V Congreso Internacional de Lingüística de Corpus
CILC 2013
5th International Conference on Corpus Linguistics

Bilingual Lexicography: Relevance of Corpora

The example of German-Basque as language pair

David Lindemann
PhD cand.
UPV-EHU University of the Basque Country
david.lindemann.soraluze@gmail.com

EUSKAL HIZKUNTZA ETA KOMUNIKAZIOA SAILA
UPV/EHU | www.ehk.ehu.es

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

Motivation for this paper

The Problem

Translation Studies, Philology, Lexicography vs. Computational Linguistics

- Experts on both sides, lack of comprehensive discipline-linking publications with focus on practical issues

Low or Medium Density Languages and more or less 'exotic' Language Pairs

- DE-EN is not DE-EU is not FI-EU

The Intention

Useful information for lexicographers that start from scratch

- as computational linguist
- as corpus linguist
- with no or little work done on their language pair

The Example

- German-Basque

Overview

- 1. Lexicography and Corpus Linguistics
- 2. Bilingual Lexicography and Corpus Linguistics
  - 2.1 Corpus based lemma list retrieval (German)
    - 2.1.1 Comparing corpus-based vs. editorial-only lemmalist
  - 2.2 Corpora and Corpus based lemma list retrieval (Basque)
  - 2.3 Usage examples from monolingual corpora
  - 2.4 Translation examples from parallel corpora
- 3. New DE-EU dictionary: Editing, Corpora, Tools
  - 3.1 New DE-EU dictionary: frontend design
  - 3.2 New DE-EU Literary Corpus
  - 3.3 Verb entry
  - 3.4 Noun entry
  - 3.5 Adjective and Adverb
- 4. Automatic pairing of Translation Equivalents or SADD
  - Corpus-based methods, Wikimedia, WordNet
- 5. Conclusions

1. Lexicography and Corpus Linguistics

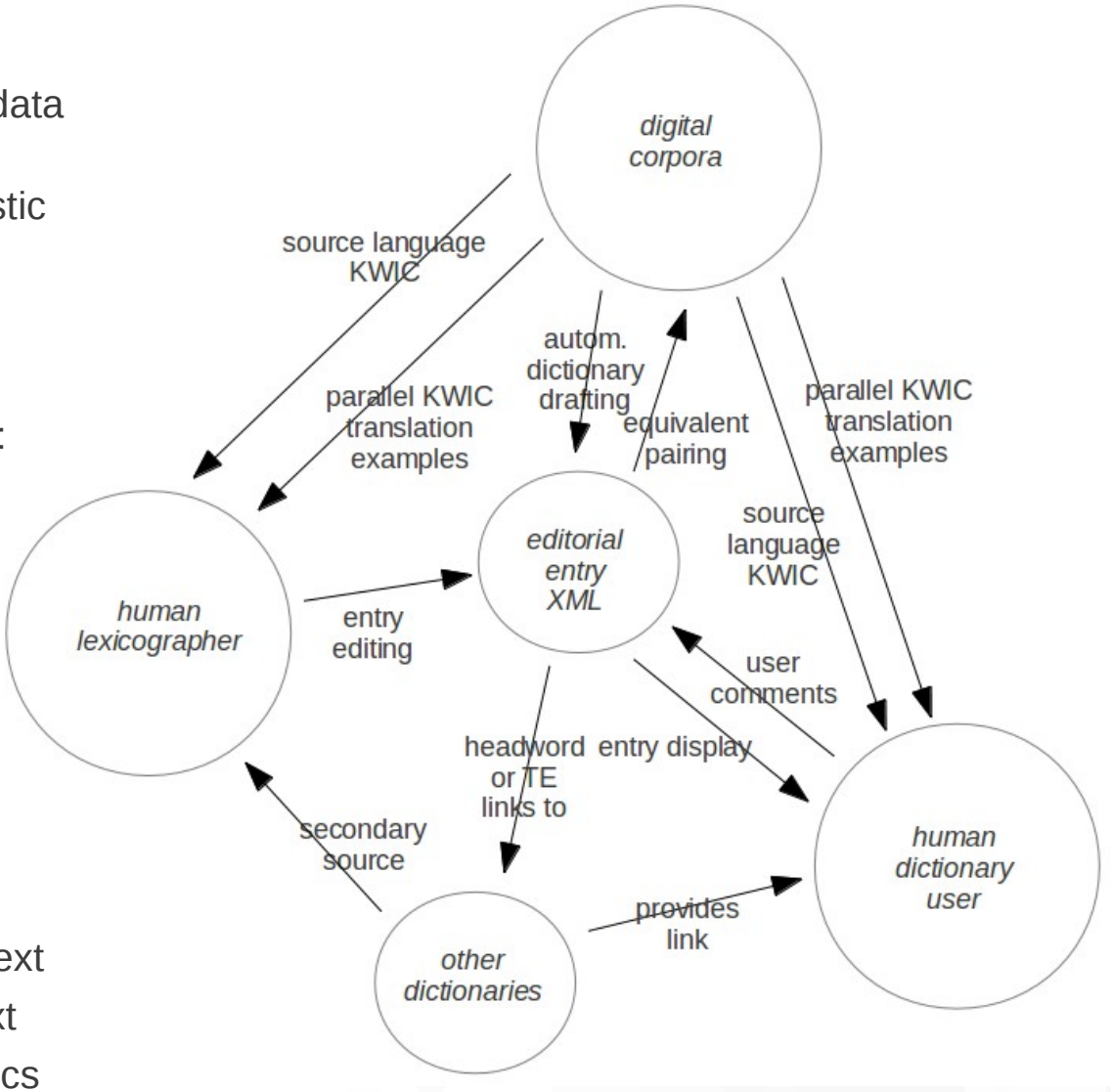
- Making dictionaries from corpora
  - Introductions: Atkins & Rundell 2008: 53-96; Svensén 2009:43-58
  - “Corpus Linguistics and Lexicography” (Teubert 1999, 2002, 2007)
  - “Corpus Lexicography” (Kilgarriff & Tugwell 2002, Kilgarriff 2012)
  - “Korpusgestützte Lexikographie” (Klosa 2007)
  - Quantity: Large electronic corpora, much larger than in pre-computer age
  - Quality: Human made index cards vs. e-corpora: The rare rather than the common
  - Automatic retrieval of frequency lists / lemma lists
  - Lexicographers' Documentation process when editing entries
    - Concordances (KWIC)
    - Coocurrence / Collocation Statistics / Multiword LU
    - 'Wordsketch' incl. grammar (eg for verbs: Subjects, objects, prepos., conjunct.)
  - “Corpus-based” vs. “corpus-driven” lexicography: Trust Corpora for WSD?
  - Evidence from parallel corpora: TE proposed by biling. dics vs. translators' choice
- Data from large corpora in electronic dictionary publishing
  - On-the-fly KWIC generation as part of dictionary search result page
    - as shown on some (still not many) dictionary websites

2. Bilingual Lexicography and Corpus Linguistics

- Monolingual and Bilingual Lexicography:
  - Corpus-based lemma-list retrieval (frequency lists)
    - Frequency list becomes lemma list combining automated and hand-craft methods (IDS 2009)
    - Corpus frequency = look-up frequency (in the top few thousand) (De Schryver 2010)
  - Usage examples from monolingual corpora
    - Commonplace in Lexicography: Useful for dictionary entry editing
    - Corpus Concordances shown on monolingual dictionary websites (e.g. http://www.dwds.de)
- Bilingual Lexicography:
  - Automatic pairing of Translation Equivalents
    - Different strategies and software tools (examples Héja 2010, Nazar 2012)
  - Translation examples from parallel corpora
    - Useful for bilingual entry editing
  - Parallel KWIC as part of the dictionary search-result page
    - Teubert, *op. cit.;* first mention in: Atkins 1996, Dickens & Salkie 1996.
    - Recent, maybe leading example: http://www.linguee.com/ (no German-Basque)
    - Mainly based on software localization files: http://en.glosbe.com/de/eu (also German-Basque)

Corpora in bilingual Lexicography

- Lexicographer or Dictionary editor
  - “reasoned condensation” of linguistic data provided by computers
  - “introspection”, use of their own linguistic competence
- Digital Corpora
  - monolingual: automatically built
  - parallel corpus building and alignment:
    - automatically
      - statistical methods
      - starting from (reliable) bilingual lexicon
    - semi-automatic or by hand
- Dictionary user
  - condensed in one search result:
    - editorial entry
    - headword in source language context
    - headword and TE in parallel context
    - headword or sense links to other dics



Density: Availability of digitally stored material (Varga et al. 2005)

- Approximation indicators
  - Number of speakers of a language
  - Web size of a language

	German	Basque
Speakers	98 million	0,8 million
Biggest Corpus	5,4 billion	0,12 billion
Wikipedia Articles	1,6 million	0,15 million
ELRA Products	444	6

- Parallel Corpus Building methods
  - Web mining
    - Fine for monolingual corpora even for low density languages
    - For parallel corpora more difficult:

      “When the other language is much lower density, the actual number of automatically detectable parallel pages is considerably smaller”
  - Other sources
    - Literary text, religious text, international law, movie dubs/subtitles, software internationalization, bilingual magazines, annual reports, corporate home pages
    - Apply to German-Basque / Don't apply
    - UPV-EHU: German-Basque literature corpus (81 books, 2 million tokens)

2.1 Corpus based lemma list retrieval (DE)

- German: DeReWo (IDS 2009)
  - Based on DeReKo (5,4 bill. words)
  - DeReWo 40.000: Lemmatized
  - Comp. w. DUDEN (2011 [240.000])

  First sample: abc-first 2000

  61 sorted out

  23 not found in DUDEN

  (29 found in another form)

  20 female human gender

  15 proper names (orgs, geogr.)

  4 other cases

  35 added

  Homonyms: 8

  Infinitives (corr. Part. or *-ung*): 27

  2000 - 61 + 35 = 1974 (98,7%)

  76 added acc. to lexicographer's subjective relevance criteria

  2000 - 61 + 35 + 76 = 2050 (102,5%)

Most freq.	Least freq.	abc-first	abc-last
der/die/das	Kindheitstraum	á	Zwischenspiel
und	Gulaschsuppe	à	Zwischenspurt
sein	verschreibungspflichtig	Aal	zwischenstaatlich
in	Superminister	Aar	Zwischenstand
ein	Seniorentreffen	ab	Zwischenstopp
werden	Schlampe	abändern	Zwischenzeit
von	Rinnsal	abarbeiten	zwischenzeitlich
mit	Kronprinzessin	Abbau	Zwist
haben	Hockeyspieler	abbauen	zwitschern
zu	Heimaufsicht	abbekommen	zwölf
für	Führungsmitglied	Abberufung	Zwölf
er/sie/es/sie	Wegweisung	Abbestellung	Zwölfer
auf	Umbaukosten	abbezahlen	zwölfjährig
nicht	Straßenbahnhaltestelle	abbiegen	zwölfköpfig
eine	Stahlrohr	Abbild	zwölfmal

DeReWo-40.000	EAH-2007	LDS	DeReWo Pos.	DeReWo-Freq	Leipzig Freq.
Befehl	Befehl	Befehl	5926	13	12
befehlen	befehlen	befehlen	10262	14	16
	Befehlsform	Befehlsform			21
	Befehlsgewalt				16
Befehlshaber	Befehlshaber	Befehlshaber	18747	16	14
		Befehlsverweigerung			16
befestigen	befestigen	befestigen	8204	14	15
befestigt			13643	15	13
Befestigung	Befestigung	Befestigung	21221	16	15
	befeuchten	befeuchten			18
befeuern			33303	17	16
befinden	befinden	befinden	569	9	10
Befinden	Befinden	Befinden	24672	16	15
befindlich		befindlich	5499	13	19
Befindlichkeit		Befindlichkeit	13423	15	15
		befingern			21
		beflaggen			21
	beflecken	beflecken			18
beflissen			39294	17	17
beflügeln		beflügeln	8895	14	13
beflügelt			25788	16	13
		befohlen			14
befolgen	befolgen	befolgen	12262	15	14
	Befolgung				17
	Beförderer				20
befördern	befördern	befördern	4647	13	12
befördert			12377	15	12
Beförderung	Beförderung	Beförderung	9731	14	13
	Beförderungsmittel	Beförderungsmittel			19
		befrachten			20
befragen	befragen	befragen	2829	12	12
befragt			14584	15	11
Befragte			4367	13	14
Befragung	Befragung	Befragung	5226	13	11
befreien	befreien	befreien	2234	11	11
Befreier	Befreier	Befreier	30667	17	15
befreit			8821	14	11
Befreiung	Befreiung	Befreiung	4917	13	11
Befreiungsarmee			23679	16	16
Befreiungsbewegung		Befreiungsbewegun	30167	17	16
		Befreiungskampf			17
Befreiungsschlag			14791	15	13
Befreiungstiger			33275	17	14
	Befreiungsweg				21
befremden		befremden	15373	15	15
		Befremden			15
befremdlich		befremdlich	19219	16	14
	befreunden	befreunden			19
befreundet		befreundet	5158	13	13

2.1.1 Comparing corpus-based list with dictionary lemmalists

- Not on freq.list but freq. (according to other sources)
- Not on freq.list, less freq.
- Not on freq.list, not freq.
- On freq.list, not on DUDEN lemmalist, to be erased (proper names)
- Gaps from a human lexicographer's point of view

...on a freq.list sample of 33:

- 2 to be erased
- 3 frequent to be included
- 4 less freq. to be included

German Lemma List: Problems (1)

- Schnorr 1991
- Lötzsch 1991

- Partizipien
  - Partizip I mit und ohne adjektivischer Bedeutung
    - Eigenes Lemma, wenn im Duden als "Adjektiv": *alleinerziehend*
    - Infinitiv auf Derewo, Part I. nicht, aber Part. I im Duden, Verbinfinitiv nicht: Geänderte Aufnahme (*andersdenken>andersdenkend*)
  - Partizip II mit und ohne adjektivischer Bedeutung
    - Wie im Duden als "Adjektiv" und/oder "Partizip"
    - Link vom Partizip zum Verbinfinitiv (der bei Fehlen aufgenommen wird)
- Weibliches Human Gender
  - Zusammen mit der männlichen Form als Lemma-Nebenform
- "Transparente Komposita" (Mehrwortsyntagmen)
  - *Abschlusstabelle*, *Abschleppwagen*, *Holzhaus, Küchentisch*
    - Wenn auf DEREWO und im Duden-230.000, dann als Lemma aufnehmen (auch hier: Frequenz als Kriterium)
    - "Postulat der optimalen Platzverwendung" gilt nicht
    - Erscheinen im WB für DE L2 Produktion sinnvoll, über EU-Register suchbar.

German Lemma List: Problems (2)

- Substantiv-Nennform immer Singular?
  - (Quasi-)Plurale Tantum
    - Aufgenommen wie im Duden, DEREWO-Form bei Bedarf geändert
      - *Abbrucharbeiten, Achtzigerjahre, Arbeitsbedingungen*
- Nicht auf DEREWO enthaltene Lemmata
  - Gewonnen aus Abgleich mit den Lemmastrecken der drei Sekundärquellen
    - *Aas, Aasfresser, abbinden, abbüßen, Abendrot, Abendschule, abgefeimt, abgekartet, abgepackt, abgeschmackt, abgesehen, abgespannt, abgestanden, abgetakelt, abgetreten, abgewetzt, abgezehrt, abgießen, Abguss, abhärten, abkacken, abkehren, Abklatsch, Abkömmling, abküssen, ablecken, abmessen, Abrede, abreiben, Abreibung, Abreise, Abriss, abrüsten, Abschleppwagen, Absolutismus, Absonderlichkeit, absondern, Abspann, abspülen, abstillen, abstinent, abstumpfen, abteilen, abwertend, abwracken, achtsam, Adoptivbruder, Adoptivkind, Adoptivschwester, Adoptivsohn, Adoptivtochter, Afroamerikaner, Ähre, Akkordarbeit, Aktiv, Alabaster, Albaner, Alias, allwissend, altklug, am, amen, Amphibie, Andalusier, andersartig, anfeuchten, anflehen, Angeberei, angeberisch, Angemessenheit, ankern, Anschnallgurt, anspitzen, Antiquitätenhändler*

German Lemma List: Problems (3)

- Nicht übernommene DEREWO-Einträge:
  - Nicht im DUDEN: *á, abend, Abfahrts, abgründen, ablaufend, aborigines, Abstiegsrunde, Abstimmungskampf, abwartend, Abwässer, accessoires, achtelfinal, Achtelfinals, achtzehnt, Acrylbild, Adventkonzert, Adventsbasar, adventsingen, afro, Aktienkursus, alleinbleiben, allround, Allstar, allzuoft, allzusehr, allzuviel, Alpenliga, Alpenstraße, Alpin, Altenheimen, Alterssiedlung, altertum, Altertümer, älterwerden, Altstadtfest, Altstadtfreund, Ampelregelung, Anders, Anerkennungspreis, Angeln, anglo, Angriffsbemühung, anlaufend, Anwaltskosten, Arbeits, Arbeitsamtsbezirk, Arbeitsjahr, Arbeitsmarktdatum, Arbeitsmarktexperte, Arbeitsmarktlage, Arbeitspartei, Arbeitsschwerpunkt*
  - Spez. Eigennamen: *Achensee, Ackermann, Akademietheater, Alaska, Alb, Allen, Alster, Am, Amnesty, AOK, Aphrodite, Apostelkirche, Appenzell, Appenzeller, Apulien*

German Lemma List: Prediction on [2000] sample

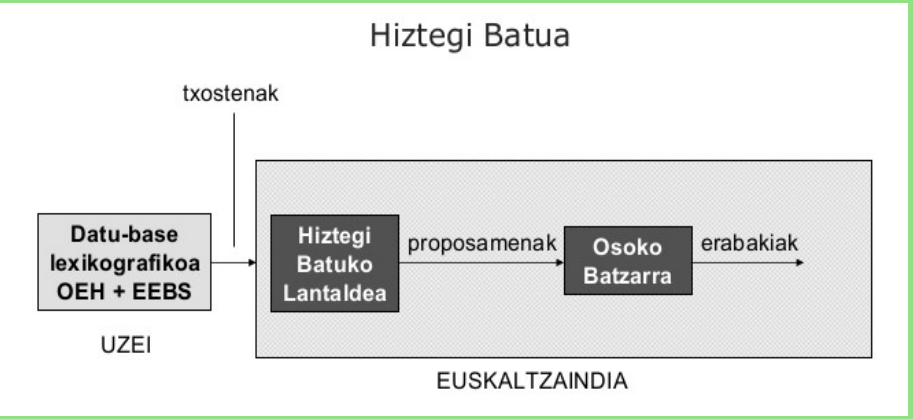
- German Lemmata:
  - [DEREWO-40.000] ∩ [DUDEN-230.000] ≈ 38.700
  - + 5-6% from Secondary Sources
  - results in approx. 42.000 Lemma signs

- Basque Lemma Sources (future)
  - *Hiztegi Batua* 2011 (52.615 lemmata)
    - freely available XML (GPL)
  - *Basque WordNet* (26.727 lemmata)
    - ELRA resource (not free)
  - Elhuyar/EHU Science & Technology Corpus frequency list (78.600 lemmata)

2.2 Corpora and Corpus based lemma list retrieval (EU)

Euskaltzaindia (Language Academy): Lexicography

- OEH Corpus: All (304) printed books (religious and secular literature) until 1970: 5 mill. words
- UZEI 20th century corpus: 4,6 mill. words
- *Automatic filtering of lemma candidates, final decision one-by-one by editorial board*
- *Hiztegi Batua: 52.615 entries*

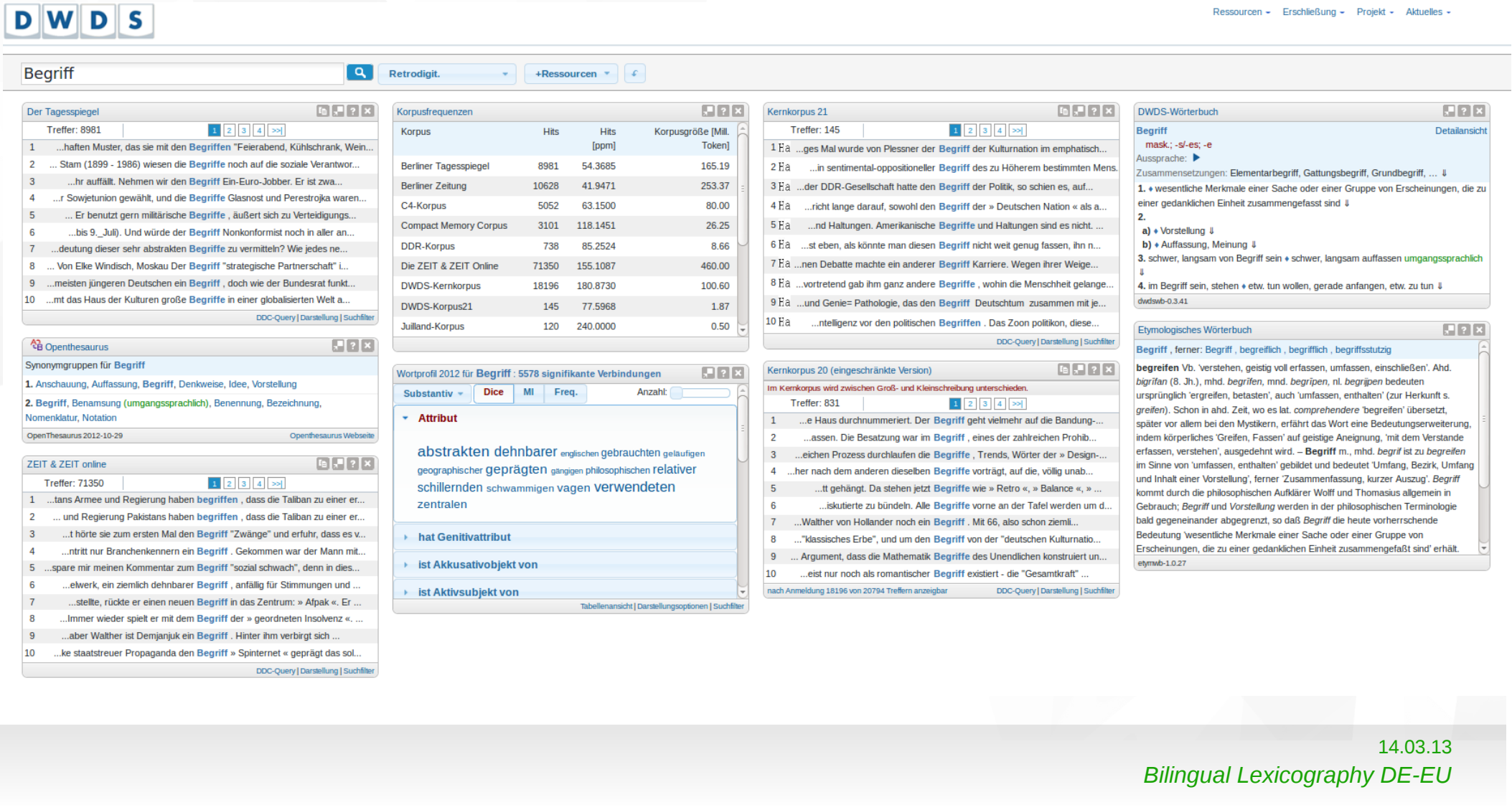


UPV-EHU/Elhuyar

- Science and Technology Corpus
  - 1990-2002 scientific pub.
  - 8,5 mill. words
    - filtered lemmalist: 78.600 words
- EPG Contemporary Prose Corpus
  - from 2000 onwards
  - 13,1 mill. words from 287 books
  - 12 mill. words from press
    - frequency lists or lemmalists?
- XUXEN spell-checker/EDBL
  - 120.000 lemmata, flexion-morphemes and irregular word forms

2.3 Usage examples from monolingual corpora

- German DWDS monolingual: Entries in several editorial dictionaries, wordsketch, KWIC of several corpora, hit statistics. Appearance fully customisable.
- For German, leading in dictionary publishing



2.4 Translation examples from parallel corpora

„Begriff“	
kontzeptu	18
ideia	11
hitza	2
adigai	1
burutapena	1
ezagutza	1
gai	1
hitz	1
ikusmolde	1
pentsakera	1
termino	1
ulerbide	1
ulerkera	1

„im Begriff sein“			
-tzera joan	5	*-tzekotan izan*	1
-tzeko zorian	4	*-tzera apoderatzen izan*	1
-tzear egon	3	*ekarri*	1
-tzeko asmoa izan	3	*gutxi falta + subjkt.*	1
-tzeko asmotan egon	2	*hasia izan*	1
-tzea pentsatu	1	*inf. nahi izan*	1
-tzear izan	1	*inf. nahian*	1
-tzeko duda egin	1	*egin behar izan*	1

LU and TE candidates in Parallel Corpus (SkE). Example: “Begriff” (noun)

kein Begriff sein	*ez ezagutu*	1
kein Begriff sein	*horren entzuterik ere ez izan*	1
keinen andern Begriff haben als ...	*... baino ez pentsatu*	1
nicht allzu schnell von Begriff sein	*ez oso azkarra izan*	1
schwer von Begriff sein	*burugogorra izan*	1
sich einen stillen Begriff machen	*gutxi gora behera irudika ahal izan*	1
sich kaum einen Begriff machen	*ozta-ozta ideia bat izan*	1
sich keinen Begriff machen	*ez jakin*	1
sich keinen Begriff machen	*ezin imajinatu ere egin*	1
über alle Begriffe	*ezin esan bezain*	1
Der Begriff X	*X*	1
einen Begriff geben	*aditzera eman*	1

Parallel KWIC from lemmatized and POS-tagged DE-EU corpus using SkE. Example: “Begriff” (noun)

Corpus: **DE_1**
Hits: **86** (37.3 per million) [EU_1] [EU_2]
Page 1 of 5 Go Next | Last

#95733	war frech , davon machen Sie sich keinen **Begriff** . Denken Sie sich , ich hab meinen Überzieher
	EU_2: *Zoragarria zen ... Maiatza ... Bero egiten zuen ... Eta ni lotsagabe sentitzen nintzen , ez dakizu ondo .*
#110678	Gedächtnis , weil ich gerade zu dieser Zeit im **Begriffe** war , die Reisebeschreibung des berühmten
	EU_2: *Oroimenean geratu zitzaidan ; izan ere , hain zuzen garai hartan , izen bereko ibiltari ezagunaren bidaiei buruzko kontakizunak irakurtzekotan nintzen .*
#136122	sonnverbrannte Wüste geht . Und schon war Andreas im **Begriffe** , den Helden des Films sympathisch und
	EU_2: *Eta Andreas hasia zenean heroiart sinpatia eta atxikimendua hartzen , filmea uste ez bezala zorion aldera makurtu , handik pasatzen zen zientifiko espedizio batek desertuan galdutako gizona salbatu , eta Europako zibilizazioaren gerizpera ekartzen zuen atzera .*
#136185	den Helden des Films . Und schon war er im **Begriff** , sich zu erheben , als auf der Leinwand
	EU_2: *Eta jeikitzeko zorian zegoela , hara non pantaHan agertu zen lehentxeago , barran zegoela , tabernariaren bizkar atzean ikusitako eskola lagunaren irudia .*
#139853	Hier stieß er auf einen Mann , der eben im **Begriffe** war , die Treppe hinaufzusteigen , und
	EU_2: *Han , eskaileretan gora hastear zen gizon batekin topo egin zuen . Oso ezaguna egin zitzaion .*
#151414	aufgewachsen und hatte von dorther eine Menge von **Begriffen** und Schablonen beibehalten . Er hatte theoretisch
	EU_2: *Gainera hezimolde burges txikian hazia zen , eta ordudanik kontzeptu eta kate andana gordetzen zituen .*
#154557	den Dichtungen des alten Indien ist dieser **Begriff** ganz unbekannt , die Helden der indischen
	EU_2: *India zaharreko poesietan kontzeptu hori erabat ezezaguna da , Indiako epopeietako heroeak ez dira pertsonak , pertsona matazak , enkarnazio sailak baizik .*
#155264	Leben . Was die Menschen jeweils unter dem **Begriff** " Mensch " verstehen , ist stets nur eine
	EU_2: *Gizonek " gizaki " kontzeptuaren bidez gehienetan ulertzen dutena , akordio burges iragankor bat besterik ez da betiere .*
#160170	und erkannte mich erst , als ich schon im **Begriff** war , an ihm vorüberzugehen . Er stürzte
	EU_2: *Parez pare zetorkidan jakituna , zurrun eta nahiko begilauso , eta jaramonik egiteke albotik pasatzea pentsatzen nuen unean ezagutu ninduen .*
#202482	Ihre Arbeit zu Ende zu führen . Wenn der **Begriff** der Pflicht Ihnen unbekannt ist ... " Er
	EU_2: *Egitekoaren ideia ezezaguna bazaizu ... .*
#202528	Warten Sie ! " sagte Gustav höflich . " Den **Begriff** der Pflicht allerdings kenne ich nicht
	EU_2: *- Egitekoaren ideaia egiaz ez dut ezagutzen , ez gehiago .*
#202602	Gegenteil getan . Aber wenn ich auch den **Begriff** der Pflicht nicht mehr kenne , so kenne
	EU_2: *Baina eginbeharraren ideia gehiago ezagutzen ez badut ere , ongi ezagutzen dut ordea kulparena ; agian biak gauza bera izango dira .*
#203637	Menschen , einst ein hohes Ideal , ist im **Begriff** , zu einem Klischee zu werden . Wir Verrückten
	EU_2: *Gizairudia , garai batean ideal altu bat , azken batean klitxe bihurtzeko zorian dago .*
#204784	lückenhafte Seelenlehre der Wissenschaft durch den **Begriff** , den wir Aufbaukunst nennen . Wir zeigen
	EU_2: *Guk horregatik , akatsez beteriko jakintza horren arimari buruzko irakaskintza osatu egiten dugu , eraikuntz deritzogu n ideiaren bidez .*
#392066	der Herr Vorsteher schwieg . Ich war im **Begriff** , ihm irgendeine grobe Beleidigung ins
	EU_2: *Irain satsuren bat aurpegira botatzear egon nintzen .*

Begriff *m.* kontzeptu | adigai ▸ ich stand im Begriff, das Haus zu verlassen: *Etxea uzteko abian nengoen*

editorial EAH 2007

2.4.1 Summary: Functions

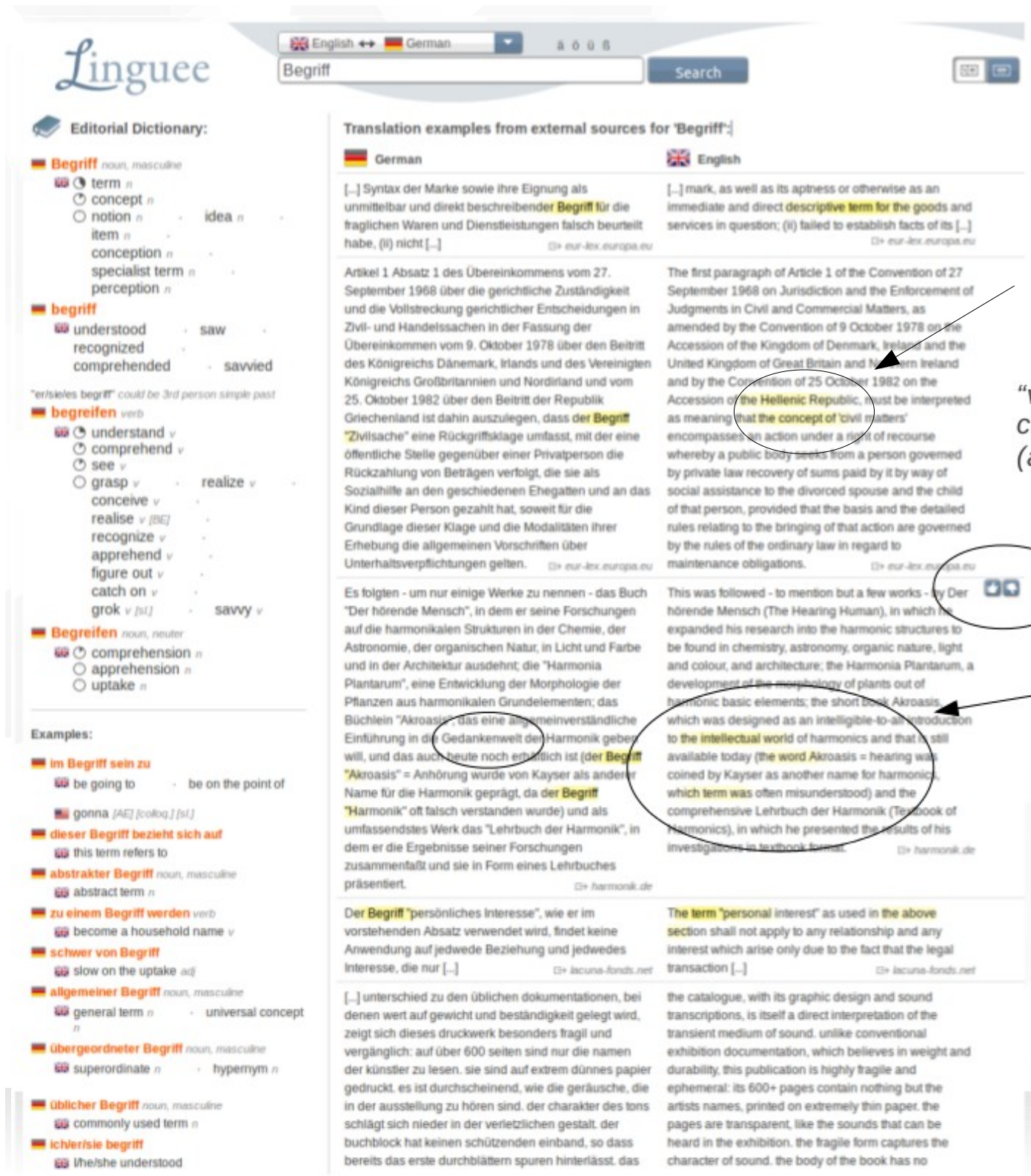
- Parallel KWIC display:
  real translation examples
  - TEs for the headword that are not listed in the editorial entry
  - TEs in context
    - domain, register, regional
  - TEs for MWE and collocations
    - MWE: “im Begriff sein”
    - phraseological coll.: “üblicher Begriff”
    - idiom: “schwer von Begriff”
  - POS-transposition
    - in editorial entry POS is maintained even if translators don't
  - other adaptations done by human translator

berücksichtigen (V)	consider (V)	aintzat hartu (PosP+V) kontuan izan (PosP+V)
Berücksichtigung (N)	consideration (N)	aintzat hartze?? (PosP+N) kontuan izate?? (PosP+N)
unter Berück-sichtigung +Gen. (Prep+N+Gen.)	*considering sth.; taking sth. into account (Part. Phrases)*	*zbt. aintzat harturik; zbt. kontuan izanik (Part. Phrases)*

Abfrageergebnis für Ihre Anfrage nach »Begriff«

30 Treffer aus dem **deutsch->englisch** Lexikon:		#Belege
Begriff	**notion**	6
Begriff	**concept**	3
Begriff	**term**	2
begriff	**comprehended**	2
Begriff	**idea**	2
begriff	understanded	2
Begriff	**sense**	2
Begriff	**realization**	2
Begriff	**acquaintance**	2
Begriff	**knowledge**	2
Begriff	**item**	2
Begriff	**apprehension**	1
Begriff	seidea	1
Begriff	**conception**	1
Begriff	**perception**	1
im **Begriff** sein	to be about to	1
üblicher Begriff	**commonly** used **term**	1
aktrakter **Begriff**	**abstract idea**	1
im **Begriff** sein zu	be about to	2
abstrakter Begriff	**abstract term**	2
im **Begriff** sein zu	to be on the **point** of	2
im **Begriff** sein zu	to be about to	2
abstrakter Begriff	**abstract concept**	1
schwer von **Begriff**	**slow** on the **uptake**	1
im **Begriff** sein zu	to be **going** to	1
im **Begriff** sein zu	**gonna**	1
numerischer Begriff	**numeric character**	1
langsam von **Begriff**	**slow-witted**	1
der **Begriff** 'Pferd'	the **concept** 'horse'	1
mehrdeutiger Begriff	**ambiguous concept**	1

dict.uni-leipzig.de: DE-EN parallel corpus TE extraction and coocurrence statistics as on-the-fly generated dictionary entry (under development; still noisy)



Editorial dictionary entry and parallel KWIC combination as e-dictionary search result (linguee.com)

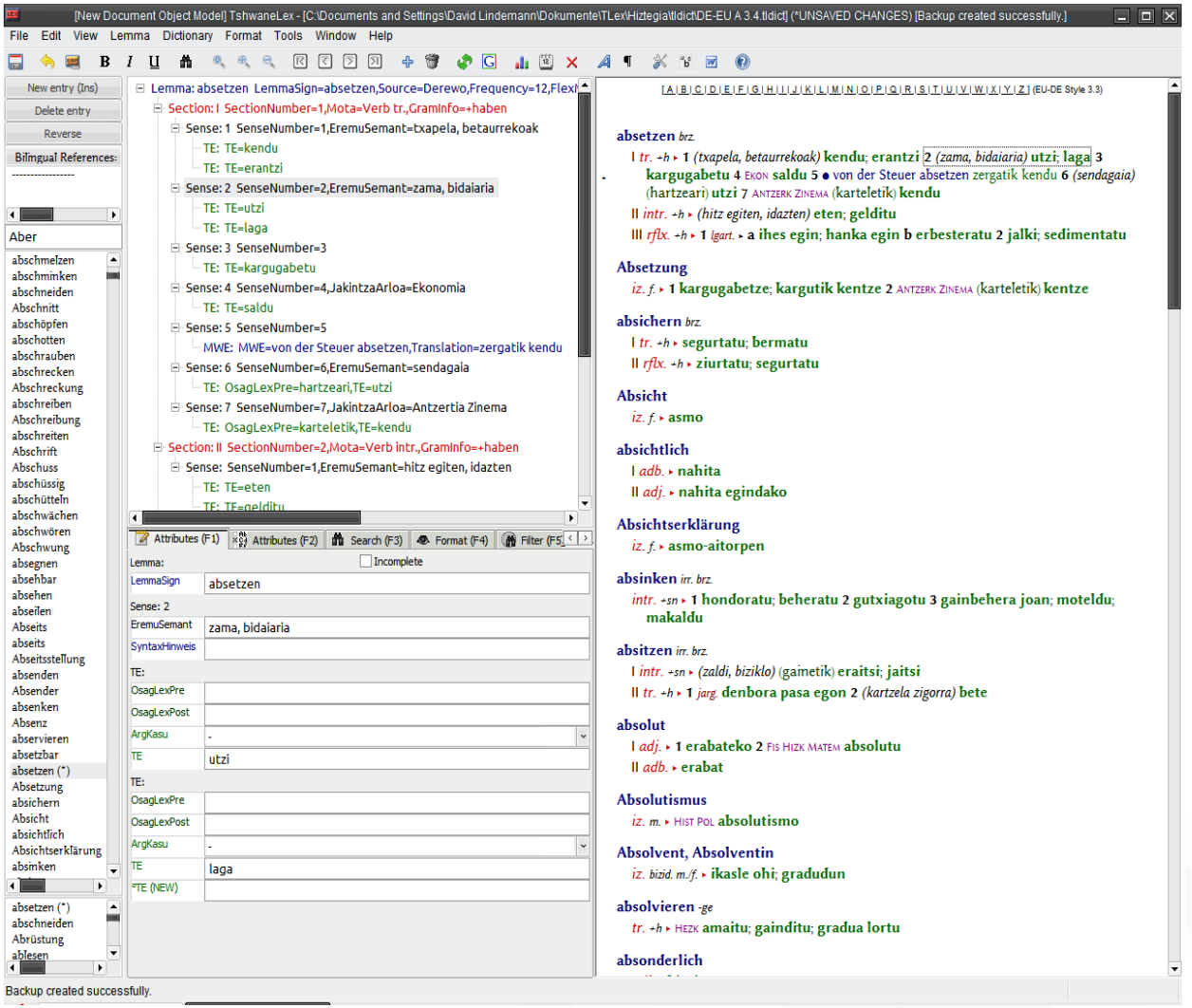
Noise: word highlighted without clear reason

"vote up" / "vote down" buttons: crowd votes for good examples (affects display order)

Noise: 3 TEs for 2 appearances of the HW: "word", "term" plus "intellectual world" (TE for "Gedankenwelt"="Begriffswelt")

3.0 New DE-EU dictionary: Editing, Corpora, Tools

- Editing Software: TshwaneLex (TLex) (de Schryver & Joffe 2005, 2010)
  - Freely designable DTD, Styles
  - Output as XML, html, rtf, tsldict
  - Attribute Lists, Label Sets
  - Import, Reverse, Compare, Merge
  - CrossRefs, Hyperlinks
- Software for Corpus Tagging and KWIC: SketchEngine (SkE) (Kilgarriff et al. 2004)
  - Queries in large annotated corpora for DE and EU
  - Wordsketch, GDEX
  - Parallel Concordances



3.1 New DE-EU Literary Corpus

- Created at UPV-EHU
  (Sanz Villar 2011, Zubillaga 2012)
- 81 Digital and OCR-ed literary DE originals and official EU versions
- about 2 million tokens per language
- 146.457 sentence pairs
- Sentence alignment
  one-to-many alignment revision by hand
  - Improved sentence alignment tools like *hunalign* (Varga et al. 2005) require seed lexicons (DE-EU still not available)
- Exported to TMX format, imported in the SketchEngine
- DE: Lemmatized with TreeTagger
- EU: Tokenized, not lemmatized (at the moment)

lemma	Freq
d	145113
und	57846
sein	54259
ich	46686
er	43723
ein	39059
haben	25846
sie	25470
in	22647
zu	22589
nicht	22374
es	19506
er|es|sie	18760
mit	15777
auf	15423
du	14433
ihr	13580
wie	11903
von	11372
an	10944
so	10854
als	10526
aber	10518
sagen	10474
wir	10217
werden	9393
Sie|sie|sie	9335
noch	8271
was	7530

word	Freq
eta	58220
zuen	29551
ez	28415
zen	21353
bat	16811
ere	16025
egin	14295
da	10926
bere	9961
zion	9338
izan	8328
du	6907
baina	6592
behar	6149
egiten	6032
nahi	5746
Eta	5647
esan	5407
Ez	5143
bezala	5089
ziren	4948
zegoen	4688
Baina	4664
nuen	4493
hori	4374
dut	4182
zituen	4134
nire	4130
zuten	3887

3.2 New DE-EU dictionary: frontend design

Begriff ~s, ~e

Substantiv m. s. Verbinf. begreifen ▸ **1 kontzeptu; ideia; hitz; ezagutza 2** ● im Begriff sein -tzera joan; -tzeko zorian egon; -tzeko asmoa izan **3** ● sich (k)einen Begriff (von etw.) machen (ez) jakin

Article has cross-references to >>

begreifen *unregelmäßig* begreift, begriff, begriffen

Verb Transitiv +haben ▸ **1** = verstehen **ulertu; aditu; barrendu 2** *[als etw.]* **-tzat hartu**

Hyperlinks Block:

- German monolingual dics
  - DWDS, Leipzig, Openthesaurus...
- Wiktionary & Wikipedia
- Comment form (user comments)

Corpus: **deTenTen**
Hits: **280464** (98.6 per million)

Page 1 of 7012 Go Next | Last

#2706173999	denjenigen vernünftig reden, die unseren	Begriff	von Öffentlichkeit nicht teilen? Nachvollziehen
#2709650368	Leistungen sind aber nur durchschnittlich. Der	Begriff	Zeitgeschichte ist aus dem Informationsinteresse
#2271543246	wiederum hinterließ dem 18. Jahrhundert den	Begriff	einer dynamischen, offenen Geschichte.
#2100412987	ausgewählte Forum wird dann nach den eingegebenen	Begriffen	durchsucht. Gib einfach den Benutzernamen
#645973149	. Zumindest gegen Ihr Verständnis dieses	Begriffs	. Ein sportlich gehaltener Kinderrodel in
#2719021251	Projektbeschreibung taucht immer wieder der	Begriff	Modul auf. Ein Modul ist letztlich ein
#1864856840	Beipackzettel finden Sie dies unter dem	Begriff	Gegenanzeigen. Auch weitere Hinweise zu
#843383439	nicht zusammenpassen. Und entfernte den	Begriff	?gesund? von der Verpackung. Im Internet
#1488077412	Kommunikationspartner einander sind, desto mehr werden auch	Begriffe	aus dem umgangssprachlichen Bereich verwendet
#284107654	Expedition@bo.drs.de bestellt werden. Der	Begriff	Narzissmus ist so vieldeutig, dass ihm
#1453968627	Text::Abbrev Erzeugt aus einer Liste von	Begriffen	eine Abkürzungstabelle. Text::Balanced
#1055825351	Friedrichstraße liefen. Präsenzpflicht Der	Begriff	"City-Filiale" bürgerte sich ein. Hier
#1545778620	programmatische Vorgabe für die nächsten Jahre. Der	Begriff	ist ehrlich. Denn Talent bedeutet nicht
#626917325	eines tragenden Bauteils nach § 24 . Der	Begriff	"raumabschließend" ist somit nicht im landläufigen
#998849916	Tschernobyl hat die tödliche Dimension des	Begriffs	Restrisiko spüren lassen. Nicht Überheblichkeit
#2484747439	spirits" nur am Rande auftaucht und der	Begriff	behavioral finance" nicht einmal Erwähnung
#1865327519	Kulturforum: der Melancholie. Obwohl als	Begriff	durchaus gebräuchlich, dürften die meisten
#1066195727	Eine Firmware? Ein ROM? Ich hoere immer die	Begriffe	Firmware und ROM, wie soll man sich das
#931180247	nichts Ernstes. Nie, sagt sie, würde ihr der	Begriff	"Provinz" über die Lippen kommen. Nicht
#151488829	einer einfachen Suche. Dokumente, welche den	Begriff	'keine' enthalten, werden unverändert gewichtet
#1315067624	meint: (17.2.2008 um 23:29) Antworten Der "	Begriff	WLAN-Kabel" ist ein Oxymoron, wie ich neulich
#1896232155	Weder trifft es bei mir zu noch ist der	Begriff	"Bumser" in meinen Augen sozialadäquat.
#2635746810	Vollends absurd wird eine Orientierung an	Begriffen	, wenn die benutzten Begriffe mit den abgebildeten
#2202616937	Erkrankung ist der medizinisch korrekte	Begriff	, unter dem über 800 derzeit bekannte Krankheitsformen
#1695537921	von Absatz 1 dient der Legaldefinition des	Begriffs	" internationale Koproduktionen ". Zu §
#1536755453	Zugangsdaten und Standortdaten. Bezüglich des	Begriffes	"Zugangsdaten" wird auf § 92 Abs. 3 Z 4
#558406229	geförderten, betrieblichen Altersversorgung. Der	Begriff	steht für eine Lebensversicherung, die
#176324671	Entwicklungsumgebungen und Werkzeuge nimmt der	Begriff	der Produktivitätssteigerung eine zentrale
#112512653	eine Krümmung dieser Raumzeit zurück. Der	Begriff	der Gravitationskraft wird dabei ersetzt
#90941217	allerdings nur, wenn man einen sehr weiten	Begriff	von Techno heranzieht. Noch stärker als
#2761581102	neue Bedeutung der Biologie z.B. durch den	Begriff	der Bionik einer größeren Öffentlichkeit
#2654703085	diese Weise sollte es möglich sein, den	Begriff	der Regelstudienzeit ad absurdum zu führen
#2598941734	natürliche oder künstliche Allergene aus der Der	Begriff	der Umwelt ist geprägt durch die anthropogene
#1001387296	Leben der Menschen und die Welt durch den	Begriff	der Gerechtigkeit zurückgebunden an Gott
#74499997	ich die Politik. Nehmen wir den Marxschen	Begriff	: hier ist Politik die Herrschaft des Menschen
#1040568325	zwischenmenschliche Kommunikation mit den	Begriffen	aus der Signalübertragung beschrieben wird
#2074189024	Diskussion so bedeutsamen aber auch schwierigen	Begriff	des Habitus zu erläutern und in seinen

Corpus: **DE_1**
Hits: **86** (37.3 per million) [EU_1] [EU_2]

Page 1 of 5 Go Next | Last

#95733	war frech , davon machen Sie sich keinen	Begriff	. Denken Sie sich , ich hab meinen Überzieher
	EU_2: *Zoragarria zen ... Maiatza ... Bero egiten zuen ... Eta ni lotsagabe sentitzen nintzen , ez dakizu ondo .*		
#110678	Gedächtnis , weil ich gerade zu dieser Zeit im	Begriffe	war , die Reisebeschreibung des berühmten
	EU_2: *Oroimenean geratu zitzaidan ; izan ere , hain zuzen garai hartan , izen bereko ibiltari ezagunaren bidaiei buruzko kontakizunak irakurtzekotan nintzen .*		
#136122	sonnverbrannte Wüste geht . Und schon war Andreas im	Begriffe	, den Helden des Films sympathisch und
	EU_2: *Eta Andreas hasia zenean heroiart sinpatia eta atxikimendua hartzen , filmea uste ez bezala zorion aldera makurtu , handik pasatzen zen zientifiko espedizio batek desertuan galdutako gizona salbatu , eta Europako zibilizazioaren gerizpera ekartzen zuen atzera .*		
#136185	den Helden des Films . Und schon war er im	Begriff	, sich zu erheben , als auf der Leinwand
	EU_2: *Eta jeikitzeko zorian zegoela , hara non pantaHan agertu zen lehentxeago , barran zegoela , tabemariaren bizkar atzean ikusitako eskola lagunaren irudia .*		
#139853	Hier stieß er auf einen Mann , der eben im	Begriffe	war , die Treppe hinaufzusteigen , und
	EU_2: *Han , eskaileretan gora hastear zen gizon batekin topo egin zuen . Oso ezaguna egin zitzaion .*		
#151414	aufgewachsen und hatte von dorther eine Menge von	Begriffen	und Schablonen beibehalten . Er hatte theoretisch
	EU_2: *Gainera hezimolde burges txikian hazia zen , eta ordudanik kontzeptu eta kate andana gordetzen zituen .*		
#154557	den Dichtungen des alten Indien ist dieser	Begriff	ganz unbekannt , die Helden der indischen
	EU_2: *India zaharreko poesietan kontzeptu hori erabat ezezaguna da , Indiako epopeietako heroeak ez dira pertsonak , pertsona matazak , enkarnazio sailak baizik .*		
#155264	Leben . Was die Menschen jeweils unter dem	Begriff	" Mensch " verstehen , ist stets nur eine
	EU_2: *Gizonek " gizaki " kontzeptuaren bidez gehienetan ulertzen dutena , akordio burges iragankor bat besterik ez da betiere .*		
#160170	und erkannte mich erst , als ich schon im	Begriff	war , an ihm vorüberzugehen . Er stürzte
	EU_2: *Parez pare zetorkidan jakituna , zurrun eta nahiko begilauso , eta jaramonik egiteke albotik pasatzea pentsatzen nuen unean ezagutu ninduen .*		
#202482	Ihre Arbeit zu Ende zu führen . Wenn der	Begriff	der Pflicht Ihnen unbekannt ist ... " Er
	EU_2: *Egitekoaren ideia ezezaguna bazaizu ... .*		
#202528	Warten Sie ! " sagte Gustav höflich . " Den	Begriff	der Pflicht allerdings kenne ich nicht
	EU_2: *- Egitekoaren ideaia egiaz ez dut ezagutzen , ez gehiago .*		
#202602	Gegenteil getan . Aber wenn ich auch den	Begriff	der Pflicht nicht mehr kenne , so kenne
	EU_2: *Baina eginbeharraren ideia gehiago ezagutzen ez badut ere , ongi ezagutzen dut ordea kulparena ; agian biak gauza bera izango dira .*		
#203637	Menschen , einst ein hohes Ideal , ist im	Begriff	, zu einem Klischee zu werden . Wir Verrückten
	EU_2: *Gizairudia , garai batean ideal altu bat , azken batean klitxe bihurtzeko zorian dago .*		
#204784	lückenhafte Seelenlehre der Wissenschaft durch den	Begriff	, den wir Aufbaukunst nennen . Wir zeigen
	EU_2: *Guk horregatik , akatsez beteriko jakintza horren arimari buruzko irakaskintza osatu egiten dugu , eraikuntz deritzogu n ideiaren bidez .*		
#392066	der Herr Vorsteher schwieg . Ich war im	Begriff	, ihm irgendeine grobe Beleidigung ins
	EU_2: *Irain satsuren bat aurpegira botatzear egon nintzen .*		

3.3 XML element hierarchy

```
<entry>
    <sense>
        <gramGrp>
        </gramGrp>
    </sense>
</entry>
```

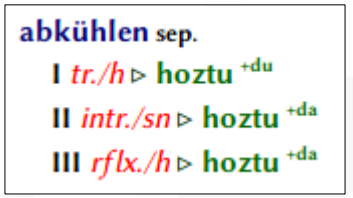
- Bilingual entry (TEI)
  - syntactical information below WSD
  - TEI Standard
  - most common
  - redundancies kept on the syntax level

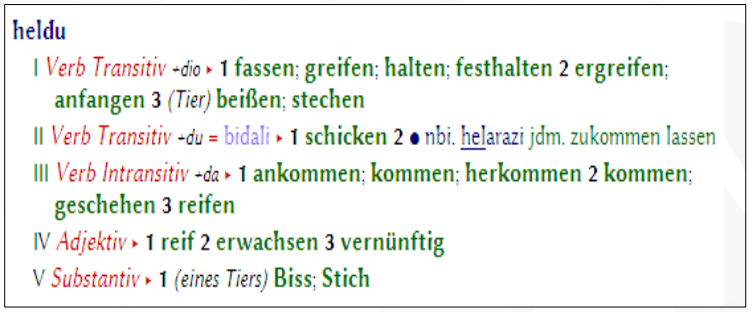
```
<entry>
    <synt>
        <sense>
        </sense>
    </synt>
</entry>
```

- Bilingual entry (2)
  - WSD below syntactical disambiguation
  - Redundancies
  - Common in German
    - monolingual for learners (PONS GW DaF)
    - bilingual (Langenscheidt)
    - never different POS
  - Useful for Basque(?)
    - possible to group different POS below the same headword

```
<synset>
    <LexUnit>
    </LexUnit>
</synset>
```

- WordNet
  - Lexical Headword below WSD





3.4 Verb entry

```
<Lemma id="19375" LemmaSign="befinden" Source="Derewo" DerewoRang="569" Frequency="9"
        FlexMorfTag="unregelmäßig, ohne ge">
        <Section id="211053" SectionNumber="1" POS="Verb Transitiv" GramInfo="+haben">
                <Sense id="211054" SenseNumber="1" SyntaxHinweis="jdn./etw. für +adj.">
                        <TE id="211055" TE="-tzat eman"/>
                        <TE id="289097" TE="iritzi"/>
                        <MWE id="289101" MWE="etw. für gut ~" Translation="ontzat eman; -i oniritzi"/>
                </Sense>
        </Section>
        <Section id="289102" SectionNumber="2" POS="Verb Intransitiv" GramInfo="+haben">
                <Sense id="289103" SenseNumber="1">
                        <TE id="289104" TE="erabaki"/>
                        <TE id="289105" TE="erabakia hartu"/>
                </Sense>
        </Section>
        <Section id="289098" SectionNumber="3" POS="Verb +&amp;apos;sich&amp;apos;" GramInfo="+haben">
                <Sense id="289099" SenseNumber="1">
                        <TE id="289106" TE="izan"/>
                        <TE id="289100" TE="egon"/>
                        <TE id="289107" TE="aurkitu"/>
                </Sense>
        </Section>
</Lemma>
```

befinden Derewo 569 - *unregelmäßig, ohne ge.*

I *Verb Transitiv* +haben ▸ **1** *[jdn./etw. für +adj.]* **-tzat eman; iritzi**
- etw. für gut befinden ontzat eman; -i oniritzi

II *Verb Intransitiv* +haben ▸ **1 erabaki; erabakia hartu**

III *Verb +'sich'* +haben ▸ **1 izan; egon; aurkitu**

3.5 Noun entry

Ausdruck[1] Derewo 1918 ~s, Ausdrücke

Substantiv m. s. Verbinf. ausdrücken ▸ **1 adierazpen 2 termino; esamolde; esapide 3 ezaugarri; adierazgarri 4 adierazkortasun; adierazgarritasun**

Ausdruck[2] Derewo 1918 ~s, Ausdrucke

Substantiv m. s. Verbinf. ausdrucken ▸ **1 inprimatze; inprimaketa**

3.6 Adjective and Adverb

anständig Derewo 8054

I *Adjektiv* ▸ **1 zintzo**

II *Adjektiv adverbial* ▸ **1 zintzo; zintzoki; jendetasunez**

anteilig Derewo 20466

I *Adjektiv* ▸ **1 zatikako; proportzional**

II *Adjektiv adverbial* ▸ **1 zatika; proportzionalki**

arg Derewo 3262

I *Adjektiv* ▸ **1 latz; gaitz 2** *veraltend* **gaizto; maltzur**

II *Adjektiv adverbial* ▸ **1** *ugs. südd.* **oso**

anders Derewo 596

I *Adverb* ▸ **1 ezberdin; bestela; beste era batean**

II *Adverb präd./att.* ▸ **1 ezberdin; bestelako; beste era bateko**

4.0 Automatic TE pairing for Bilingual Dictionary Drafting

- Corpus based
  - 1. GIZA++
  - 2. Bifid
  - 3. Elhuyar Pivot Dictionaries
- Others
  - 4. Wikimedia
  - 5. WordNet

4.1 Corpus-based strategies: GIZA++

- Bilingual Parallel Corpus, **GIZA++**
  - English-Swedish 700.000 sentence pairs, 82,3% precision, 73,3& recall (Holmqvist 2010)
  - Lithuanian 1.765.000 tokens, Hungarian 2.121.000 tokens, 4026 TE-candidates (Héja 2010)
  - DE-EU:
    - DE: 2.303.307 tokens,
    - EU: 1.948.504 tokens,
    - Extremely low recall:
      - 573 TE-candidates,
      - 153 after revision
    - Low precision
    - First attempt, Gorka Labaka, UPV-EHU
    - Second attempt: After lemmatizing EU corpus

abend	*gauean*	aber	*ordea*
abend	*gau*	aber	*-baina*
abend	*arrats*	abgesehen	*utzita*
abend	*arratsean*	abgesehen	*aparte*
abend	*iluntzean*	abgrund	*amildegira*
abend	*arratsaldean*	abgrund	*amildegiaren*
abend	*gauerako*	abgrund	*amildegia*
abend	*arratsalde*	abgrund	*amildegi*
abend	*gauez*	abholen	*bila*
abendessen	*afaria*	abrupt	*-batean*
abendessen	*afaltzeko*	abschied	*agur*
abends	*gauean*	abschied	*adio*
abends	*iluntzean*	absicht	*asmoa*
abends	*gauez*	absicht	*nahita*
abends	*arratsean*	absicht	*asmo*
abends	*gauetan*	absicht	*asmorik*
abends	*arratsalde*	absichten	*asmo*
abenteuer	*abentura*	absichtlich	*nahita*
abenteuer	*abenturak*	absichtlich	*propio*
aber	*baina*	abstand	*tartea*

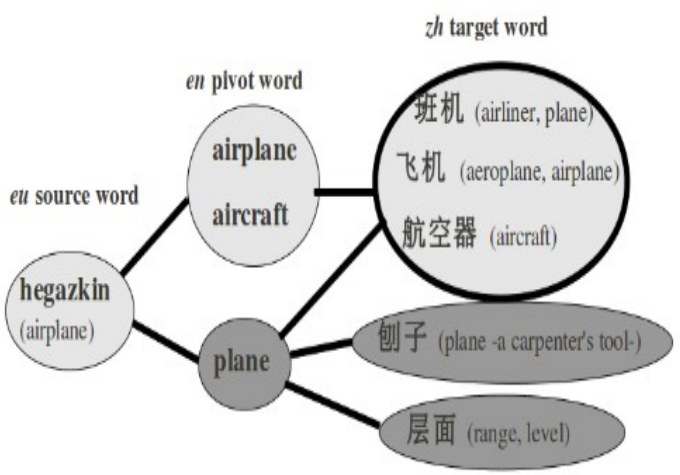
4.2 Corpus based strategies: Bifid

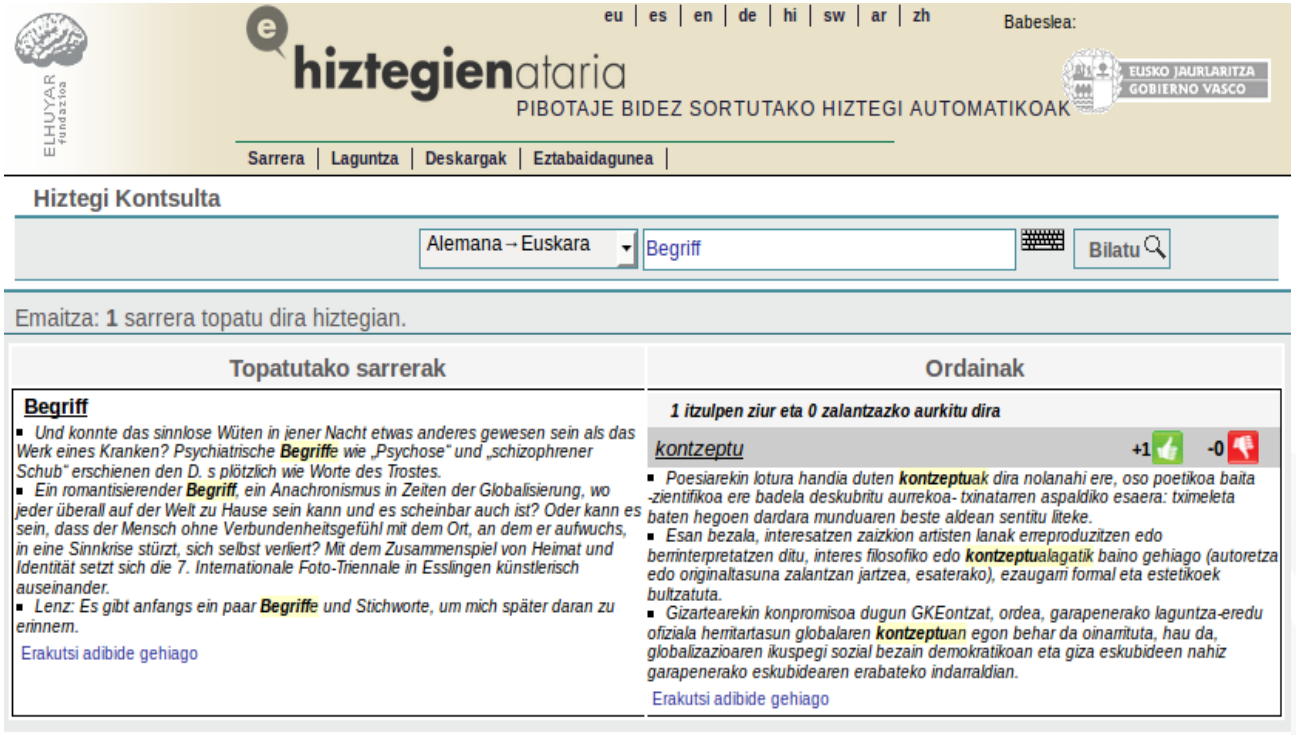
- Bilingual Corpus, **Bifid** (*cf.* Nazar 2012)
  - First attempt done: 2089 pairings,
  - Works much worse than with other language-pairs, evaluation in course
  - Insufficient corpus size
  - Second attempt: After lemmatizing EU corpus

wills gesicht	*willen aurpegia*
bildhauer	*eskultorea*
einem tiefen	*hasperen sakon*
individuum	*gizabanako*
eines hundes	*txakur baten*
mensch ärgere	*partxisean*
rechts links	*eskuin ezker*
schranktür	*armairuko*
sagte irene	*irenek*
wolfgang	*amadeo*
fünfunddreißig	*hogeita hamabost*
fußgänger	*oinezko*
tante	*izeba*

4.3 Corpus-based strategies: Elhuyar Pivot Dictionaries

- Pairing by combining pivot and corpus based methods
  - Research in course at Elhuyar (Saralegi et al., 2012): Pivot-based bilingual dictionary building
    - Pairings obtained by two reference dictionaries with English as a pivot
    - Inverse Consultation (IC) and Distributional Similarity (DS) in bilingual comparable corpora
    - Problems resolving polysemy (words with ambiguities are skipped)
    - Done for Basque and Chinese, German, Hindi, Swahili, Arabic
    - German-Basque: In the actual version, 13.000 DE lemmata with 1-25 basque TE each
    - Evaluation in course. Still noisy.





4.4 TE extraction from Wikimedia

- First attempts: TE extraction from Wikimedia database dumps (source: http://www.dicts.info)
  - Wiktionary (by 2012):
    - DE-EN 2161 pairings. DE 3066 LU (905 syn.), EN 2161 LU (0 syn.)
    - DE-EU 2131 pairings. DE 2316 LU (185 syn.), EU 2338 LU (207 syn.)
      - *Eval. by hand: 80%+ nouns, very freq. words and toponyms, nearly no wrong translations*
      - *after MWE filtering and synonym cut-off: 2030 pairings. 1330 yes, 700 not on our lemmalist.*
  - Omegawiki interlanguage links (by 2012):
    - DE-EN 5603 pairings. DE 6836 LU (1233 syn.) EN 6393 (790 syn.)
    - DE-EU 4603 pairings. DE 5362 LU (759 syn.), EU 6771 LU (2168 syn.)
      - *Eval. by hand: mostly rare terms, some wrong translations*
  - Wikipedia interlanguage links (wikipedia article titles) (by 2012):
    - DE-EN 7433 pairings
    - DE-EU: ?
      - *Eval. by hand: only nouns, frequent and rare terminology, no wrong translations*
      - Erdmann 2007: Much better results!

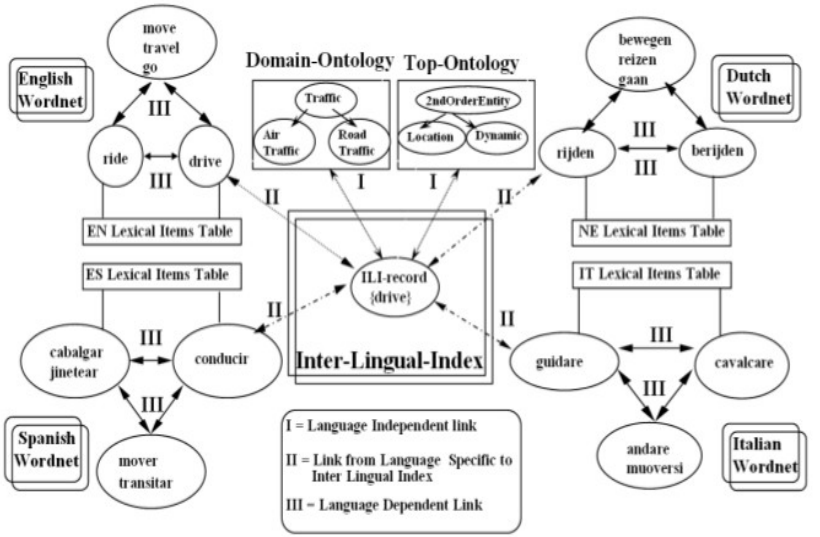
→ no significant difference between the two language pairs

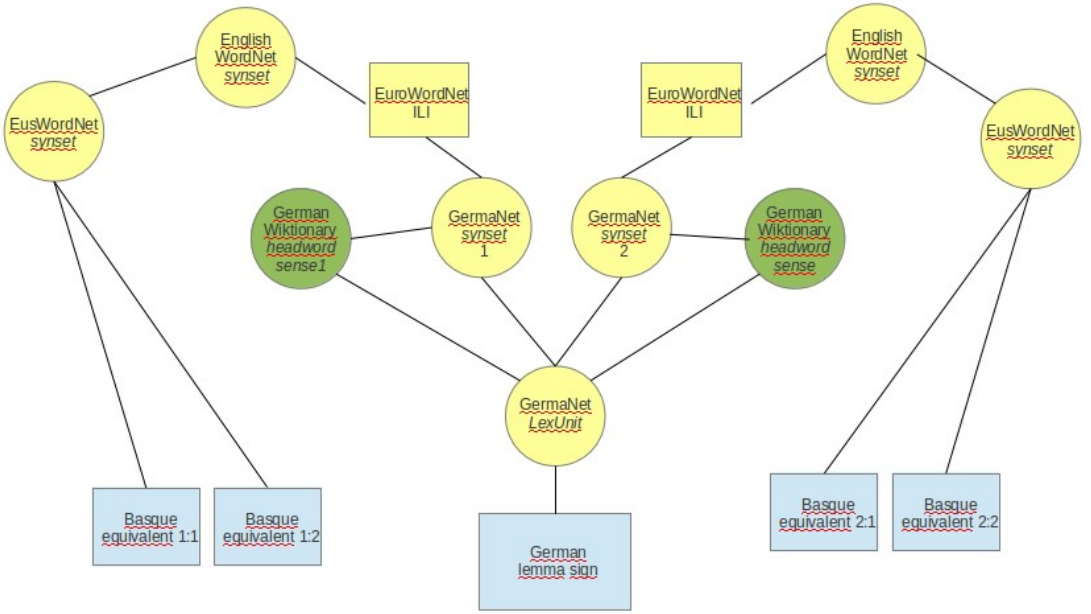
→ Wiktionary & Omegawiki (still) not useful for automatic dictionary drafting

→ Wikipedia bilingual term extraction: More research for DE-EU required

4.5 TE pairing via EuroWordNet

- Stops on the way:
  - German lemma-sign
  - GermaNet Lexical Unit
  - GermaNet synset(s)
  - EuroWordNet ILI record(s)
  - Princeton WordNet synset(s)
  - EusWN synset(s)
  - EusWN lemma-signs
- Advantages:
  - large databases
  - link on sense level
  - human-made WSD
- Known Problems:
  - Different approaches interling. WN
  - Different versions of PWN





5. Conclusions

- Importance of corpus methods in Lexicography increases
  - Indispensable in dictionary editing
  - Upcoming trend in dictionary publishing
  - Automatic Dictionary Drafting
- Not without large bilingual corpora
- Human Lexicographers (still) needed...

> *... The main resource remains the lexicographic knowledge of the project members in combination with large annotated text corpora which serve as a reference in all cases of doubt.*
>
> (Introduction to German Wordnet)

Thanks for your attention

Vielen Dank, Eskerrik asko

david.lindemann.soraluze@gmail.com