

# PARALLEL CORPORA

**LIDUN HAREIDE** | Researcher at Møreforsking Molde | 14.03.2018

UPF SERISS seminar, Barcelona

# CORPUS LINGUISTICS: LEECH

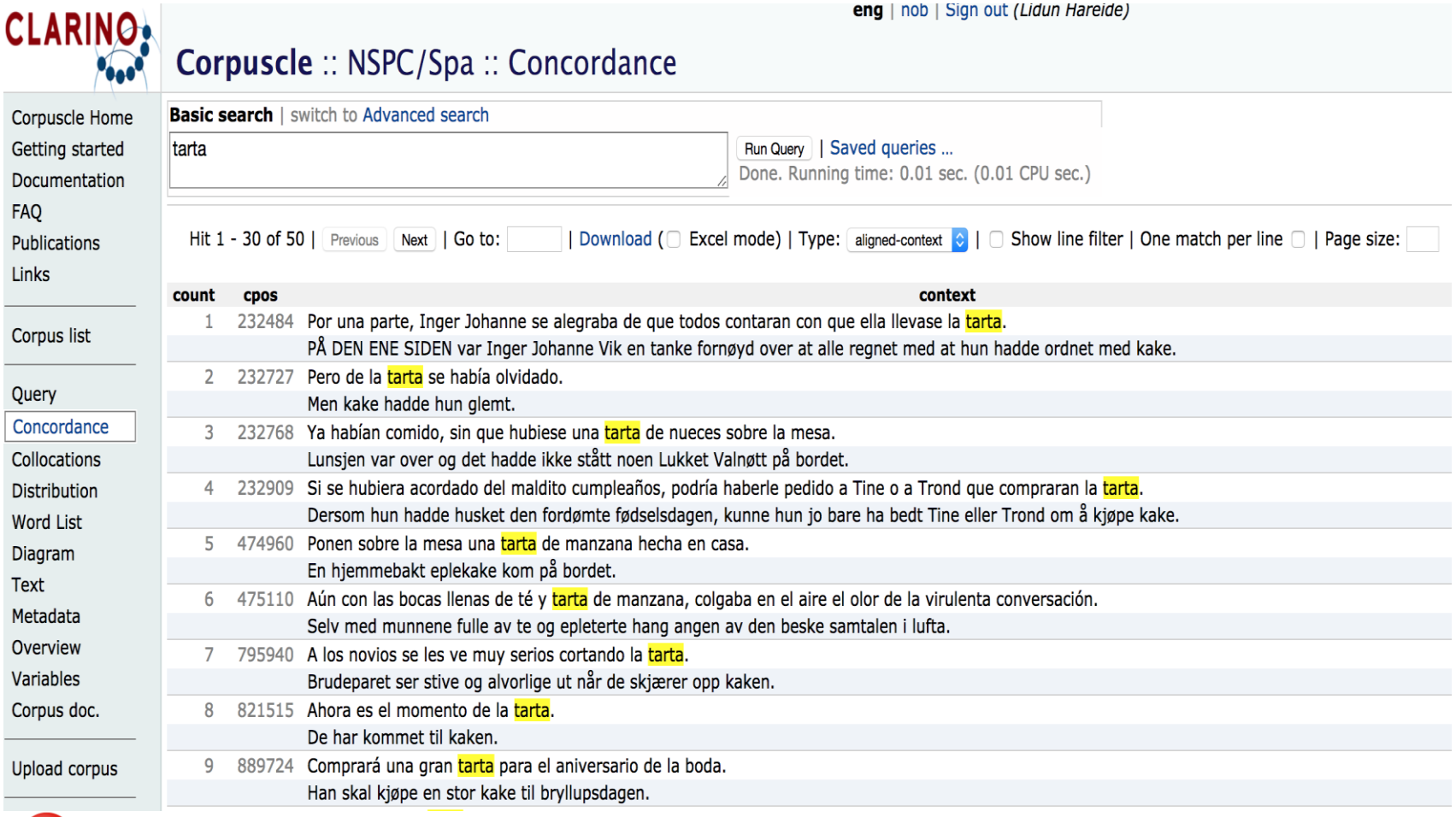
- "a significant advantage of the corpus linguistic methodology is that it allows for the analyst to approach the study of language from the context of the scientific method" (Leech 1992: 112).
  - Quantitative method
  - Hypothesis testing

- Corpora shows us how language really *is,* not how we have learned it should be
  - Caused a shift from *prescriptive* grammars to *descriptive*
  - Most new dictionaries are corpus-based.

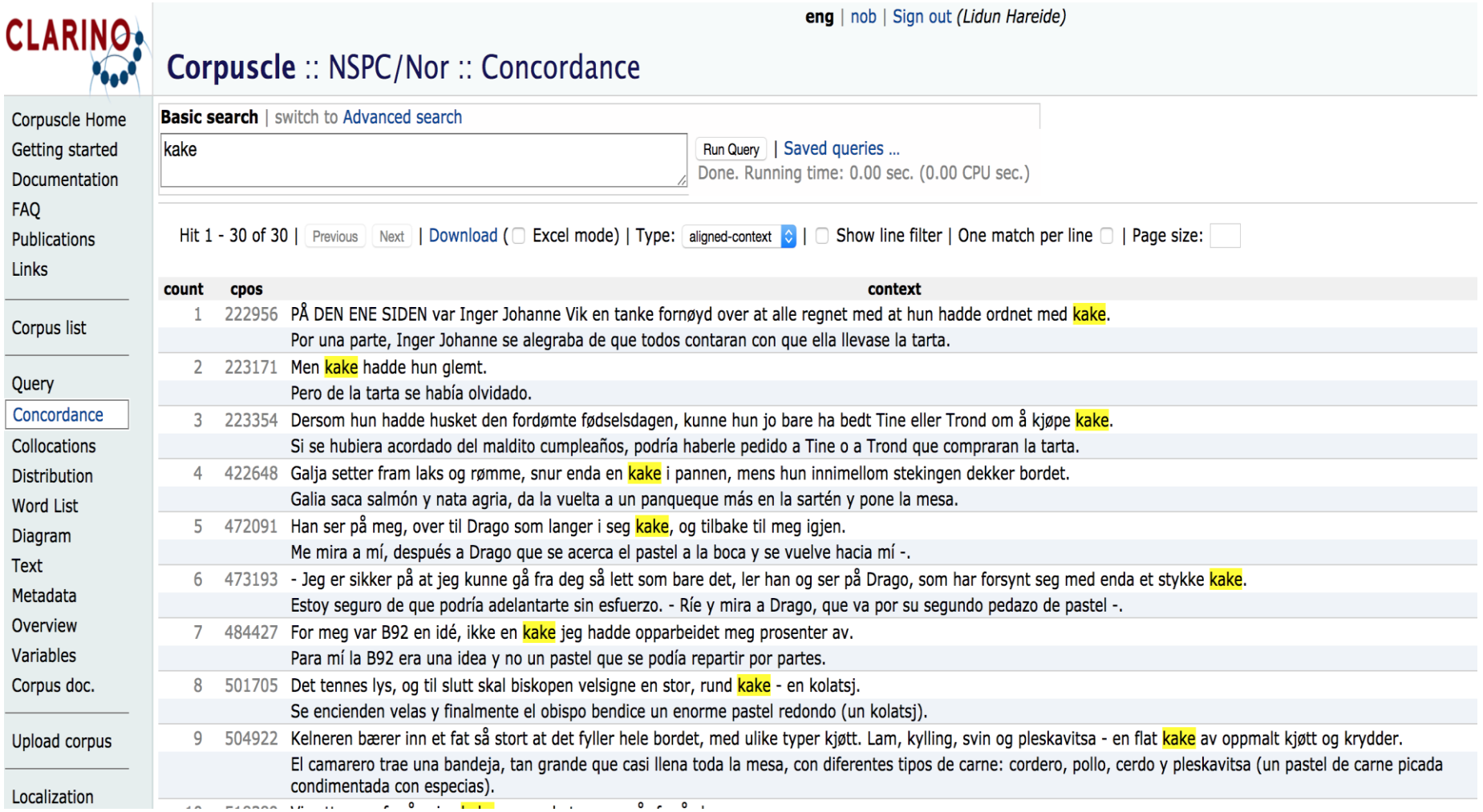
# PLAN FOR THIS PRESENTATION

- Definition of corpora
- Which universe of language is the corpus meant to represent?
- What determines the choice of corpus/corpora?
- Different types of corpora
- Parallel corpora: what can they be used for?
- How to construct a corpus; the case of the NSPC
- Measures taken to create a freely downloadable corpus
- New technology for DIY parallel/multilingual corpora

# EXAMPLE OF A PARALLEL CORPUS: NSPC: ALIGNED CONTEXT SPANISH-NORWEGIAN



# NSPC: ALIGNED CONTEXT NORWEGIAN-SPANISH



CLARINO

eng | nob | Sign out *(Lidun Hareide)*

## Corpuscle :: NSPC/Nor :: Concordance

Corpuscle Home
Getting started
Documentation
FAQ
Publications
Links

Corpus list

Query
Concordance
Collocations
Distribution
Word List
Diagram
Text
Metadata
Overview
Variables
Corpus doc.

Upload corpus

Localization

**Basic search** | switch to Advanced search

kake

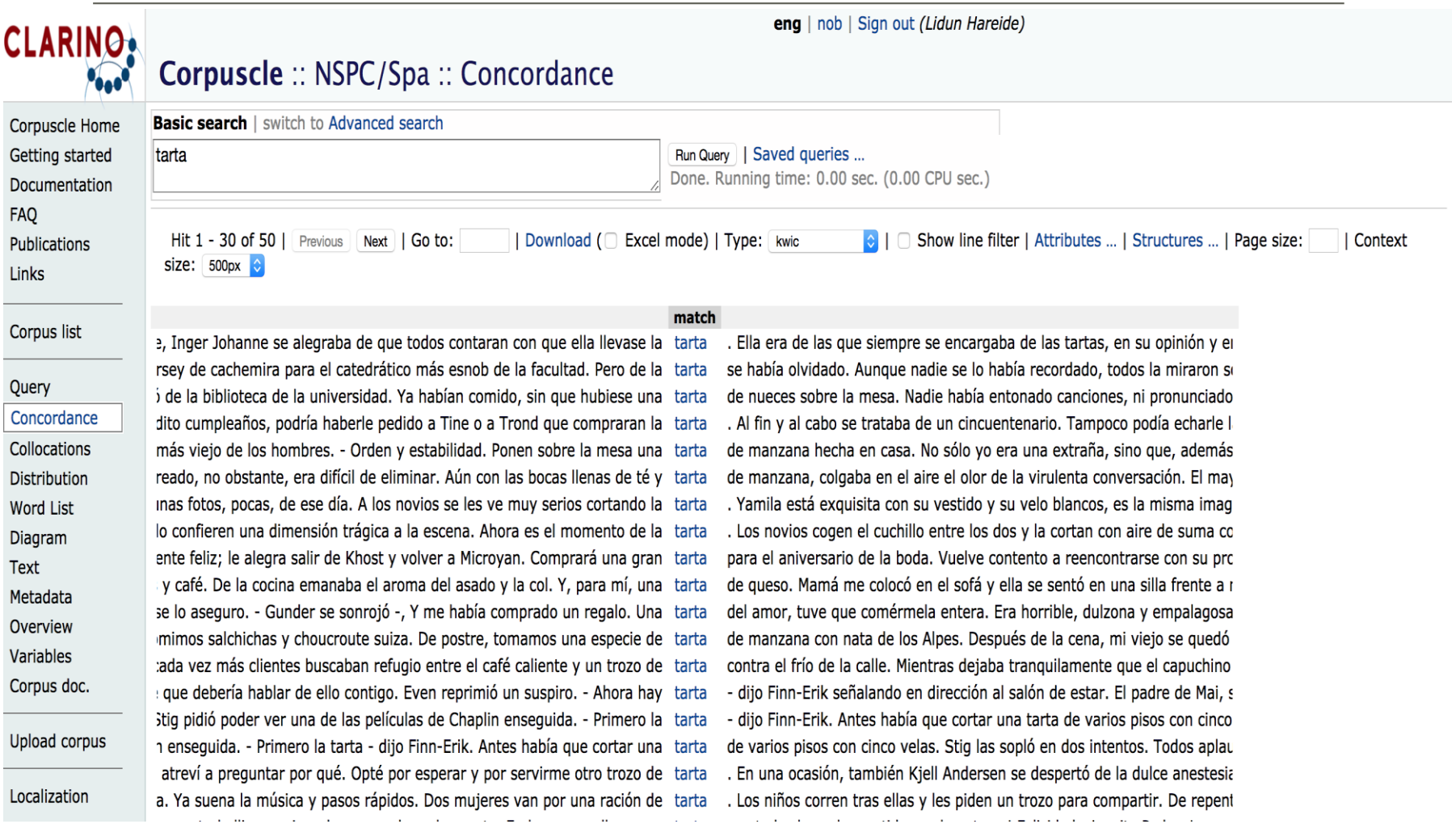
Run Query | Saved queries ...

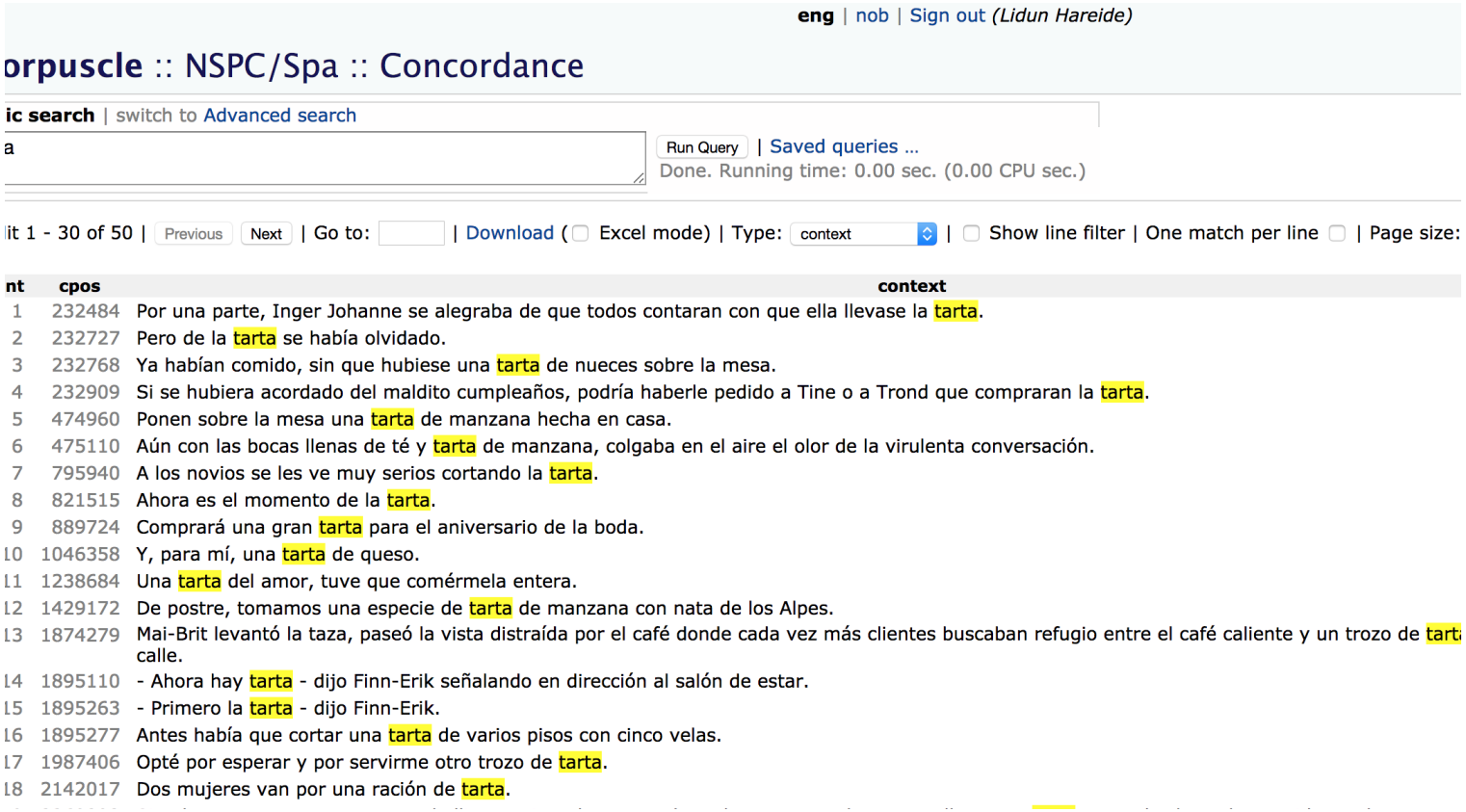
Done. Running time: 0.00 sec. (0.00 CPU sec.)

Hit 1 - 30 of 30 | Previous Next | Download ( Excel mode) | Type: aligned-context | Show line filter | One match per line | Page size:

| count | cpos | context |
|---|---|---|
| 1 | 222956 | PÅ DEN ENE SIDEN var Inger Johanne Vik en tanke fornøyd over at alle regnet med at hun hadde ordnet med kake. |
| | | Por una parte, Inger Johanne se alegraba de que todos contaran con que ella llevase la tarta. |
| 2 | 223171 | Men kake hadde hun glemt. |
| | | Pero de la tarta se había olvidado. |
| 3 | 223354 | Dersom hun hadde husket den fordømte fødselsdagen, kunne hun jo bare ha bedt Tine eller Trond om å kjøpe kake. |
| | | Si se hubiera acordado del maldito cumpleaños, podría haberle pedido a Tine o a Trond que compraran la tarta. |
| 4 | 422648 | Galja setter fram laks og rømme, snur enda en kake i pannen, mens hun innimellom stekingen dekker bordet. |
| | | Galia saca salmón y nata agria, da la vuelta a un panqueque más en la sartén y pone la mesa. |
| 5 | 472091 | Han ser på meg, over til Drago som langer i seg kake, og tilbake til meg igjen. |
| | | Me mira a mí, después a Drago que se acerca el pastel a la boca y se vuelve hacia mí -. |
| 6 | 473193 | - Jeg er sikker på at jeg kunne gå fra deg så lett som bare det, ler han og ser på Drago, som har forsynt seg med enda et stykke kake. |
| | | Estoy seguro de que podría adelantarte sin esfuerzo. - Ríe y mira a Drago, que va por su segundo pedazo de pastel -. |
| 7 | 484427 | For meg var B92 en idé, ikke en kake jeg hadde opparbeidet meg prosenter av. |
| | | Para mí la B92 era una idea y no un pastel que se podía repartir por partes. |
| 8 | 501705 | Det tennes lys, og til slutt skal biskopen velsigne en stor, rund kake - en kolatsj. |
| | | Se encienden velas y finalmente el obispo bendice un enorme pastel redondo (un kolatsj). |
| 9 | 504922 | Kelneren bærer inn et fat så stort at det fyller hele bordet, med ulike typer kjøtt. Lam, kylling, svin og pleskavitsa - en flat kake av oppmalt kjøtt og krydder. |
| | | El camarero trae una bandeja, tan grande que casi llena toda la mesa, con diferentes tipos de carne: cordero, pollo, cerdo y pleskavitsa (un pastel de carne picada condimentada con especias). |

# NSPC: KEY WORD IN CONTEXT





eng | nob | Sign out *(Lidun Hareide)*

**orpuscle :: NSPC/Spa :: Concordance**

**ic search** | switch to Advanced search

a

Run Query | Saved queries ...

Done. Running time: 0.00 sec. (0.00 CPU sec.)

it 1 - 30 of 50 | Previous Next | Go to: | Download ( Excel mode) | Type: context | Show line filter | One match per line | Page size:

| nt | cpos | context |
|---|---|---|
| 1 | 232484 | Por una parte, Inger Johanne se alegraba de que todos contaran con que ella llevase la tarta. |
| 2 | 232727 | Pero de la tarta se había olvidado. |
| 3 | 232768 | Ya habían comido, sin que hubiese una tarta de nueces sobre la mesa. |
| 4 | 232909 | Si se hubiera acordado del maldito cumpleaños, podría haberle pedido a Tine o a Trond que compraran la tarta. |
| 5 | 474960 | Ponen sobre la mesa una tarta de manzana hecha en casa. |
| 6 | 475110 | Aún con las bocas llenas de té y tarta de manzana, colgaba en el aire el olor de la virulenta conversación. |
| 7 | 795940 | A los novios se les ve muy serios cortando la tarta. |
| 8 | 821515 | Ahora es el momento de la tarta. |
| 9 | 889724 | Comprará una gran tarta para el aniversario de la boda. |
| 10 | 1046358 | Y, para mí, una tarta de queso. |
| 11 | 1238684 | Una tarta del amor, tuve que comérmela entera. |
| 12 | 1429172 | De postre, tomamos una especie de tarta de manzana con nata de los Alpes. |
| 13 | 1874279 | Mai-Brit levantó la taza, paseó la vista distraída por el café donde cada vez más clientes buscaban refugio entre el café caliente y un trozo de tart calle. |
| 14 | 1895110 | - Ahora hay tarta - dijo Finn-Erik señalando en dirección al salón de estar. |
| 15 | 1895263 | - Primero la tarta - dijo Finn-Erik. |
| 16 | 1895277 | Antes había que cortar una tarta de varios pisos con cinco velas. |
| 17 | 1987406 | Opté por esperar y por servirme otro trozo de tarta. |
| 18 | 2142017 | Dos mujeres van por una ración de tarta. |

# SO WHAT IS A CORPUS?

- The word *corpus* means *body* as in ”a body of text”
- ”A large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria” (Bowker and Pearson 2002: 9)
- ”A helluva lot of text, stored on a computer” (Leech 1992: 106)
- ”A corpus is a collection of texts when considered as an object of language or literary study” (Kilgarriff and Grefenstette 2003: 334)

# DEFINITION OF CORPUS

- There are many ways to define a corpus..., but there is an increasing consensus that a corpus is a collection of (1) *machine-readable* (2) *authentic* texts (including transcripts of spoken data) which is (3) *sampled* to be (4) *representative* of particular language or language variety.

  (McEnery, Tono, and Xiao 2006: 5) (emphasis original)

# MACHINE READABLE

- Earlier corpora: filing cards
  - Most field linguists made corpora before the Chomsky era of “armchair linguists”.

- Brown corpus 1964:  (American English): the first electronic corpus
- Available on http://clarino.uib.no/korpuskel/

# AUTENTIC TEXTS

- What are authentic texts?
  - Naturally occurring text (spoken or written)
    - *NOT* constructed examples
    - Entire texts or samples?
      - Traditionally samples (storage capacity)
      - Now: entire texts
    - Is the language identical all through a novel?

- Is literature authentic text? Even though it is fiction
  - How about adaptions?

- Are survey questions authentic texts?

- Are translations authentic texts?
  - Depends on what the corpus is meant to represent – i.e. what population the corpus is sampled from

# SAMPLE

- *Statistics* A portion drawn from a population, the study of which is intended to lead to statistical estimates of the attributes of the whole population

# REPRESENTATIVENESS IN CORPUS LINGUISTICS

- The importance of representativeness resides in the fact that data sampled from a representative corpus can to be generalized to a larger universe, which is one of the core uses of quantitative methodology (Hareide 2014:54).

# APPROACHES TO REPRESENTATIVENESS

- Traditionally:
  - As large a variation of text genres or text types as possible
- The modern approach:
  - As large a corpus as possible

- Both approaches are problematic

# WHAT UNIVERSE OF TEXT IS THE CORPUS MEANT TO REPRESENT?





# WHAT UNIVERSE OF TEXT IS THE CORPUS MEANT TO REPRESENT? (2)

- If our object of study is the universe “domestic animals”, will a corpus of cats be of use?
  - Will the corpus be more representative for the universe of “domestic animals” if we assemble all races, types and colours of cats? (greatest possible variation?)
  - Will the corpus be more representative of the universe of “domestic animals” if we gather a 100 million cats?
    - or even if we gather 100 million chihuahuas?

# SIZE MATTERS, BUT… THE CASE OF THE NORWEGIAN NEWSPAPER CORPUS

- Largest corpus of Norwegian in existence (1.5 billon tokens/words)
- Harvests text automatically from a series of Norwegian newspapers.
- Covers both written varieties

- Can findings be generalized back to the Norwegian language in general?

# BIBER'S PROPOSITION: DEFINE YOUR POPULATION

- The most important factor in ensuring representativeness is that the population the texts are to be drawn from is well defined, and that the entire variability of this population is represented in the corpus (Biber 1993: 243).

# WHAT DETERMINES THE CHOICE OF CORPUS?

- What is my research question/hypothesis ?
- Can the corpus I have chosen answer this research question?
  - Is the corpus representative?
  - Can my findings/results be generalized back to the universe of language I am studying?

- Users must be aware of the inner variations of a corpus

# TYPES OF CORPORA



# SIZE: FILLMORE

- I have two observations to make. The first is that I don't think there can be any corpora, however large, that contain all the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had the chance to examine, however small, has taught me facts I couldn't imagine finding out any other way… (Fillmore 1992:35)

# TYPES OF CORPORA

- Oral vs. Written
- General/reference corpora  vs. tematic
- Single language vs. multilanguage

# ORAL CORPORA

- Contains transcribed naturally occurring conversations
- Very time-consuming to transcribe speech, so oral corpora are normally very small
- ICAME-COLT: Bergen Corpus of Teenage Language and ICAME-LLC: London-Lund Corpus of Spoken English are oral corpora
  - Most corpora are corpora of written text.

# WRITTEN CORPORA

- Single language vs. many languages
- Single language:
  - Reference corpora
  - Tematic corpora
  - Corpora of translated text

# REFERENCE CORPORA

- Very large corpora designed to provide comprehensive information about a language.
- represent all/most language varieties and text types
- General: are supposed to represent entire languages, a reference that more specific corpora can be compared to
- Provide data for grammars, dictionaries, thesauri and diacronic studies
- http://www.collins.co.uk/page/The+Collins+Corpus
- http://www.rae.es/recursos/banco-de-datos/corpes-xxi

- Several English large corpora are freely available, many created by Mark Davies
- http://davies-linguistics.byu.edu/personal/

# TEMATIC CORPORA

- Normally smaller corpora with text from one single genre

- Exception: The Norwegian Newspaper Corpus
  - Automatically harvests texts from all the larger newspapers in Norway daily.
  - 1.5 billion words
  - The language in newspapers is *NOT* identical to general language, so this corpus cannot be used as a reference corpus.

# CORPORA OF TRANSLATED TEXT

- Collectons of translations from several languages into one target language
- Only the target text is included
- Used in Translation Studies to study the differences between original and translated text.
- Most famous: TEC Translational English Corpus (Mona Baker and Sara Laviosa, University of Manchester)

# BILINGUAL AND MULTILINGUAL CORPORA

- Comparable corpora
- (Bilingual) parallel corpora
- Multilingual corpora
- Comparable parallel corpora

# COMPARABLE CORPORA:

- “original texts in each language, matched as far as possible in terms of text type, subject matter and communicative function” (Altenberg & Granger 2002: 7–8).

  – The texts are *not* translations of each other

# (BILINGUAL) PARALLEL CORPORA (TRANSLATION CORPORA)

- Original text in one language and their translations into one or several other languages (Baker 1993:248, Altenberg and Granger 2002: 7-8)

- The two texts are *aligned* at sentence level/word level

| | context |
|---|---|
| 4 | Por una parte, Inger Johanne se alegraba de que todos contaran con que ella llevase la tarta. |
| | PÅ DEN ENE SIDEN var Inger Johanne Vik en tanke fornøyd over at alle regnet med at hun hadde ordnet med kake. |
| 7 | Pero de la tarta se había olvidado. |
| | Men kake hadde hun glemt. |
| 8 | Ya habían comido, sin que hubiese una tarta de nueces sobre la mesa. |
| | Lunsjen var over og det hadde ikke stått noen Lukket Valnøtt på bordet. |
| 9 | Si se hubiera acordado del maldito cumpleaños, podría haberle pedido a Tine o a Trond que compraran la tarta. |
| | Dersom hun hadde husket den fordømte fødselsdagen, kunne hun jo bare ha bedt Tine eller Trond om å kjøpe kake. |
| 0 | Ponen sobre la mesa una tarta de manzana hecha en casa. |
| | En hjemmebakt eplekake kom på bordet. |
| 0 | Aún con las bocas llenas de té y tarta de manzana, colgaba en el aire el olor de la virulenta conversación. |
| | Selv med munnene fulle av te og epleterte hang angen av den beske samtalen i lufta. |
| 0 | A los novios se les ve muy serios cortando la tarta. |
| | Brudeparet ser stive og alvorlige ut når de skjærer opp kaken. |
| 5 | Ahora es el momento de la tarta. |
| | De har kommet til kaken. |
| 4 | Comprará una gran tarta para el aniversario de la boda. |
| | Han skal kjøpe en stor kake til bryllupsdagen. |
| 8 | Y, para mí, una tarta de queso. |
| | Og en osteform til meg. |
| 4 | Una tarta del amor, tuve que comérmela entera. |

# USES OF BILINGUAL/MULTILINGUAL CORPORA (CONT.)

- Illuminate differences between source text and translations, and between native and non-native texts
- Can be used for a number of practical applications, e.g. Lexicography, language teaching and translation (Aijmer and Altenberg: 1996)

# USES OF PARALLEL CORPORA: TRAINING STUDENTS

- Students of translation, interpreting, general language have used parallel corpora for:
  - Term extraction
  - Collocations
  - Use of prepositions
  - Idioms/idiomatic expressions
  - Lexis

- Parallel corpora can raise the students awareness of professional translator's strategies (Pearson 2003)
  - Have proven to be very effective in the acquisition of translation expertise (Bowker 1998)

# TRANSLATOR TRAINING

- Bilingual parallel corpora are extremely valuable
  - enhancing the translation skills of trainee translators
  - refining their contrastive knowledge of both the source and target language. (Laviosa, 2003: 105-117; Zanettin, 2002; Bowker, 2002: 43-76)
  - Retrieving and examining lexical, terminological, phraseological, syntactical and stylistic equivalents (Bowker & Pearson 2002, Laviosa 2011)
  - Improve subject field understanding, correct term choice and fluency along the lines of Bowker's (1998) experimental study.
  - Raise linguistic and cultural awareness (Bernadini 1997, Hunston 2002)

# STUDIES OF TRANSFER/SOURCE LANGUAGE INFLUENCE

- Transfer/ source language influence:
- Mauranen (2004) The distribution of very frequent words in translations is different from their distribution in original texts in the same language
  - Their profiles of deviation is correlated to specific source languages

# TRANSLATION OF PHRASEOLOGICAL UNITS

- Marco Borillo´s (2009) study of phraseology
- Used the English-Catalán subcorpus of the COVALT (Valencian Corpus of Translated Literature) corpus (1.16 million words)
  - Studies the translation of English phraseological units (idioms/fixed expressions) into Catalán.
  - Descriptive studies would also be beneficial:
    - Norwegian lacks dictionaries of idioms
    - A comparative study of idioms would be of great benefit to translators/translation students.

# TRANSLATIONAL STYLISTICS/TRANSLATOR STYLE

- Translator style: “a kind of thumb-print that is expressed in a range of linguistic as well as non-linguistic features” (Baker 2000)
  - Studies translated text from the target text persepctive
  - Establishes patterns in several translations by the same translators
  - Filters out the variables that may affect these patterns
  - Which can be attributed to the translator?
- (Malmkjær 2003, Munday 2008, Saldanha 2011)

# USES FOR PROFESSIONAL TRANSLATORS - APPLICATIONS

- ”From the professional translator’s point of view, parallel corpora are seen as a multi-purpose resource with enormous potential for assisting the translation activity” (Lagudaki 2007)
  - Compare their own translation solutions to those of published translators
  - extract bilingual terminology with the help of term extraction tools, and build specialised glossaries
  - can be used to validate the translator’s choices, by cross-checking them with word frequency data or by being able to see the chosen terms in a variety of contexts simultaneously (Lagoudaki 2007)

# USES FOR PROFESSIONAL TRANSLATORS (CONT.)

- Check for domain-specific terminology in the two languages in order to maintain consistency
- for acquiring/cultivating linguistic or subject-specific knowledge
- ”In fact, such corpora give the translator the ideal opportunity to discover underlying patterns for the two working languages and to obtain an X-ray of all kind of equivalences (semantic, stylistic, structural, pragmatic) that exist between two parallel texts” (Lagoudaki 2007).

# LEXICOGRAPHY

- Bilingual lexicography is one of he most obvious applications of cross-linguistic research (Johansson 2007)
- Lexicographers regard parallel corpora as the means to discover semantic lexical information in an effort to build multilingual lexicons (Sahlgren & Karlgren, 2005; Lenci et al., 2000; Grefenstette, 1999)
- The field of lexicography has gained substantially from the development of new corpora and tools for monitoring ongoing language development (Atkins and Rundell 2008; Grefenstette 2002; Ooi 1998; Pulcini 2008 Andersen 2013).

# USES IN NATURAL LANGUAGE PROCESSING

- Parallel corpora aligned at sentence level are very useful for NLP purposes
- Parallel corpora are the input for both example-based (Somers 2001) and statistical machine translation systems (SMT) (Mihalcea & Simard, 2005; Koehn 2005, Tou Ng et al., 2003; Menezes and Richardson, 2001)
  - is used to train machine translations systems like the EU-financed MOSES (http://www.statmt.org/moses/)
  - Manually checked corpora represent the gold standard

# WORD SENSE DISAMBIGUATION

- If a lexical item in one language is ambiguous, and each meaning corresponds to a distinct lexical item in another language, we can use use parallel corpora to extract a large number of "disambiguated examples" in the first language (Resnik and Yarowsky 2007)
- The disambiguated examples constitute a corpus aligned at word level, which in turn can be as input in a statistical Word Sense Disambiguation too which can be be used on monolingual text (Resnik and Yarowsky 2007, Kilgarriff 1998)

# OTHER USES OF PARALLEL CORPORA IN NLP

- Bilingual vocabulary extraction: multi-word compounds (compound nouns) (Somers 2001, Parra, Ahrenberg et. al 1998)
- Anafora resolution: the problem of resolving what a cohesive element (i.e. a pronoun or a noun phrase) refers to.
- Bilingual Lexicon Induction
- Temporal Expression Recognition
- Terminology Extraction

# OTHER POSSIBLE USES FOR PARALLEL CORPORA

- Discourse analysis (Baker 2006)
- Research on ideology in corpus based translation studies
  - Munday (2002) studied three translations of García Márquez article about the Cuban six year old boy Elián Gonzalez
- Lexical priming ”the psycholinguistic phenomenon whereby speakers have been observed to understand a given word more quickly if it is presented immediately after a semantically related word, rather than an unrelated word” (McEnery and Hardie 2012, Hoey 2005)
  - The NSPC can bes used for the extraction of examples for both monolingual (Norwegian) and bilingual lexical priming

# COMPARABLE PARALLEL CORPORA

- 2 or more parallel/ translation-corpora specifically created to be comparable

- “A comparable corpus is defined as “original text in each language, matched as far as possible in terms of text type, subject matter and communicative function” (Altenberg and Granger 2002: 7-8). The P-ACTRES and the NSPC are comparable parallel corpora, that is, they contain original text in one language and their translation into one other language, and they are matched as far as possible in terms of text type, subject matter and communicative function” (Hareide 2014: 209)

# WHAT MAKES CORPORA COMPARABLE?

**Sampling frame:**

- time frame of sampling,
- the language variety sampled
- genres/text types represented
- specialized vs. non-specialized language
- original vs. translated language
- size (?)

**A corpus is only as good as its sampling frame**

"products of human sensibility, connected with human interests and self-interests. All the more reason, therefore, to consider the objects of study, the data gathered in the databases, and the parameters defining the corpora themselves very carefully" (Tymockzo 1998: 3).

# WHY ARE COMPARABLE PARALLEL CORPORA IMPORTANT? (1)

- allow for the testing of the same hypothesis on at least two different language pairs.
  - in 2005 Rabadán called for studies in other languages than English, and for transparency in methodology so studies can be replicated with other data sets (Rabadán 2005: 165– 166).

  - Even today most parallel corpora are compiled to solve specific research questions.
  - This often means that the research done on these corpora is hard to replicate on other language pairs
  - in a strictly scientific sense, the value of these studies is limited
  - If we want CBTS to be considered a science, we must comply with the basic requirements of science.

# WHY ARE COMPARABLE PARALLEL CORPORA IMPORTANT? (2)

- allow for the testing of complex hypotheses like Halverson's Gravitational Pull hypothesis (Halverson 2003, 2007, 2009, 2010).
  - Full testing requires data from two comparable parallel corpora with the same target language.
  - relate to each other differently with respect to perceived patterns of linkage in the bilingual's mental lexicon
    - test item must be a unique item in one language pair,
    - Must exist in both languages in the other (Hareide 2014, 2017)

# PARALLEL CORPORA VS. TRANSLATION MEMORY

- Parallel corpus:
  - A searchable database of translation choices
  - Public/shared

- Translation memory:
  - a linguistic database that captures translations as you work for future use
    www.sdltrados.com/solutions/translation-memory/
  - Personal tool

# THE NORWEGIAN-SPANISH PARALLEL CORPUS (NSPC)



# SAMPLING CHALLENGE – THE NSPC

- Designed to be comparable to verson 1 of the English-Spanish P-ACTRES corpus compiled at the University of León as described in Rabadán (2005, 2005 - 2008, 2007) Izquierdo, Hofland and Reigem (2008) and Rabadán, Labrador and Ramón (2009).

- A freestanding parallel unidirectional translation corpus of published contemporary Norwegian original texts and their published translations into Spanish

# NSPC –VITAL DATA

- http://clarino.uib.no/korpuskel/clarino-metadata?session-id=24178469476927 6&corpus=nspc/nor&default-corpus=nspc/nor&resource=nspc/nor

- Version 1 described in Hareide and Hofland (2012) consists of 3.01 million running words, originating from 31 entire text pairs (21 volumes) consisting of mainly books.
  - In 2013 the corpus was expanded to include the entire population of 41 text pairs and now consists of 4.1 million words.
- Language varieties sampled:
  - Norwegian written standards: Nynorsk and Bokmål
  - Spanish: Iberian Spanish (Castellano)

# NSPC –VITAL DATA (2)

- Designed to be a parallel unidirectional translation corpus of contemporary Norwegian texts and their published Spanish translations

- Version 1 was designed to be roughly comparable to version 1 of the English-Spanish P-ACTRES corpus compiled at the Universidad de León described in Rabadán (2005, 2005 - 2008, 2007), Izquierdo, Hofland and Reigem (2008) and Rabadán, Labrador and Ramón (2009).
  - Both corpora have later been expanded

- The NSPC consists exclusively of published written material
- Text formats: books, brochures and booklets are represented in the target population, as well as collections of texts by several authors.
- The corpus is compiled for research in Linguistics and Translation Studies, and is therefore coded for translation relevant meta-data.

# ENSURING REPRESENTATIVENESS

- Traditionally one would want as wide a variety as possible of text genres.
- Biber: the most important factor in ensuring representativeness is that the population the texts are to be drawn from is well defined, and that the entire variability of this population is represented in the corpus (1993: 243).

- Additional problem with parallel corpora:
  - not all genres are translated between all languages (McEnery, Tono & Xiao 2006: 95)
  - "translations are heavily biased towards certain genres but these biases are rarely symmetrical for any language pair" (Mauranen 2004: 74; Biber 1993).
- Easy to find translated books, whereas translated journal and magazine articles are not so easily found in some language pairs (Zanettin 2000: 108; Izquierdo, Hofland & Reigem 2008b: 36) even between major languages such as English-European Spanish and Italian-English.

# TEXT SELECTION

- I set rigorous external criteria for defining the population and ensuring representativeness.
- Exploration: The Norwegian National Library *Norbok* database of texts by Norwegian authors published abroad.
- A query for texts originally written in Norwegian and published in Spanish between 1 January 2000 and 31 December 2009.
  - This query returned 67 volumes, which included information booklets, children's literature, novels, poetry, theatre plays, cultural history, science
- Indeed very difficult to find journal and magazine articles originally written in Norwegian and translated into Spanish

# CRITERIA FOR TEXT SELECTION

The target population of the corpus was operationalized as all texts in the *Norbok* database meeting the following criteria:

1. List texts originally published in Norwegian and translated into Spanish in the time- period 2000 to 2009.
2. Remove all texts that are not verified translations in accordance with Toury's definition of translation "presented or regarded as such within the target culture, on whatever grounds" (Toury 1995: 20).
3. Exclude all texts not translated by translators operating under their full name in accordance with criteria set by Halverson (1998:500) and Laviosa (1997a)
4. Include all genres represented in the population in accordance with Biber's advice to include the entire variability of the population. (1993: 243).
5. Include source texts in both the Norwegian written standards in accordance with Biber's principle of preserving the full variability of the population (1993: 243).

# METADATA AND CLASSIFICATION OF THE TEXTS

- De Sutter et al. (2012: 326) criticize research in Corpus-based Translation Studies for not adhering rigorously enough to the corpus linguistic methodology.
  - claim that researchers in Corpus-based Translation Studies tend to study only one variable, thereby demonstrating insensitivity to other factors such as text type and source- language variation that may influence the variable studied or have explanatory power.

The NSPC can be divided into sub-corpora based on parameters that could represent possible variables in a corpus-based study
  - allows for testing of multiple variables as suggested by de Sutter et al.
- Genre/text strata
- Author's gender
- translator's gender
- Translator's mother tongue
- Norwegian written standard
- Year of publication (original and translation)
  - See Jenset and Hareide (2013) for an example case. (https://boap.uib.no/index.php/bells/article/view/356)

# NSPC 1 – COMPARED TO THE P-ACTRES 1

- I have compiled the NSPC to be as maximally comparable to version 1 of the English-Spanish P-ACTRES corpus with regard to:
  - time frame of sampling,
  - the language variety sampled (European Spanish),
  - genres represented,
  - size
  - and a focus on non-specialized language “directed at the general public” (Rabadán 2005: 161).

# SUBDIVISION INTO SUBCORPORA

- The comparability is further enhanced by allowing for the subdivision of the NSPC into sub-corpora, a strategy that has already been successfully employed by the P- ACTRES corpus (Rabadán 2005; Izquierdo, Hofland & Reigem 2008: 36).
- This means that the books section of the P-ACTRES is comparable to the NSPC's book- section and both corpora are comparable to the books section of the Spanish CREA reference corpus.
- The CREA also allows for specified queries based on genre, geography and chronology, where the researcher enters only the desired variables in the query.
- For an account of the P-ACTRES' comparability to the CREA, see Rabadán (2005: 162). We believe that this procedure ensures the representativeness of both the P- ACTRES and the NSPC while at the same time ensuring their comparability.

# DIFFERENCES BETWEEN THE P-ACTRES AND THE NSPC

## P-ACTRES

- Press editorials
- "Ephemera": short translated texts such as product sheets
- P-ACTRES had to use text fragments due to legal restrictions (copyright regulations) (Rabadán 2008: 110)

## NSPC

- No press editorials, but contain journalism
- Ephemera fall outside the population of the NSPC as they are not translated by translators operating under their full names.
- The NSPC is based entirely entirely on full texts in accordance with the recommendations of Sinclair (1991: 19), and Johansson (p.c. 16.10.2008). The NSPC was originally conceived as a tool for the study of translation, where the object of study is the full text (Zanettin 2000: 107), a view which is supported by Laviosa (1997a: 296) and Baker (1995: 240).

# COMPARABILITY WITH THE P-ACTRES AND CREA (VERSION 1 OF NSPC AND P-ACTRES)

| Corpora | NSPC | P-ACTRES | CREA |
|---|---|---|---|
| Number of words | 3.01 million | 2.5 million | 160 million (2008)[32] |
| Full texts/exerpts | Full texts | exerpts | Full texts |
| Language variety | European Spanish | European Spanish | European Spanish subcorpus |
| Genres | | | |
| Fiction | X | X | X |
| Non fiction | X | X | X |
| Popular science | X | X | X |
| Journalism/editorials | X | X | X |
| Childrens literature | X | X | ?[33] |
| Magazine articles/press | 0 | X | X |
| Theatre | 0 | 0 | X |
| Ephemera | 0 | X | X |

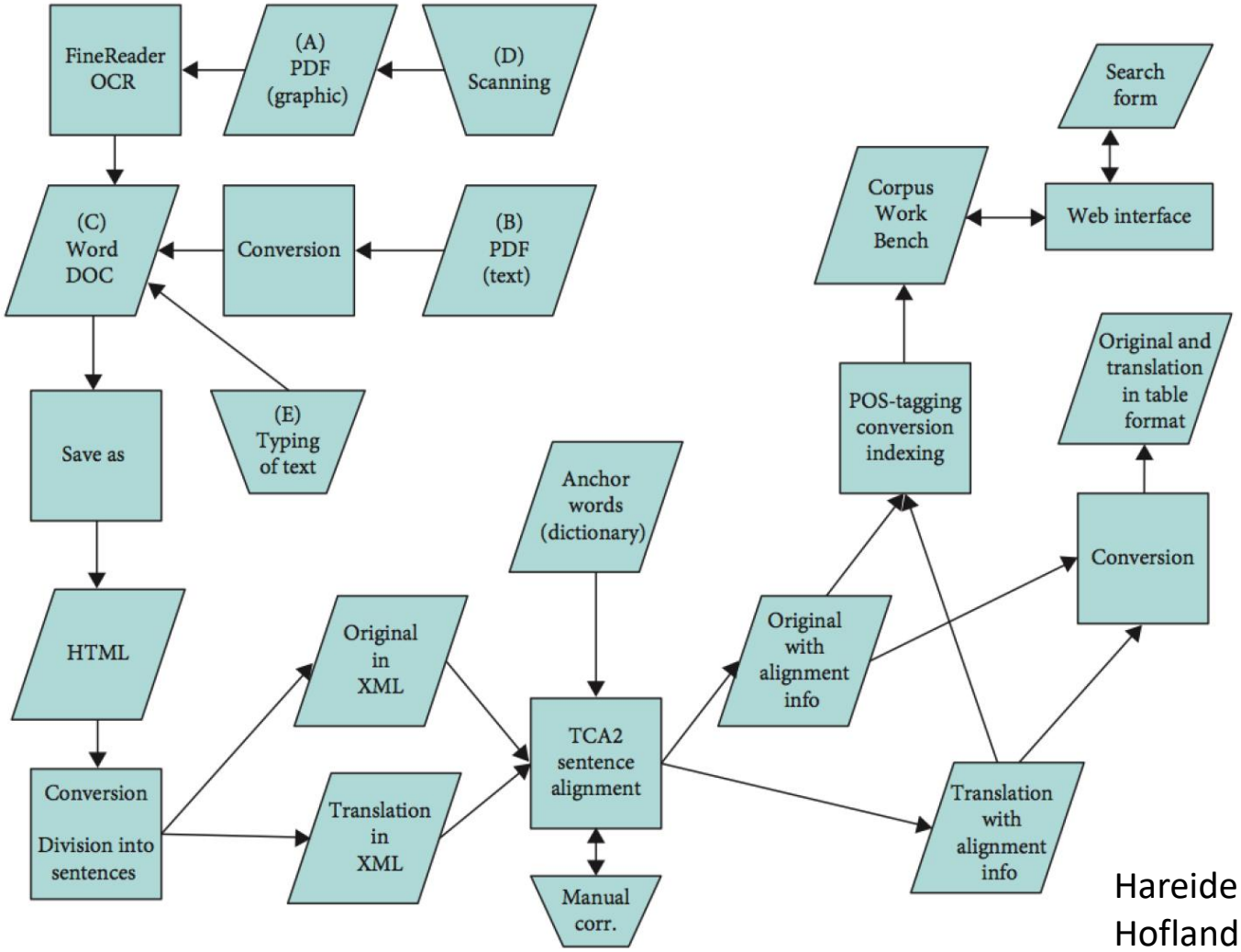
Based on Rabadán (2005: 162)

# MEASURES TAKEN TO CREATE A FREELY CORPUS

1. Written consent from all publishing houses and all authors of the original texts
   - Carefully worked out a letter of agreement with the help of experts in this field
     - Law department at Aschehoug (one of the largest publishing companies in Norway)
     - Head of Uni Digital at the University of Bergen
   - Phoned all the editorials, talked them trough the agreement
   - Mailed it to them as we spoke
     - The editorials contacted the authors
   - Normally I got an answer within a week.

2. Contacted all the Spanish editorials
3. Contacted *Centro Español de Derechos Reprográficos* (CEDRO) Ramón Gavidia Saelices.
   - under art. 32, *Ley de propiedad intelectual,* consent from authors and publishing houses is not neccessary for use in investigation, as long as bibliographical information is provided.
4. Renegotiated the contracts for CLARIN.NO
5. Contacted CEDRO again, art. 32 Ley de propiedad intelectual applies.

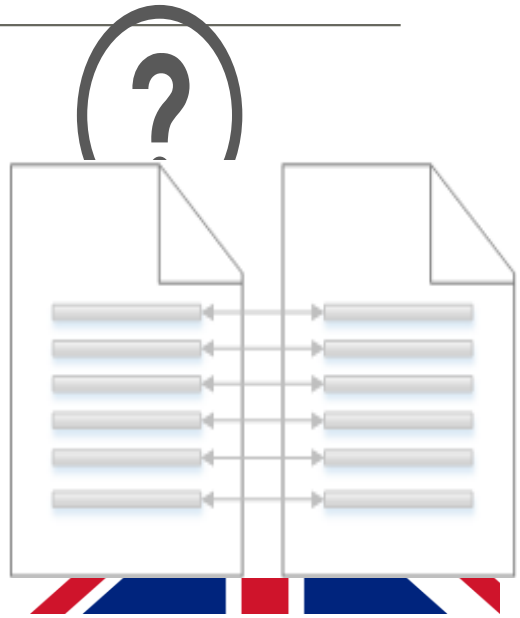
# THE STEPS FROM TEXT TO CORPUS (NSPC)




Hareide and Hofland (2012)

# TECHNICAL LIMITATIONS OF CORPUS ANALYSIS SOFTWARE

- Critical activities such as **aligning** and **tagging** must be carried out by technical specialist.
- **Multiplicity** of software resources.
- Preponderance of **English** language.




Linguist

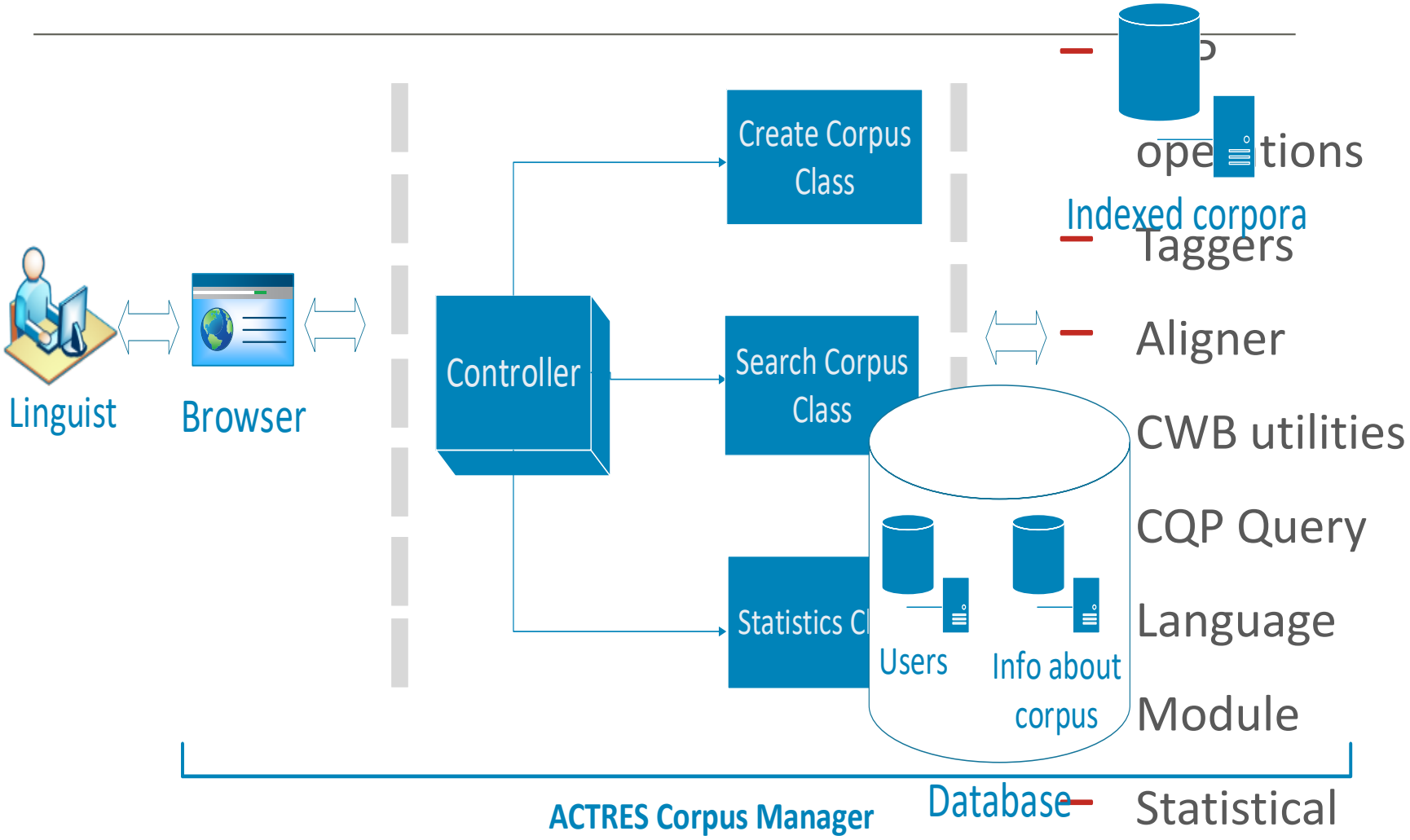
# ACTRES CORPUS MANAGER (ACM)

Video Demo:
http://actres.unileon.es/ACM/videodemo/

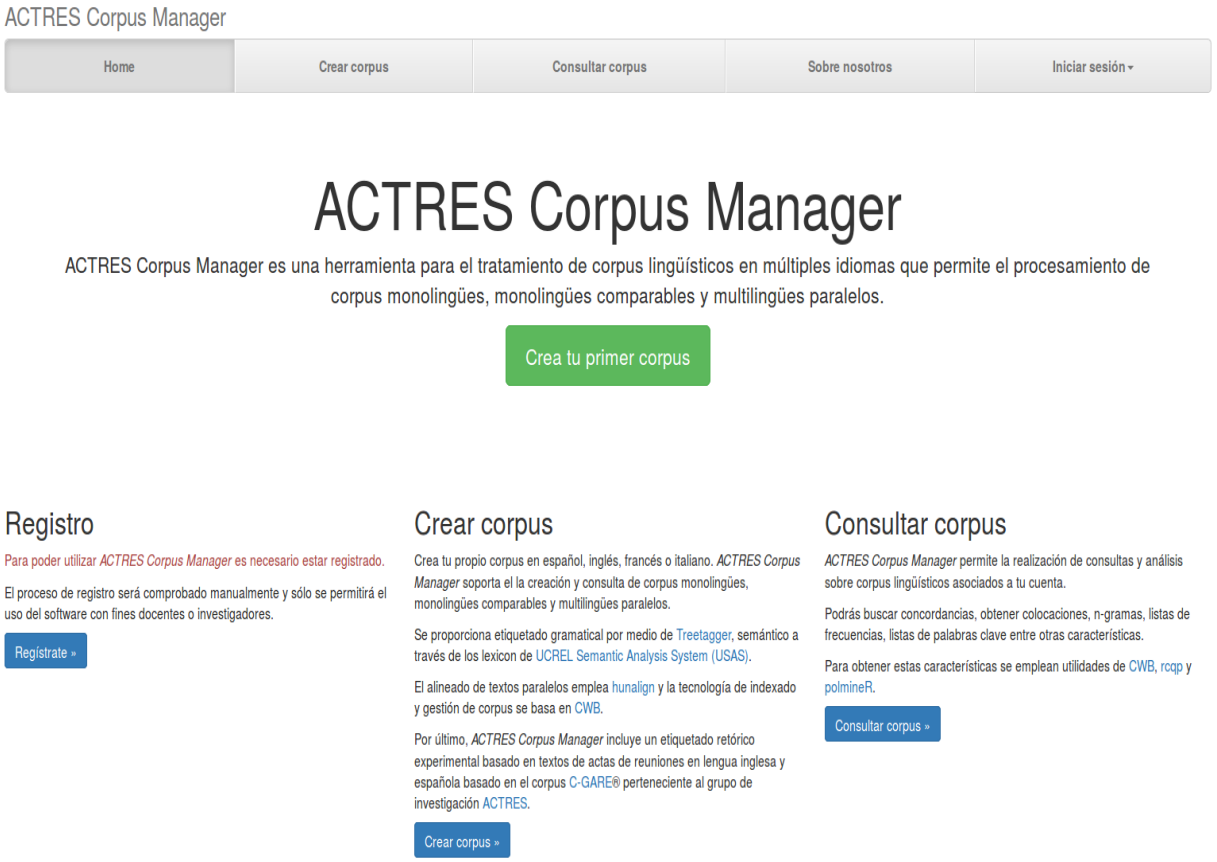
# ACTRES CORPUS MANAGER (ACM)

- It allows users:
  - to create their own corpora (monolingual, bi/multilingual parallel and comparable) with several annotation layers (grammatical, semantic and rhetorical),
  - make linguistic queries,
  - and obtain the most common statistics without technical assistance during the process and regardless of the technical skills of the user.
- It takes usability from a linguistic user point of view as a key factor based.

# ACM: OVERALL ARCHITECTURE



# ACM: CREATE CORPUS

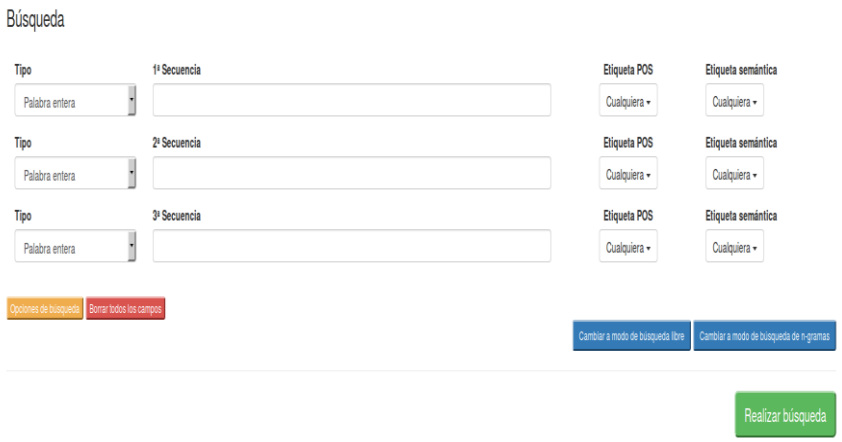


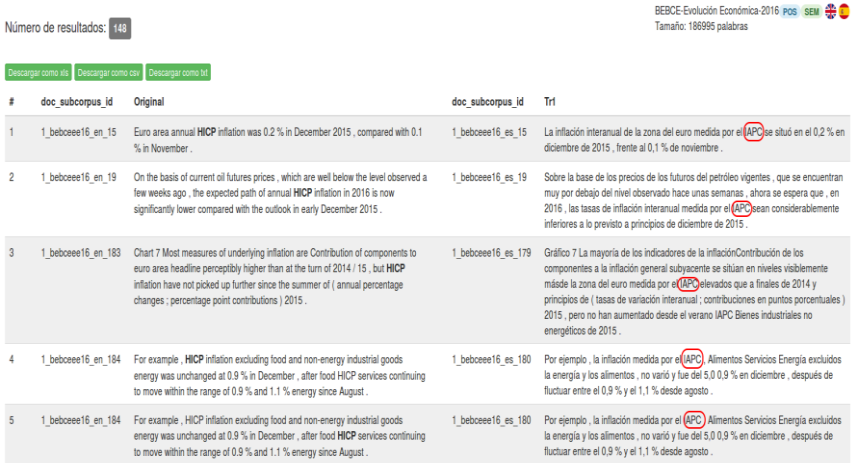

universidad de león

ACTRES

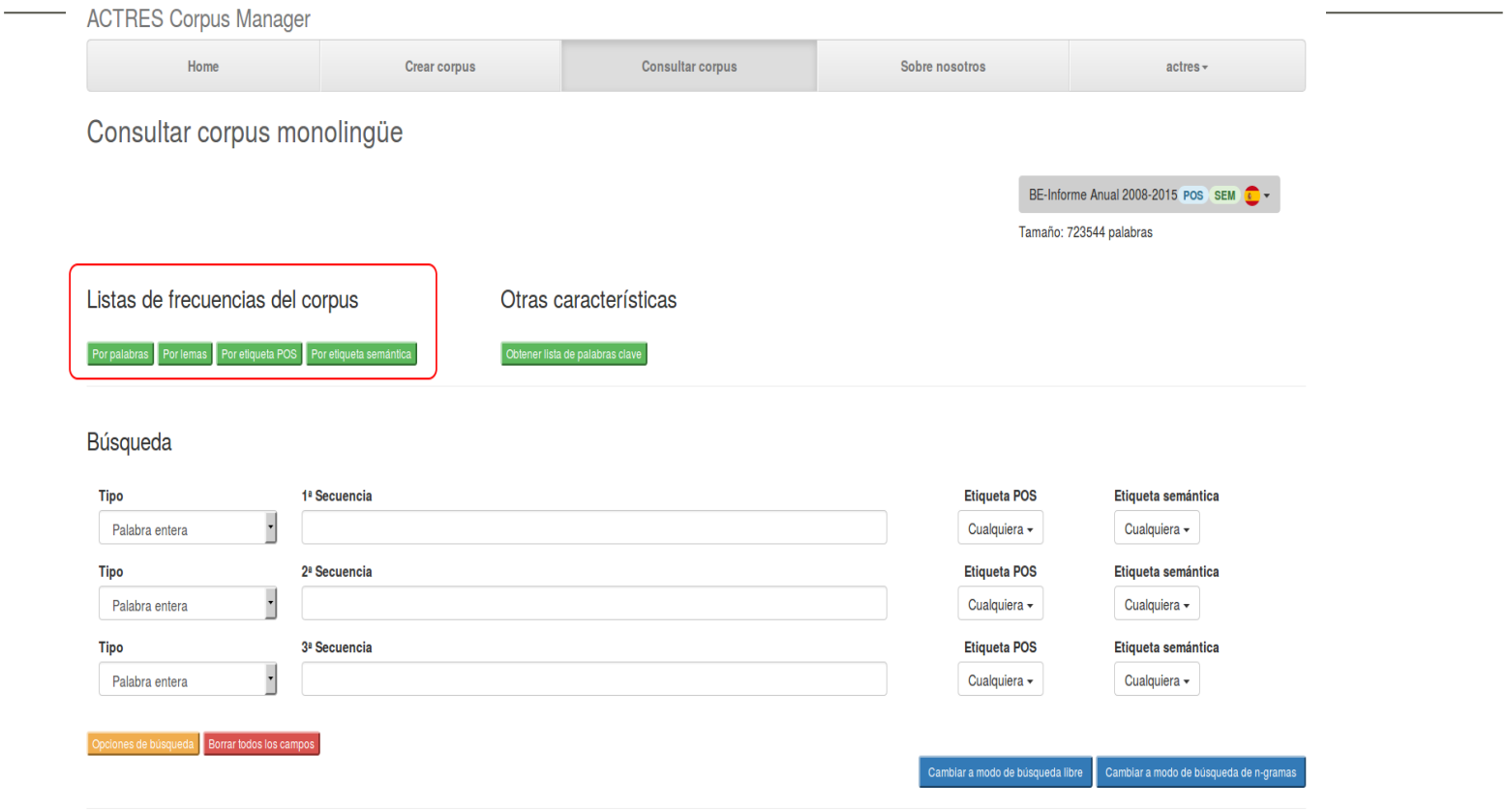
MØREFORSKING

# ACM: SEARCH CORPUS



# ACM: GET STATISTICS



# ACM: RELEVANT FEATURES FOR THE PROJECT

## Corpora

- Multilingual parallel corpora.
- Medium-large corpora size support.
- POS and semantic annotation.

## Search

- Custom user-friendly interface.
- Semi-assisted concordance search.
- Mapping between different grammatical tagsets.

## Statistics

- Collocations
- Frequency lists.
- Keywords.
- Word n-grams.
- Excel format.

# ACM: PRESENT LIMITATIONS FOR THE PROJECT

- Support 4 languages.
- Support 1 original corpus + 3 corresponding translations.
- Statistic are based on subcorpus, there are not specific parallel statistics.
- Texts of the corpus must be in plain text format.