Description and representation in language resources of Spanish and English specialized collocations from Free Trade Agreements

Pedro Patiño García
Multilingual Lexical Resources and Tools
UiB, Bergen, 22.06.2011

Definition of collocation:

Palmer (1933)
"a succession of two or more words that must be learned as an integral whole and not pieced together from its component parts."

Firth (1957)
"You shall know a word by the company it keeps".

Pedro Patiño García
Multilingual Lexical Resources and Tools
UiB, Bergen, 22.06.2011

Definition of collocation:

Choueka (1988)
"a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components".

Pedro Patiño García
Multilingual Lexical Resources and Tools
UiB, Bergen, 22.06.2011

Another definition of collocation:

Bartsch (2004)
"lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactic relation with each other".
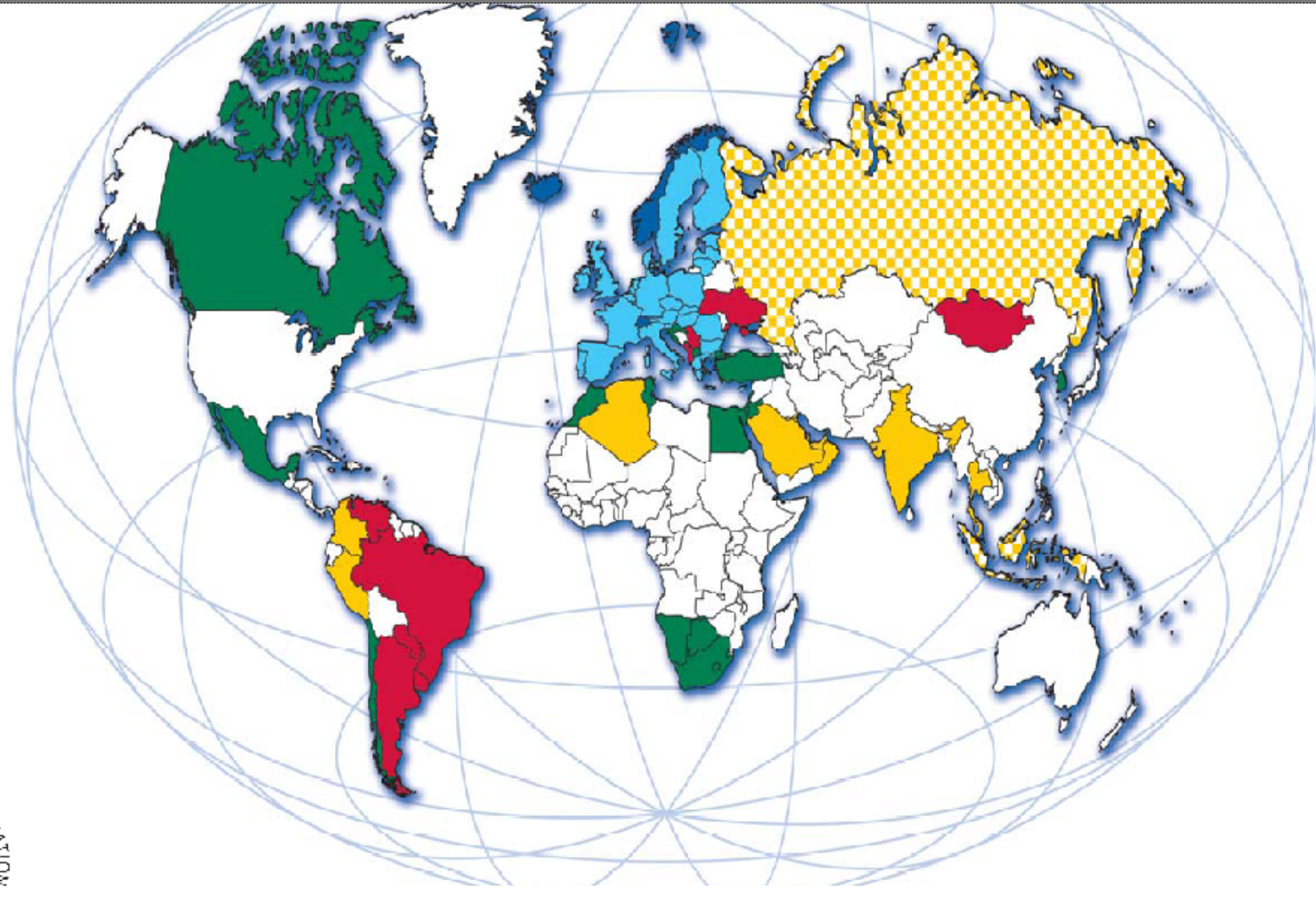
Pedro Patiño García
Multilingual Lexical Resources and Tools
UiB, Bergen, 22.06.2011

Pedro Patiño García
Multilingual Lexical Resources and Tools
UiB, Bergen, 22.06.2011

Research questions (1)

What is the "termhood" of the collocations that appear in FTA texts so that they can be considered as specialized?

How can this termhood be determined from the linguistic features of constituents such as head and modifier of specialized collocations from the subdomain of Free Trade?

Pedro Patiño García
Multilingual Lexical Resources and Tools
UiB, Bergen, 22.06.2011

Research questions (2)

Which grammatical categories are used to identify specialized collocations in the encoding of language resources?

How can these units be represented in term bases and computational lexicons so the data can be interchangeable and interoperable?

Pedro Patiño García
Multilingual Lexical Resources and Tools
UiB, Bergen, 22.06.2011

Objectives

Theoretical objective: **To determine the terminological value of the specialized collocations from the domain of economics that appear in English and Spanish FTA texts.**

Applied objective: **To assess the applicability of linguistic annotation schemes for the representation of specialized collocations in term bases and computational lexicons.**

Pedro Patiño García
Multilingual Lexical Resources and Tools
UiB, Bergen, 22.06.2011

Specific objectives (1)

1. To perform a linguistic classification, description and comparison of FTA specialized collocations as they appear in a corpus of English and Spanish from American and European FTA texts.

2. To compare the behavior of specialized collocations from the area of FTA in general and specialized English and Spanish corpora.

Pedro Patiño García
Multilingual Lexical Resources and Tools
UiB, Bergen, 22.06.2011

Specific objectives (2)

3. To verify the representativeness and inclusion of these collocations in general and specialized English and Spanish dictionaries and term bases.

4. To determine how the constituents of specialized collocations from both languages in the sub-domain of free trade contribute to delineate the domain of economics.

Pedro Patiño García
Multilingual Lexical Resources and Tools
UiB, Bergen, 22.06.2011

Specialized collocation: collocate + base (term):

{impose | extend | maintain | apply | adopt} a safeguard measure

{incrementar | adoptar | aumentar | aplicar | eliminar | reducir} un arancel aduanero

Pedro Patiño García
Multilingual Lexical Resources and Tools
UiB, Bergen, 22.06.2011

Examples:

{adoptar medidas tributarias; adoptar medidas de salvaguardia; aplicar un arancel aduanero}

{apply a customs duty; to effectively enforce its environmental laws; arrive at a mutually satisfactory resolution}

Pedro Patiño García
Multilingual Lexical Resources and Tools
UiB, Bergen, 22.06.2011

Examples:

Collocates of arancel (SketchEngine):
aplicar, imponer , introducir, reducir, adoptar, establecer
Similar words to "aplicación":
adopción, implementación

Pedro Patiño García
Multilingual Lexical Resources and Tools
UiB, Bergen, 20-23.06.2011

Methodology

Corpus-driven and corpus-based approach. English-Spanish parallel corpus of FTA texts (1,3 m words per lang.) being aligned with TCA2.

Specialized bi- monolingual dictionaries and termbases like IATE (free trade, economics)

Reference corpora

NLP software

Pedro Patiño García

Multilingual Lexical Resources and Tools
UiB, Bergen, 22.06.2011

Corpus interface

Corpus	Book/Chapter		
Texts ▾	All ▾	Write book/chapter: [_____]	Example: SN.*|SA1|MH10

	Word Original (English)		Word Translation (Spanish)	
	Whole word ▾	support	Whole word ▾	
	POS	Any ▾	POS	Any ▾
Search / Ignore case ☑ / **RESET FORM**	Stop words		Stop words	
	2nd word			
	Whole word ▾		Whole word ▾	
	POS	Any ▾	POS	Any ▾
	3rd word			
	Whole word ▾		Whole word ▾	
	POS	Any ▾	POS	Any ▾

Pedro Patiño García

Multilingual Lexical Resources and Tools
UiB, Bergen, 22.06.2011

Corpus interface

Economic Community of 9 February 1970, as amended ; (UE_1992e.s2608)	Estados miembros de la Comunidad Económica Europea el 9 de febrero de 1970, en su versión modificada ; (UE_1992s.s2602)
All claims and liabilities arising from the very short-term financing mechanism and the short-term monetary **support** mechanism, under the Agreements referred to in Article 6.1, shall be settled by the first day of the third stage. (UE_1992e.s2770)	Todos los créditos y deudas resultantes del mecanismo de financiación a muy corto plazo y del mecanismo de ayuda económica a corto plazo estipulados con arreglo a los Acuerdos a que se refiere el artículo 6.1 se satisfarán no más tarde del primer día en que entre en vigor la tercera fase. (UE_1992s.s2763)
1. With a view to achieving the objectives of Article 1, the Community shall **support** and complement the activities of the Member States in the following fields: (UE_1992e.s2914)	1. Para la consecución de los objetivos del artículo 1, la Comunidad apoyará y completará la acción de los Estados miembros en los siguientes ámbitos: (UE_1992s.s2906)
1. The Commission shall have the task of promoting the consultation of management and labour at Community level and shall take any relevant measure to facilitate their dialogue by ensuring balanced **support** for the parties. (UE_1992e.s2934)	1. La Comisión tendrá como cometido fomentar la consulta a los interlocutores sociales a nivel comunitario y adoptar todas las disposiciones necesarias para facilitar su diálogo, velando porque ambas partes reciban un apoyo equilibrado. (UE_1992s.s2926)
- **support** for national criminal investigation and security authorities, in particular in the coordination of investigations and search operations ; (UE_1992e.s3232)	- asistencia a las autoridades nacionales encargadas de la persecución penal y de la seguridad, en particular en materia de coordinación de diligencias e investigaciones ; (UE_1992s.s3224)

Pedro Patiño García

Multilingual Lexical Resources and Tools
UiB, Bergen, 22.06.2011

IMS Corpus Workbench – CQP
[pos="V.*"] []? "medida" [pos="ADJ"];

```
FTA-ES> [pos="V.*"] []? "medida" [pos="ADJ"];
   20144: aguardia textil . La Parte que <adopte la medida arancelaria> la aplicará solamente por el
  111943: garantía . 3 El tribunal puede <ordenar una medida provisional> de protección para preservar
  127077:  Más Favorecida ) , deberá ser <tratada como medida disconforme> no sujeta al Artículo 12 . 2
  184124: gor de este Acuerdo , e ( ii ) <incluye cualquier medida subordinada> , adoptada o mantenida bajo
  184958: trato nacional y por ello será <listada como medida disconforme> : a un banco extranjero de un
  186205: e al sector general al cual se <aplica la medida disconforme> ; ( b ) Subsector se refiere
  186222: l sector específico al cual se <aplica la medida disconforme> ; ( c ) Obligaciones Afectada
  186381: gor de este Tratado , y ( ii ) <incluye cualquier medida subordinada> , adoptada o mantenida bajo
  186478: e al sector general al cual se <aplica la medida disconforme> ; ( b ) Subsector se refiere
  186495: l sector específico al cual se <aplica la medida disconforme> ; ( c ) Obligaciones Afectada
  188681: eserva el derecho de adoptar o <mantener cualquier medida relativa> al suministro de servicios de
  193340: gor de este Acuerdo , e ( ii ) <incluye cualquier medida subordinada> , adoptada o mantenida bajo
  381962: á a definir y , en su caso , a <aplicar cualquier medida oportuna> para reforzar la base industr
  382881: uir a definir y , en su caso , <aplicar cualquier medida oportuna> para reforzar la base industr
  486013: ento de la Unión Europea , que <modifiquen una medida existente> que sea vinculante para dicho
  488517: ento de la Unión Europea , que <modifiquen una medida existente> que sea vinculante para dicho
  561468:  de la Parte importadora podrá <adoptar cualquier medida necesaria> para salvaguardar los interes
  586778: iquen que el supuesto subsidio <es una medida gubernamental> de asistencia , directa o ind
  594225:  desarrollado Parte que decida <aplicar una medida definitiva> en contra de productos origin
  621618: sin necesidad de que esa Parte <tome alguna medida legislativa> adicional . No obstante lo di
  639130: a sea directa o indirectamente <mediante] medida [s]> equivalentes a la expropiació
  639743: inatoria de aplicación general <es una medida equivalente> a la expropiación de un valor
  647655:  el párrafo 29 . 3 . a ) podrá <solicitar una medida provisional> de protección que no incluya
  654333: r a las partes contendientes , <[una medida provisional> de protección] [medidas provi
  654481:  ] [44 . 1 . Un Tribunal podrá <recomendar una medida provisional> de protección para preservar
```

Future work

- Finish corpus alignment (TCA2)
- Derive specialized collocations
- Compare specialized collocations found in FTA corpus with reference corpora and specialized dictionaries
- Test the suitability of current representation schemes (LMF, TBX, commercial software | private companies) for encoding specialized collocations.

Pedro Patiño García
Multilingual Lexical Resources and Tools
UiB, Bergen, 22.06.2011

Encoding of a Spanish phraseological unit using LMF

```xml
1 <?xml version="1.0" encoding="iso-8859-1"?>
2 <LexicalResource dtdVersion="16">
3 <GlobalInformation>
4        <feat att="label" val="Representación de locuciones"/>
5        <feat att="comment" val="Locución española daños y perjuicios"/>
6        <feat att="languageCoding" val="ISO 639-3"/>
7 </GlobalInformation>
8 <Lexicon>
9        <feat att="language" val="spa"/>
10 <LexicalEntry mwePattern="NConjN">
11        <feat att="partOfSpeech" val="LocNom"/>
12 <Lemma>
13        <feat att="writtenForm" val="daños y perjuicios"/>
14 </Lemma>
15 <ListOfComponents>
16        <Component entry="E1"/>
17        <Component entry="E2"/>
18        <Component entry="E3"/>
19 </ListOfComponents>
20 </LexicalEntry>
21 <LexicalEntry id="E1">
22        <feat att="partOfSpeech" val="nombre"/>
23 <Lemma>
24        <feat att="writtenForm" val="daños"/>
25 </Lemma>
```

Applications

Harmonization of language resources across domains and languages.

Creation of ontologies, corpus-based extraction of terms, domain recognition of text from ontologies.

Specialized translation, terminology, lexicography, technical writing.

Development of NLP tools for semi-automatic collocation extraction.
Teaching of specialized phraseology and translation.

Pedro Patiño García
Multilingual Lexical Resources and Tools
UiB, Bergen, 22.06.2011

References

Bartsch, S. (2004), Structural and functional properties of collocations in English: a corpus study of lexical and pragmatic constraints on lexical co-occurrence, Gunter Narr Verlag, Tübingen.

Choueka, Y. (1988). Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in Large Textual Databases. In Proceedings of RIAO'1988. (pp. 609-624).

Firth, J. R. (1957). Papers in Linguistics 1934-1951. Oxford University Press.

Palmer, H. E. 1933. Second Interim Report on English Collocations. Tokyo: Kaitakusha.

Pedro Patiño García
Multilingual Lexical Resources and Tools
UiB, Bergen, 20-23.06.2011

Tusen takk!

:-)

Your comments are welcome:

pedro.patino@nhh.no

Pedro Patiño García
Multilingual Lexical Resources and Tools
UiB, Bergen, 22.06.2011