Studying terminology and variation in specialised texts: corpus requirements and tools

Sabela Fernández Silva

Koen Kerremans

IULATERM
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra

Centrum voor Vaktaal en Communicatie
Department of Applied Linguistics
Erasmushogeschool Brussel

Aims

CORPUS LINGUISTICS	TERMINOLOGY RESEARCH	TRANSLATION STUDIES

1. Present a joint pilot project in which:
   - **terminology** and variation is studied in a
   - **corpus** of specialised texts (Ga) and their
   - **translations** (En)

2. Use this pilot project to reflect on corpus requirements (and tools)

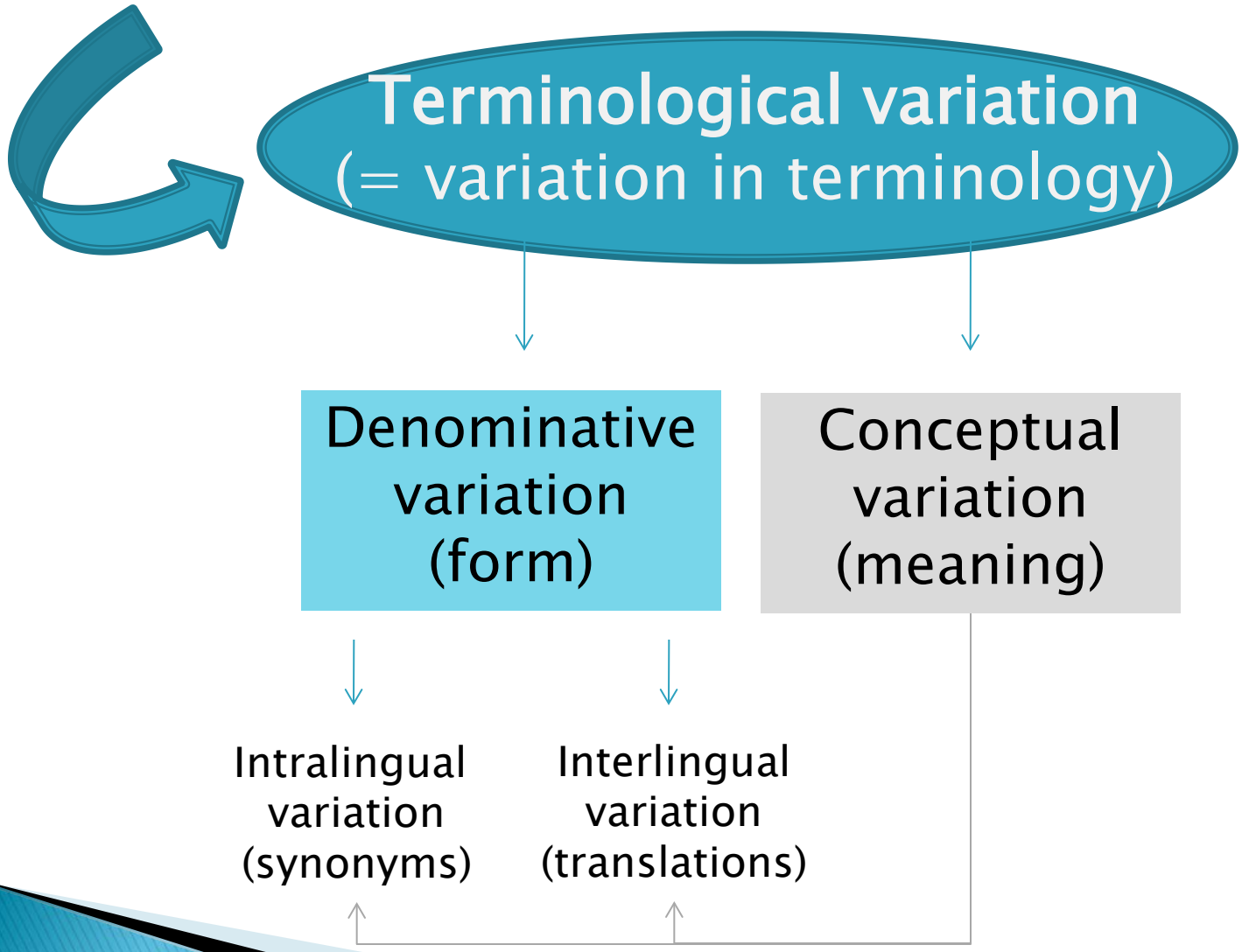
Outline

1. **Variation as object of study**
   1.1 … in terminology research
   1.2 … in a project of IULATERM
   1.3 … in a project of CVC

2. **Description of the joint pilot project**
   2.1 Methodology: corpus, object of analysis, tools
   2.2 (Intermediate) results

3. **Corpus requirements and tools**

1. Variation as object of study



1.1. Variation in terminology research

- **Prescriptive perspective** (e.g. Wüster 1979; Felber 1984)
  - Variation hampers specialised communication

- **Descriptive perspective** (e.g. Gambier 1991; Cabré 1999; Temmerman 2000)
  - Variation is functional

1.1. Variation in terminology research

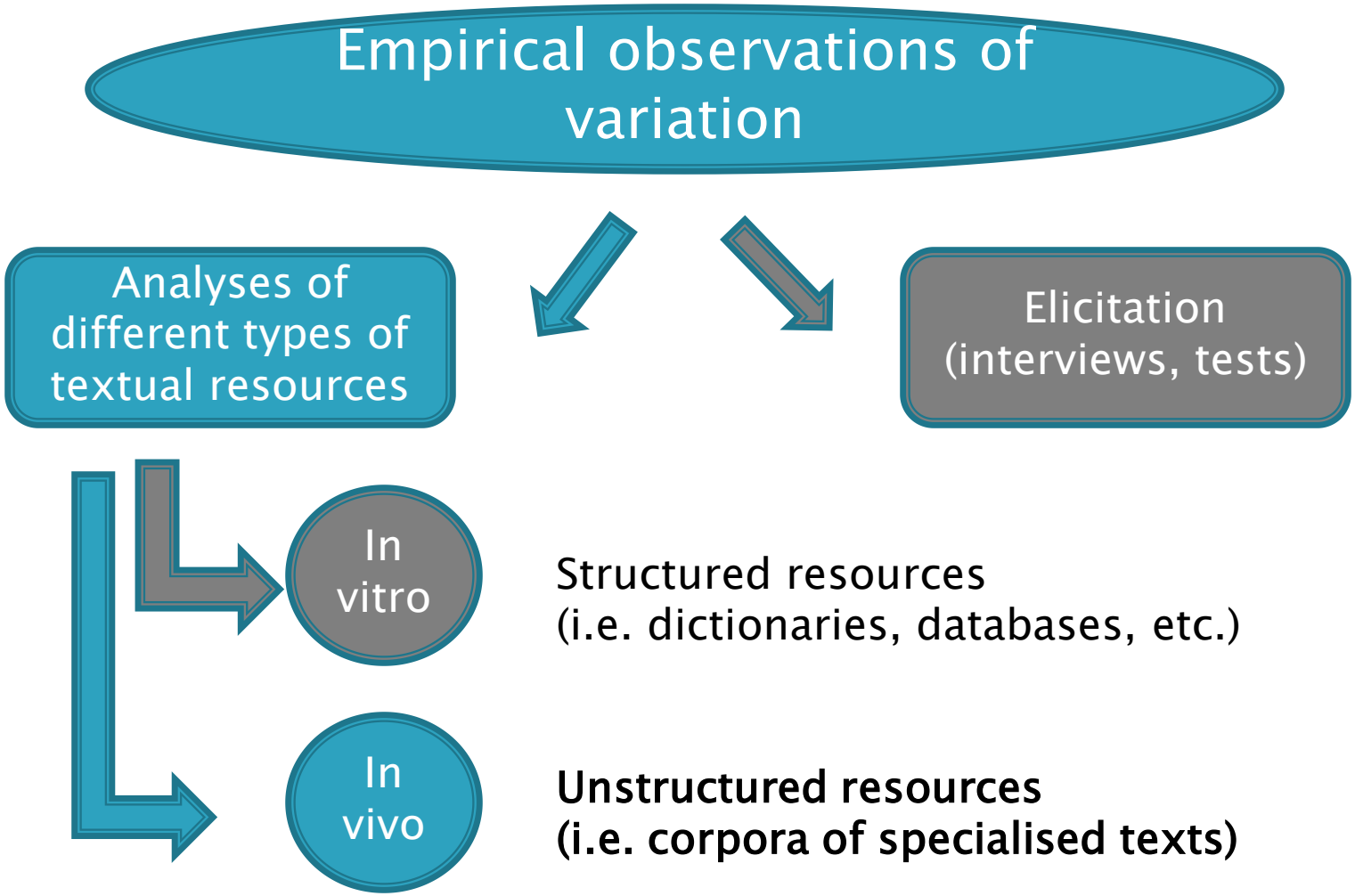
- **Consequence of descriptive view**: increased interest in the study of variation in special language

- **Possible causes** (cf. Freixa 2006):
  - Preliminary: linguistic redundancy
  - Dialectical: different origins of authors
  - Functional: different communicative registers
  - Discursive: different expressive needs
  - Interlinguistic: contact between languages
  - Cognitive: different conceptualisations

1.1. Variation in terminology research

- **Fundamental, theoretical research:**
  - Typologies: e.g. Freixa (2002), Faulstich (2002)
  - Reasons:
    - Cognitive perspective: e.g. Resche (1999), Temmerman (2000)
    - Communicative perspective: e.g. Domènech (2006); Seghezzi (2006); Faulstich (2002)
    - Discursive perspective: e.g. Suárez (2004), Ciapuscio (1999)
    - Diachronic perspective: e.g. Dury (1999)
  - ...
- **Applied research:**
  - Lexicographic representation: e.g. Collet (2004)
  - Automatic detection: e.g. Daille (2003)
  - Text clustering: e.g. Ibewke-SanJuan & SanJuan (2004)
  - ...

1.1. Variation in terminology research



1.2. IULATERM's project about variation

denominative variation with cognitive consequences: motivation, structure and behaviour in context

Iulaterm's traditional research track
(Freixa 2002, Suárez 2004, Domènech 2006)

Funding:
Spanish Ministry of Education and Science (2007-2010)

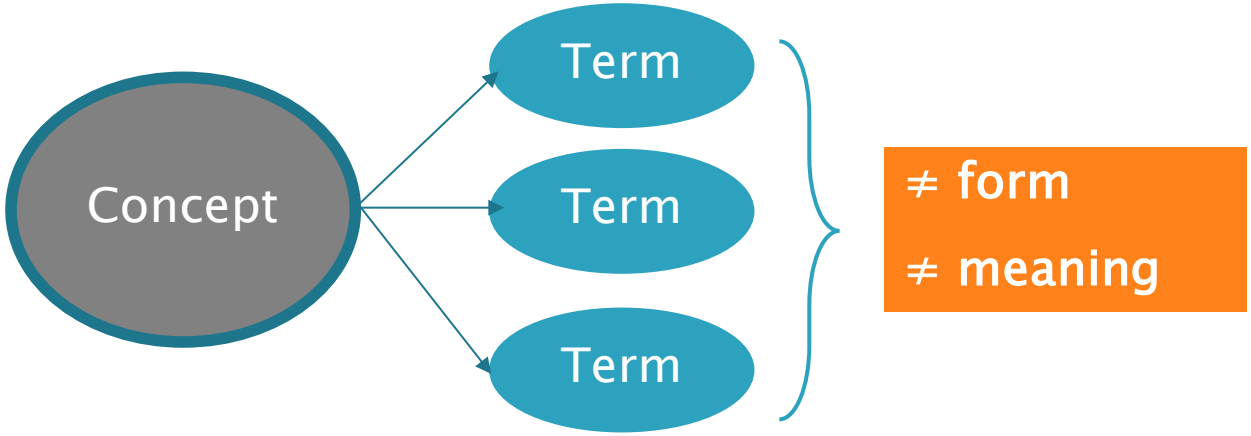
Supervisors:
Prof. Dr. Teresa Cabré (IULA)
Prof. Dr. Judit Freixa (IULA)

1.2. IULATERM's project about variation

What?

Why?

How?



- Different variants highlight different facets of the concept (e.g. reusable bottle vs returnable bottle)
- This affects the way the recipient understands the concept (i.e. denominative variation with cognitive consequences)

What are the motivations behind term choice? To which factors can it be correlated?

1.2. IULATERM's project about variation

What?

Why?

How?

Term choice is determined by conceptual characteristics and contextual features (cognitive, communicative, discursive context)

"terminology tends towards stronger systematization of its internal structure [...] at the same time it tends towards using the full flexibility of natural language"

(Kageura 2002: 15)

- Increase insight on the interaction between the conceptual realm and the linguistic realm
- Confirm the functionality of denominative variation in special communication

1.2. IULATERM's project about variation

What?

Why?

How?

- Compilation of a bilingual, comparable corpus (Galician, French)
  - Domain: Fishing
  - Diversity in perspectives, text types and levels of specialisation
- Manual and semi-automatic extraction of denominative variants
- Semantic analysis of multi-variant clusters
- Comparison of denominative tendencies

≠ concept classes
≠ taxonomical levels

Conceptual characteristics

≠ perspectives
≠ levels of specialisation

Contextual features

1.3. CVC's project about variation

"Terminological variation: a contrastive, multilingual, multidimensional, text-based analysis"

Funding:

UAB (2008 – 2011)

Supervisors:

Prof. dr. Madeline Lutjeharms (VUB)

Prof. dr. Rita Temmerman (EhB)

1.3. CVC's project about variation

What?

Why?

How?

- (How) is terminological variation in the source text reflected in the translation?

More variable use of terms in translation

Same variable use of terms in translation

More consistent use of terms in translation

- What patterns of variation can we discern in specialised translations? (cf. Rogers 2007)
- Are specialised translations characterised by more consistent use of terms? (cf. Merkel 1996)

1.3. CVC's project about variation

What?

Why?

How?

- Ignoring variation in specialised translation is sometimes problematic (see e.g. Resche 2004)

"translators may actually over-standardize, creating consistency in places where the use of variants was deliberate and well reasoned."

(Bowker and Hawkins 2006:80)

- Increase insight in specialised translation
- Optimise specialised translation dictionaries, translation processes, specialised translation courses, …

1.3. CVC's project about variation

What?

Why?

How?

- compilation of a parallel corpus (i.e. source texts and translations), taking into account different parameters:
  - ✓ languages:
    - ❑ SL (En) ↦ TL (Fr and Nl)
    - ❑ SL (Fr) ↦ TL (En and Nl)
  - ✓ degree of specialisation, text type, number of revisions, topic, …
- analysis of lexicalisations of a given unit of understanding, by following lexical chains (cf. Rogers 2007)

2. Description of the joint pilot project

- **Aims**:
  1. Share experience concerning variation research
  2. Reflect on practical issues (e.g. corpus design)
  3. Fine-tune methods and research questions by means of a small case study
     - Texts: 3 texts (Ga -> En) written by field experts and published in international journals (high degree of specialisation)
     - Topic: economic consequences of oil spill (Prestige) in the Galician fishing sector

2.1 Methodology

A) Corpus compilation	•*Pdf to txt* •*Aligner*
B) Manual term analysis in SL	*Spreadsheet*
C) Context retrieval (HTML output)	*Perl script*
D) Manual selection of terms in TL	*Text editor*
E) Calculation of variants in TL	*Perl script*

2.1 Methodology

E) Automatic counting of translation equivalents

Context Retriever

- bacallaos
- bancos de moluscos
- bancos marisqueiros
- bateas
- bateas de mexillóns
- bateas mexilloeiras
- benceno
- berberecho
- bio-diversidade
- biodiversidade
- biomasa
- biorremedicaci
- bivalvos
- bloqueo dos quimiorrecepto
- boca
- boga
- bogavantes
- breas
- buque

CLUSTER 122

SOURCE TERM	TRANSLATION	REFERENCE
bancos de moluscos	mollusc banks (2)	García-Negro(2004)_gl%20(Galician)(English).txt.html
bancos marisqueiros	. (1)	DeathCoast_(Galician)(English).txt.html
bancos marisqueiros	shellfish grounds (1)	DeathCoast_(Galician)(English).txt.html
bancos marisqueiros	shellfish banks (3)	García-Negro(2004)_gl%20(Galician)(English).txt.html

Lemma: "bancos de moluscos"

Occurrence: "bancos de moluscos"

(source) • Valoración das perdas económicas ocasionadas nos **bancos de moluscos** na zona afectada, a partires dos mostreos realizados pola Conselleria de Pesca, Marisqueo e Acuicultura: Decembro 1992 e Febreiro 1993.
• Valuation of the economic losses occurring in mollusc banks in the affected zone, based on samplings done by the Department of Fishing, Shellfish Harvesting and Aquaculture: December, 1992, and February, 1993.

2.2 (Intermediate) results

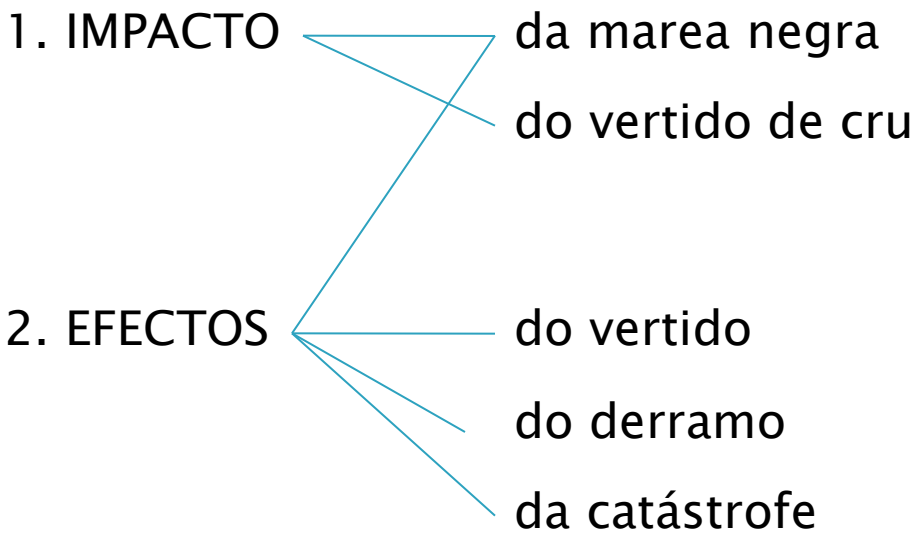
Texts	Cat. 1: Var (SL) < Var (TL)	Cat. 2: Var (SL) = Var (TL)	Cat. 3: Var (SL) > Var (TL)
DC (38 clusters)	28 (73,7%)	9 (23,7%)	1 (2,6%)
GN (38 clusters)	14 (36,8%)	20 (52,6%)	4 (10,5%)
SG (34 clusters)	16 (47,1%)	15 (44,1%)	3 (8,8%)

<u>Further qualitative analysis:</u>

- Are patterns of variation in the SL reflected in the TL ? (-> Cat. 2 is only a first indication)
- What concepts tend to have more lexical variants?

2.2 (Intermediate) results

CLUSTER 1
SOURCE TERM
impacto da marea negra
impacto da marea negra
efectos do vertido
efectos do vertido
efectos da marea negra
efeitos da catástrofe
impacto do vertido de cru
efectos do derramo
efectos da catástrofe



2.2 (Intermediate) results

CLUSTER 1		
SOURCE TERM	TRANSLATION	
impacto da marea negra	oil spill (1)	DC
impacto da marea negra	ecological catastrophe (1)	DC
efectos do vertido	spill (1)	DC
efectos do vertido	effects of the disaster (1)	DC
efectos da marea negra	oil spill effects (1)	DC
efeitos da catástrofe	effects of the catastrophe (2)	GN
impacto do vertido de cru	impact of the oil spill (1)	SG
efectos do derramo	effects of the spill (1)	SG
efectos da catástrofe	effects of the catastrophe (1)	SG

3. Corpus requirements and tools

Criteria for selecting texts?

Aim: comparability of texts

Language | *Topic* | *Text type* | *Source* | *Perspective*

derived
from text analysis

3. Corpus requirements and tools

These criteria allow us to explain reasons for term choice

Language | *Topic* | *Text type* | *Source* | *Perspective*

Genesis of texts? ⟷ **Extra information** ⟷ *People involved?*

Resources used during writing/translating?

- dictionaries, documents, standardised term lists, …

Knowledge of writer/translator/revisor?

- (non-)expert

Tools used during writing/translating?

- translation memories, technical writing tools, …

Native language of writer/translator/revisor?

- influence of native language on term choice?

3. Corpus requirements and tools

- Many large (parallel) corpora are available today:
  - e.g. **Corpus Tècnic de l'IULA** (http://www.iula.upf.edu/corpus/corpusuk.htm)
    - Texts in 5 languages belonging to 5 domains
    - Annotation: structural & morphosyntactic
    - **Extra information**: subject field specification, documentary information

- Corpus descriptions often focus on formal criteria:
  - "The ET10-63 Corpus [...] is part-of-speech tagged and also lemmatized." (http://www.clres.com/paracorp.html)
  - "Each text extract amounts to 10,000-15,000 running words [...]." (http://www.hf.uio.no/ilos/OMC/English/Subcorpora.html#En-Du)

3. Corpus requirements and tools

- **Usefulness of a workbench supporting the compilation, management and querying of (parallel) specialised corpora**
  - Term extraction
  - Alignment of translation segments in more than 2 languages (>< Wordsmith): e.g. ParaConc, MultiConcord
  - Classification of texts (according to different parameters) for advanced querying
  - Database storing
  - Statistical exploitation
- E.g. JAGUAR: workbench for compiling and exploring an LSP corpus from the web (http://jaguar.iula.upf.edu)

4. Conclusion

- A corpus is an interesting point of departure for studying terminological variation
- Availability of information about the text genesis is required to be able to explain term choices and terminological variation
- Refinement of standards describing corpora (e.g. XCES or TEI)? See Burnard 2004 (http://ahds.ac.uk/creating/guides/linguistic-corpora/chapter3.htm#section6)
  - Corpus identification (e.g. owner)
  - **Corpus derivation (description of corpus samples)**
  - Corpus encoding (e.g. XML)

Comments? Questions?

sabela.fernandez@upf.edu
koen.kerremans@ehb.be