7th SaLTMiL Workshop on
LREC 2010, Valetta, Malta
2010/05/23