RITerm 2012

XIII Simposio Iberoamericano de Terminología

Extracción automática de términos especializados en *UKSCC* y *TEC*, dos corpus específicos

Prof. María José Marín

Dra. Camino Rea

Departamento de Filología Inglesa

UNIVERSIDAD DE MURCIA

ESTRUCTURA

1. Introducción
2. Descripción y estructura de *UKSCC* y *TEC*
3. Metodología y análisis de datos
   - 3.1. Metodología
   - 3.2. Análisis de datos
     - 3.2.1. Frecuencia, *keyness* y distribución
     - 3.2.2. Terms y non-terms
4. Conclusión

1. INTRODUCCIÓN

- Conocer los términos técnicos ayuda a comprender los textos específicos.
- Rea (2008): sirven de herramienta de comunicación entre especialistas
- Spasic et. Al (2005): un término es "a textual realisation of a specialised concept"
- Cabré (2000): los términos son "unidades de forma y contenido que … adquieren un valor especializado"
- Nation y Waring (1997): los términos técnicos cubren un 5% de los textos especializados

- Los métodos ATR (Automatic term recognition methods) tratan de indentificar los términos especializados de manera automática.

- Maynard y Ananiadou, 2000; Cabré et. Al, 2001; Chung, 2003; Lemay et al., 2005; etc. muestran que los métodos ATR se centran en términos poliléxicos.

- Sin embargo los términos monoléxicos no son sencillos de identificar (no hay patrones gramaticales).

- Presentamos la aplicación y evaluación de un método ATR para la detección de términos monoléxicos diseñado por Chung (2003) en dos corpus especializados: *TEC* y *UKSCC*.

2. DESCRIPCIÓN Y ESTRUCTURA DE *UKSCC* Y *TEC*

UKSCC: *United Kingdom Supreme Court Corpus.* Corpus legal sincrónico (2008-2010) monolingüe en inglés de 2,6 millones de palabras. Compuesto por 193 textos: sentencias judiciales de la corte suprema del Reino Unido.

- Importancia de las *judicial decisions* como género textual*:* el derecho jurisprudencial en las cortes británicas.
- El Tribunal supremo es la corte de último recurso: variedad y riqueza léxica de los textos.

TEC: ***Telecommunication Engineering Corpus*****. Corpus de Ingeniería de Telecomunicaciones (5,5 millones de palabras).**

Características:

- **Inglés escrito académico/profesional**
- **Actos de comunicación con al menos un interlocutor experto/profesional**
- **Muestras producidas por hablantes nativos y no nativos extraídas de varias fuentes: revistas de divulgación y especializadas, libros, páginas web, folletos, anuncios y noticias tecnológicas.**
- **7 áreas de conocimiento: Electrónica; Arquitectura y Tecnología de Computadoras; Teoría de la Señal y Comunicaciones; Ciencia de los Materiales; Organización de Empresas; Ingeniería de Sistemas; e Ingeniería Telemática.**

Subcorpus

Subcorpus de Telemática (CT):

- Ingeniería Telemática y su especialización en Planificación y Gestión de Telecomunicaciones
- Rama que trata del funcionamiento de una red en su forma física: Fundamentos de la Telemática, Telemática, Redes y Servicios de Comunicaciones, Sistemas de Telecomunicaciones y Conmutación.
- Tamaño de las muestras: 2,2 millones de palabras (UKSCC: 2,6m).

3. METODOLOGÍA Y ANÁLISIS DE DATOS

3.1. Metodología

- Gran cantidad de métodos y herramientas de extracción e identificación terminológica: Church and Hanks (1994); Frantzi and Ananiadou (1999); Chung (2003); Drouin (2003); Kit and Liu (2008), etc.
- Chung (2003) calcula la especificidad de un término en relación a su frecuencia en el corpus específico y el general. Una ratio superior a 50 indica que una palabra es término.
- Chung llega a esta conclusión tras comparar su clasificación de términos cuantitativa con otra cualitativa.
- UKSCC y TEC se comparan con LACELL (21 millones de palabras).

3.2. ANÁLISIS DE DATOS

3.2.1. El método Chung (2003)

- Ambos corpus se procesan con *Wordsmith 5.0* : listas de tipos y frecuencia absoluta. Normalización de frecuencias para su comparación.
- LACELL (corpus general) también se procesa y las frecuencias se normalizan.
- Ratio=Frec CS/Frec CG
- Términos:
  - -Ratio >50 y elementos que no están en el corpus general (las listas han de ser revisadas)
  - -Uso de listas de corte para filtrar resultados (*Range)*
  - -Validación de resultados por comparación con gold standard.
  - -Ordenación de datos en función del nivel de especificidad

FRECUENCIA, *KEYNESS* Y DISTRIBUCIÓN

Frecuencia: Por sí sola no indica el nivel de especificidad. Señala las veces que un término se repite en un corpus.

Keyness: Aplicando el algoritmo de Dunning (1993), *log-likelihood,* indica frecuencia inusual de un término respecto de lo que se esperaría en un corpus general. Señala la representatividad de l candidato a término

Distribución: Indica en cuántos textos de corpus aparece el término.
Todos estos parámetros en combinación con los resultados obtenidos con el método Chung pueden usarse para elaborar listados ordenados en función de las preferencias del especialista.

- Pueden también utilizarse para organizar los listados de términos o en combinación con otros parámetros en función de las preferencias de lingüista, traductor o docente.

Corpus	Frecuencia media tipos	Frecuencia media términos
UKSCC	115,23	124,97
CT	28,38	115,23

Corpus	*Keyness* media tipos	*Keyness* media términos
UKSCC	102,33	504,75
CT	330	682,67

3.2.2 *Terms y non-terms*

UKSCC

Clasificación	Términos	Ejemplos
338 ratio>50 (*terms*)	146 = 43,19% aparecen en UKSCC y LACELL	*adjournal, bailee, casation, wayleave, contraventions, jurisdiction, unlawfulness*
22.269<50 (*non-terms*)	3.852 = 17,29% aparecen en un glosario especializado	Formas no especializadas*: redevelop, cellmate, inhuman, obviousness, strictness, hoarseness*

3.2.2 *Terms y non-terms*

CORPUS DE TELEMÁTICA

Clasificación	Términos	Ejemplos
3.416 ratio>50 (*terms*)	2.051= 60% aparecen en CT y LACELL	Boolean, cosine, wavelength, DBMS, debugger, electromagnetics, downstream
6.619<50 (*non-terms*)	350 = 5,3% aparecen en un glosario especializado	Formas no especializadas: *upload, director, browser, disable, nested, spam, aperture, congestion*

4. CONCLUSIÓN

- Mayor éxito en el caso del inglés de telemática que en el legal.
- Uso más extendido de términos legales en el lenguaje común, ej.: *acquittal, defendant* o *prosecution*.
- Método automático de Chung (2003) basado en las frecuencias general y específica.
- Tasas de precisión:

Inglés legal	Inglés de telemática	Inglés de anatomía
43%	60%	86%

- Fundamental la revisión manual de los listados de términos y de no-términos.

4. CONCLUSIÓN

- Organización de términos en función de su nivel de especificidad o en combinación con los parámetros de frecuencia, keyness o distribución.

- Disposición de una fuente de información fiable de léxico especializado con acceso al contexto real del uso de los términos que facilita la desambiguación de sus significados.

- Dado el tamaño de los corpus existentes, el uso de técnicas automáticas de reconocimiento terminológico se hace imprescindible como estado previo a la intervención del especialista.

RITerm 2012
XIII Simposio Iberoamericano de Terminología

Extracción automática de términos especializados en *UKSCC* y *TEC*, dos corpus específicos

Prof. María José Marín
Dra. Camino Rea
Departamento de Filología Inglesa

UNIVERSIDAD DE MURCIA