

3. Las bases de datos terminológicas

Índice

3.1.Introducción

En el capítulo anterior vimos uno de los principales recursos para la traducción: las memorias de traducción. En este capítulo veremos otro recurso de gran importancia: las bases de datos terminológicas. Comenzaremos el capítulo con una posible definición de término y presentaremos los principales conceptos relacionados con el trabajo terminológico. Veremos también la importancia de la terminología para la traducción.

Buena parte del capítulo lo dedicaremos a la revisar las técnicas de creación de bases de datos terminológicas y le daremos una gran importancia a las técnicas de extracción automática de terminología.

Presentaremos también a fondo el formato TBX para el intercambio de bases de datos terminológicas y veremos algunos recursos terminológicos libres y de libre acceso.

3.2.Terminología y traducción

Antes de comenzar el trabajo terminológico es necesario establecer una definición de término que nos ayude a decidir si una determinada unidad es o no es un término. Comenzaremos por la siguiente definición (Pazienza 2005):

Un término es una representación superficial de un concepto de un dominio específico[1]

Entendemos por *representación superficial* la denominación, es decir, la palabra o conjunto de palabras que se utilizan para referirse a este concepto. Por *dominio específico* entendemos el *campo de especialidad.* Este aspecto es muy importante, ya que la terminología trabaja con unidades de conocimiento especializado que se utilizan en campos de especialidad concretos. No consideramos, pues, el lenguaje general, aunque algunos términos referentes a conceptos de uso común están también presentes en el lenguaje no especializado.

Los términos se pueden ver como unidades formadas por un *concepto* y su *denominación.* Esta es la base de la Teoría General de la Terminología de Wüster. Tal y como explica Sánchez-Gijón (2004), la Teoría General de la Terminología estructura el conocimiento en sistemas conceptuales y dota a cada concepto de una denominación con el objetivo de establecer una comunicación inequívoca entre los expertos.

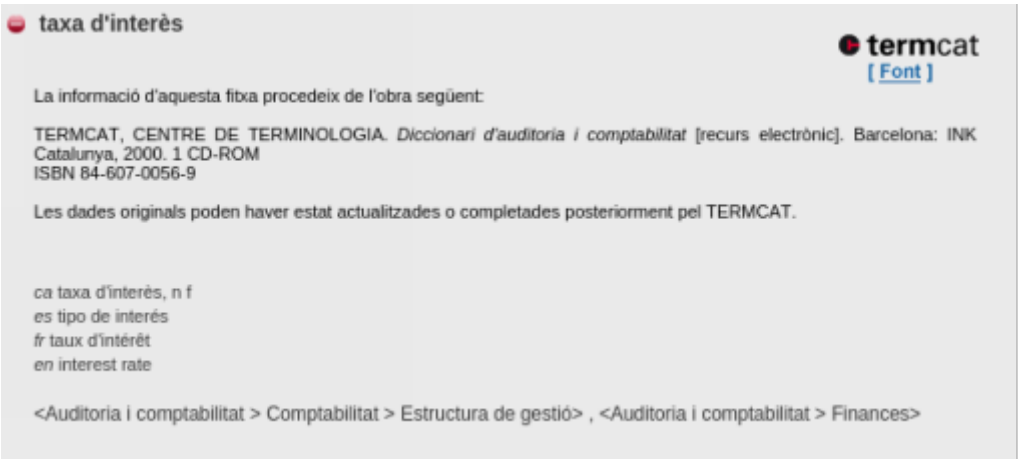
Hay muchas otras teorías sobre la terminología, pero no entraremos en detalles en este capítulo. Quien quiera profundizar en este tema puede leer la tesis doctoral de Mercè Vàzquez (2014).

A menudo el traductor olvida las teorías sobre la terminología y las definiciones de término e incluye en sus recopilaciones terminológicas unidades que no pueden ser consideradas propiamente términos, pero que precisan también de una gran consistencia en su traducción. En este capítulo veremos cómo construir y gestionar bases de datos terminológicas y tendremos en cuenta también esta aproximación más práctica.

1 *A surface representation of specific domain concept*

3.3. Bases de datos terminológicas, diccionarios generales y conocimiento enciclopédico

Las *bases de datos terminológicas* son recopilaciones terminológicas sistematizadas y generalmente (hoy en día, siempre) en un formato informático y por tanto utilizables mediante un ordenador. Las bases de datos terminológicas, pues, recogen términos, que son unidades propias del lenguaje especializado.



taxa d'interès

termcat

[Font]

La informació d'aquesta fitxa procedeix de l'obra següent:

TERMCAT, CENTRE DE TERMINOLOGIA. *Diccionari d'auditoria i comptabilitat* [recurs electrònic]. Barcelona: INK Catalunya, 2000. 1 CD-ROM
ISBN 84-607-0056-9

Les dades originals poden haver estat actualitzades o completades posteriorment pel TERMCAT.

ca taxa d'interès, n f
es tipo de interés
fr taux d'intérêt
en interest rate

<Auditoria i comptabilitat > Comptabilitat > Estructura de gestió> , <Auditoria i comptabilitat > Finances>

Los diccionarios generales incluyen palabras propias del lenguaje general. Algunos términos propios de campos especializados, que son de uso común, han entrado en el lenguaje general y por lo tanto pueden estar incluidos en los diccionarios.

interés.

(Del lat. *interese*, importar).

1. m. Provecho, Utilidad, ganancia.

2. m. Valor de algo.

3. m. Lucro Producido por el capital.

4. m. Inclinación del Ánimo Hacia un objetivo, una persona, una narración, etc

5. m. pl. **Bien.**

6. m. pl. Conveniencia o beneficio en el orden moral o material.

~ Compuesto.

1. m. **interés** de un capital al que se van acumulando suspen réditos para que produzcan Otros.

Intereses a proporcionar.

1. m. pl. Cuenta que se reduce a dividir los pagos que se Hacen a Cuenta de un capital que Produce **Intereses,** en dos partes proporcionales a la Cantidad de débito ya la suma de **Los** intereses devengados; como, por Ejemplo, si el débito fuese 20 y los **Intereses** adeudados 10, y el pago es de 6, se aplican 4 al capital y 2 en los **interesante.**

Intereses a prorrata.

1. m. pl. Cuenta que se llevaba en la Contaduría mayor de Cuentas, y consistía en supone el débito que habian de producir los **Intereses** en Cierta día; y el tiempo de pagarse una porción a Cuenta, se Cubría primeramente con ella el importe íntegro de

dichos réditos, APLICANDOSE el resto en Cuenta del débito principal, el cual se quedaba establecido en el Mismo día que se causaba, y desde el producer los Intereses que CORRESPONDIAN la Cantidad a que quedaba Reducida.

intereses creados.

1. m. pl. Ventajas, no siempre legítimas, de que Gozán varios individuos, y por efecto de las cuales se establece entre Ellos alguna solidaridad circunstancial que Reducir texto
 oponerse a alguna obra de justicia o de mejoramiento social. Uno t. siente. peyor.

Intereses de demora.

1. m. pl. *Der.* Intereses que deberan abonar el deudor moroso.

~ legal.

1. m. interés que, a falta de estipulación previa sobre sume cuantia, fija la ley.

~ legítimos.

1. m. *Der.* interés de una persona reconocida y protegido por el derecho.

2. m. *Der.* Situación jurídica que se ostenta en relaciones con la actuacion otra persona y que conllevar la facultad de exigirle, a través de un proceÂdimiento administrativo o judicial, un Comportamiento ajustada a derecho.

~ simple.

1. m. interés de un capital sin agregarle los réditos.

□ V.

dinero a interés

El *conocimiento enciclopédico* es el conocimiento del mundo acumulado por la humanidad. Una enciclopedia es un compendio de este conocimiento. Las entradas de las enciclopedias se refieren a elementos culturales de tipología muy diversa pero que, a diferencia de los diccionarios, no constituyen material estrictamente lexicográfico [fuente Wikipedia]. La enciclopedia, además de ofrecer una definición de las palabras o términos, ofrecen información más amplia y profunda.

Interés

Para otros usos de este término, véase Interés (desambiguación).

Interés es un índice utilizado para medir la rentabilidad de los ahorros o también el costo de un crédito. Se expresa generalmente como un porcentaje.

Dada una cantidad de dinero y un plazo o término para su devolución o su uso, el **tipo de interés** indica qué porcentaje de ese dinero se obtendría como beneficio, o en el caso de un crédito, qué porcentaje de ese dinero habría que pagar. Es habitual aplicar el interés sobre períodos de un año, aunque se pueden utilizar períodos diferentes como un mes o el número días. El tipo de interés puede medirse como el tipo de interés nominal o como la tasa anual equivalente. Ambos números están relacionados aunque no son iguales.

Índice [ocultar]

1 Introducción
2 Tipo de interés
  2.1 Tipo de interés (TIN)
  2.2 Tasa anual equivalente (TAE)
  2.3 Tipo de interés real o ajustado
3 Véase también
4 Enlaces externos

Introducción [editar]

En economía y finanzas, una persona o entidad financiera que presta dinero a otros esperando que le sea devuelto al cabo de un tiempo espera ser compensado por ello, en concreto lo común es prestarlo con la expectativa de que le sea devuelta una cantidad ligeramente superior a la inicialmente prestada, que le compense por la dilación de su consumo, la inconveniencia de no poder hacer uso de ese dinero durante un tiempo, etc. Además esperará recibir compensación por el riesgo asociado a que el préstamo no le sea devuelto o que la cantidad que le sea devuelta tenga una menor capacidad de compra debido a la inflación.

El prestamista fijará un tipo de interés nominal (TIN) que tendrá en cuenta los tres tipos de factores, de tal manera que al final, recibirá la cantidad inicial más un fracción de esa cantidad dada por el tipo de interés nominal:

$$K_f = K_0(1 + i_N)$$

Donde:

Estas tres fuentes: bases de datos terminológicas, diccionarios generales y enciclopedias; son herramientas de consulta muy habituales de los traductores.

3.4. Búsqueda automática en bases de datos terminológicas

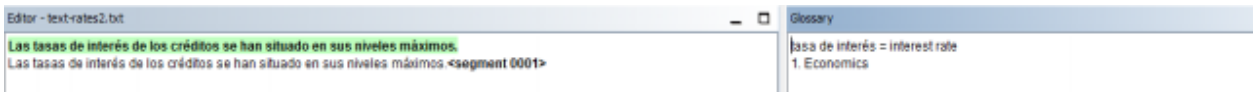
Las herramientas de traducción asistida permiten una consulta automática a bases de datos terminológicas. Si el segmento que estamos traduciendo aparece un término de la base de datos terminológica, el programa resaltará este término y nos mostrará la información relativa (como por ejemplo, la traducción) en una de las pantallas de la herramienta. A continuación vemos un ejemplo:


Editor - text-rates.txt
The interest rates on prime credits in the late 1970s and early 1980s were far higher than had been recorded.
The interest rates on prime credits in the late 1970s and early 1980s were far higher than had been recorded.<segment 0001>
Glossary
interest rate = tasa de interés
1. Economics

Fijémonos en que la herramienta tiene que ser capaz de reconocer el término de la base de datos terminológica aunque este aparezca en otra forma. Si nos fijamos, en la base de datos terminológica tenemos recogida la forma base (singular): *interest rate* pero en el texto aparece en plural *interest rates*.

Las herramientas de traducción asistida tienen que ser capaces de encontrar términos en otras formas sin tener un conocimiento demasiado profundo ni específico de la lengua. Para lenguas con una morfología compleja los sistemas genéricos de reconocimiento de términos pueden fallar.

En OmegaT el reconocimiento de términos se lleva a cabo mediante *tokenizadores* que son capaces de hacer *stemming* (es decir, eliminar los afijos morfológicos de las palabras). De esta manera, eliminando estos afijos tanto en el texto del segmento a traducir como en las entradas de la base de datos terminológica, el programa es capaz de encontrar las entradas aunque no coincidan plenamente las formas. Veamos ahora el mismo ejemplo para el castellano:

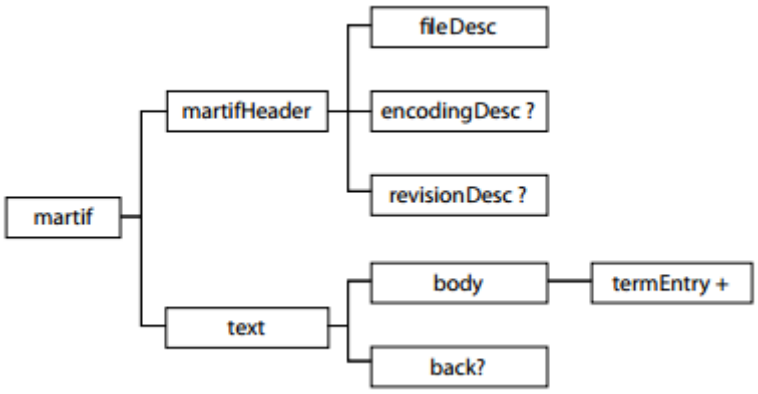

Editor - text-rates2.txt
Las tasas de interés de los créditos se han situado en sus niveles máximos.
Las tasas de interés de los créditos se han situado en sus niveles máximos.<segment 0001>
Glossary
tasa de interés = interest rate
1. Economics

Aunque reconozca el término en una forma diferente, si recuperamos la traducción de la base de datos terminológica, por regla general los sistemas de traducción asistida no serán capaces de insertar en término traducido en la forma correcta (en plural en estos ejemplos). Para poder hacer esto el sistema debería disponer de más información lingüística.

3.5.Formato de intercambio de bases de datos terminológicas: TBX

En el capítulo anterior, dedicado a las memorias de traducción, vimos el formato de intercambio TMX (*Translation Memory eXchange).En el caso de las bases de datos terminológicas hay un formato similar, también basado en XML, llamado TBX (Term Base eXchange*). La idea es exactamente la misma: aunque cada gestor de bases de datos terminológicas y cada herramienta de traducción asistida pueda trabajar con un formato interno diferente para representar las bases de datos terminológicas, podremos compartir los datos terminológicos con otras herramientas utilizando este formato de intercambio.

El TBX es un estándar internacional (ISO 30042: 2008) para la representación de datos terminológicos, publicado conjuntamente por la ISO (*International Standard Organisation*) y LISA (*Localization Industry Standard Association*). Se pueden encontrar las especificaciones completas en LISA (2008). Aquí presentaremos un pequeño resumen de las características más destacadas de este formato. En el siguiente esquema podemos observar la estructura de un documento MARTIF (*Machine-Readable Terminology Interchange Format*):

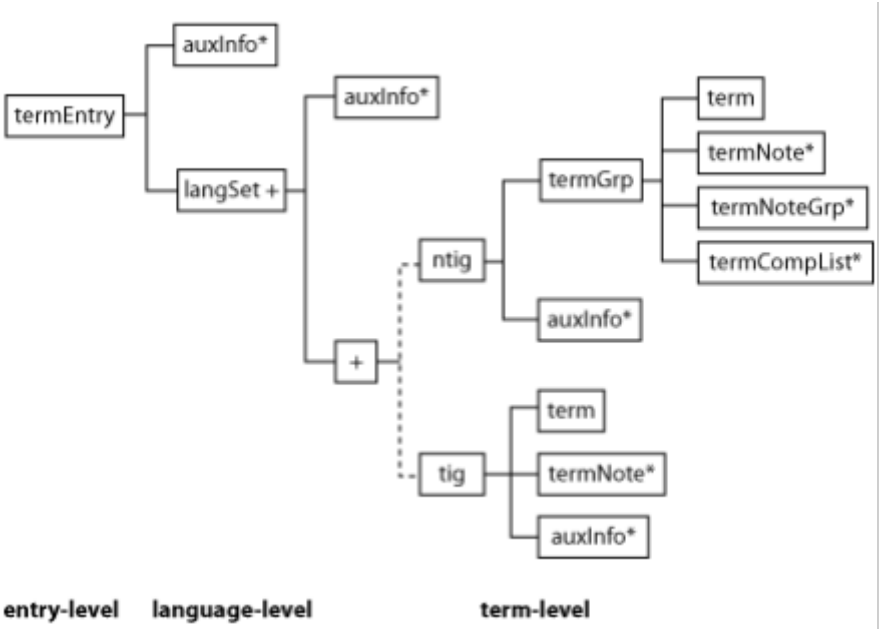


Como podemos ver en el esquema, el nivel más alto del documento XML es el elemento `<martif>`, que consiste en un elemento `<martifHeader>` y un elemento `<text>`.Los nombres de estos elementos se han tomado de la norma ISO 12200 y tiene sus raíces en la *Text Encoding Initiative*. El elemento `<text>` consiste en las entradas terminológicas, que están englobadas en un elemento `<body>` elemento e información complementaria. En TBX, la información complementaria se encuentra en el elemento `<back>`.

La información sobre la codificación de caracteres se debe incluir en la cabecera sólo cuando la codificación sea diferente de Unicode.

Componentes de una entrada terminológica

Cada entrada terminológica dentro del elemento `<body>` se denomina `<termEntry>` y sigue la estructura del metamodelo TMF. En la siguiente figura podemos observar los niveles de una entrada terminológica:



El recuadro auxInfo corresponde a información que se puede asociar a cualquiera de los tres niveles: el nivel de Entrada Terminológica (`<termEntry>`, es decir, el nivel de concepto), el nivel de Lengua (`<langSet>`) y el nivel de Término (es decir, la denominación, `<ntig>` o su versión simplificada `<tig>`). Los elementos `<termNote>` y `<termNoteGrp>` sólo pueden aparecer en el nivel de Término o por debajo.

Ejemplo de archivo TBX

A continuación podemos observar un ejemplo de archivo TBX:

```
<?xml version='1.0'?> <!DOCTYPE martif SYSTEM "TBXcoreStructV02.dtd">
<martif type="TBX" xml:lang="en">
 <martifHeader>
 <fileDesc>
 <sourceDesc>
 <p>From an Oracle corporation termbase</p>
 </sourceDesc>
 </fileDesc>
 <encodingDesc>
 <p
type="XCSURI">http://www.lisa.org/fileadmin/standards/tbx/TBXXCSV02.XCS</p>
 </encodingDesc>
 </martifHeader>
 <text>
 <body>
 <termEntry id="eid-Oracle-67">
 <descrip type="subjectField">manufacturing</descrip>
 <descrip type="definition">A value between 0 and 1 used in ...</descrip>
 <langSet xml:lang="en">
 <tig>
 <term id="tid-Oracle-67-en1">alpha smoothing factor</term>
 <termNote type="partOfSpeech">noun</termNote>
 </tig>
 </langSet>
 <langSet xml:lang="hu">
 <tig>
```



```
 <term id="tid-Oracle-67-hu1">Alfa simítási tényez </term> ő
 <termNote type="partOfSpeech">noun</termNote>
 </tig>
 </langSet>
 </termEntry>
 </body>
 </text>
</martif>
```

Dialectos del TBX

El TBX es un formato de intercambio muy bien diseñado y que se adapta perfectamente a las tareas terminológicas más complejas. Las especificaciones completas del TBX tienen más de 90 páginas e implementar aplicaciones que sean totalmente compatibles con el estándar entero resulta, sino complicado, sí muy laborioso.

Muchas de las tareas relacionadas con la terminología, y especialmente las tareas prácticas relacionadas con el trabajo del traductor, no requieren un formato de intercambio tan completo. Por este motivo se han creado *dialectos* del propio TBX con el objetivo de simplificarlo. En esta sección presentaremos dos de estos dialectos: el TBX-Basic y el TBX-Min. Un aspecto muy importante a tener en cuenta de estos dialectos es que son en sí TBX válidos. Por lo tanto, una aplicación capaz de leer y procesar los TBX completos, será también capaz de leer estos dialectos simplificados.

TBX-Basic

El TBX-Basic está diseñado para proporcionar las categorías de datos que se utilizan de forma habitual en las tareas de traducción y localización. Las principales diferencias entre el TBX completo y el TBX-Basic son las siguientes:

- El TBX dispone de los elemento s `<tig>` y `<ntig>` para los grupos de información sobre el término. En cambio, el TBX-Basic sólo dispone del elemento `<tig>`.
- El TBX-Basic no permite documentar los componentes de un término (es decir, las partes individuales de los términos). Por lo tanto los siguientes elementos no están presentes en TBX-Basic: `<termComp>`, `<termCompList>`, `<termCompGrp>`, y `<termGrp>`.
- El TBX-Basic no admite los siguientes elementos de agrupación y de sus elementos hijos: `<adminGrp>`, `<termNoteGrp>`, `<itemSet>` and `<itemGrp>`. En TBX-Basic sólo se admiten los siguientes elementos de agrupación: `<descripGrp>` y `<transacGrp>`.
- En TBX-Basic, el elemento `<descripGrp>` se utiliza sólo para asociar una fuente a una definición o un contexto. Por lo tanto, los siguientes elementos hijos no se admiten en TBX-Basic: `<descripNote>`, `<admin>`, `<adminGrp>`, `<note>`, `<ref>`, `<xref>`.
- En TBX-Basic, los valores de los atributos "DCSName" y "XCSContent" no son compatibles con la etiqueta de párrafo en el elemento `<encodingDesc>`.

TBX-Min

El TBX-Min está diseñado para facilitar el almacenamiento de glosarios monolingües y bilingües y para permitir una conversión fácil con TBX-Basic y con otros formatos para el almacenamiento de glosarios, como por ejemplo el UTX *(Universal Terminology eXchange* - http://www.aamt.info/english/utx/). Una de sus utilidades más claras es la utilización de este formato para enviar un glosario a un traductor que participa en un proyecto y que no necesita de toda la información que potencialmente está incluida en el glosario completo en formato TBX-Basic (o TBX completo). El traductor puede recibir este TBX-Min, trabajar con él introduciendo cambios o nueva información. Una vez terminado el trabajo puede enviar este TBX-Min y los cambios se pueden incluir en la TBX-Basic (o completo).

La cabecera de un TBX-Min contiene la siguiente información:

- Un ID único para identificar el documento
- El nombre del creador
- Una descripción de la base de datos terminológica
- La direccionalidad: monodireccional o bidireccional
- La lengua de partida y de llegada (sólo se pueden usar dos lenguas)
- La licencia de la base de datos terminológica
- La fecha en la que se creó la base de datos terminológica (en formato ISO 8601)

El cuerpo (`<body>`) contiene las entradas anidadas usuales `<langGroup>` y `<termEntry>`.

El elemento `<termGroup>` puede contener:

- el texto del término (denominación). Este es el único elemento obligatorio
- una nota
- la categoría gramatical
- el nombre del cliente
- el estado del término

El formato UTX *(Universal Terminology eXchange*)

El formato UTF es un formato estándar para glosarios muy simple y que puede ser fácilmente compartido y reutilizado entre diferentes herramientas. El UTX es un formato muy sencillo, ya que es simplemente texto tabulado, y se puede crear y manipular fácilmente con cualquier editor de textos u hoja de cálculo. Los términos pueden contener un campo de Term Status (que puede ser *approved* o *forbidden*).

A continuación podemos observar un ejemplo de glosario en este formato:

```
#UTX-S 1.01; en-US/ja-JP; 2009-09-08T17:55:00Z+09:00 copyright: Medical
Informatics, School of Allied Health Sciences, Kitazato University (2009);
license: CC-BY 3.0
#description: This is a medical dictionary. You may use this dictionary only
when you agreed that you and you alone are fully responsible for the results
of its use. The author of the original data, AAMT and its members do not
guarantee the contents of this dictionary including, but not limited to its
accuracy. Some part of speech properties are indicated as noun, even when
they are not. / ▪ ▪□□□□▪ □□□□□▪ □▪ □▪ □▪ □▪ □□□□▪ □□□▪ □▪ □□□□□▪ □□▪ □▪ □□□
▪ □□▪ □□□▪ □▪ □□▪ □□▪ □□▪ □□▪ □▪ □□□□▪ □□AAMT▪ ▪ ▪ AAMT ▪ □□□□□▪ □□□▪ □▪ □□▪ □▪ □□□▪ □
容について一切の保証はいたしません。品詞が名詞でない場合も名詞扱いになっていることがありま
▪ ▪
#src  tgt   src:pos     comment
1/2 T vector      1/2Tベクトル noun
1/2FF 1/2拡張分画  noun
1/3 ER mean 駆出早期1/3 ▪ ▪□□□□□▪□  noun
1/3EF 1/3駆出分画  noun
1/3ER mean  駆出早期1/3 ▪ ▪□□□□□▪□  noun
1/3FF 1/3充満分画  noun
1/3FR mean  拡張早期1/3 ▪ ▪□□□□□□□▪□noun
11-deoxycorticosterone acetate salt hypertension      DOCA食塩高血圧      noun
11-deoxycortisosterone  11-デオキシコルチコステロン noun
131I-hippurate    131I-ヒプル酸塩     noun
17 α-hydroxycorticosteroid    17αヒドロキシコルチコステロイド    noun
```



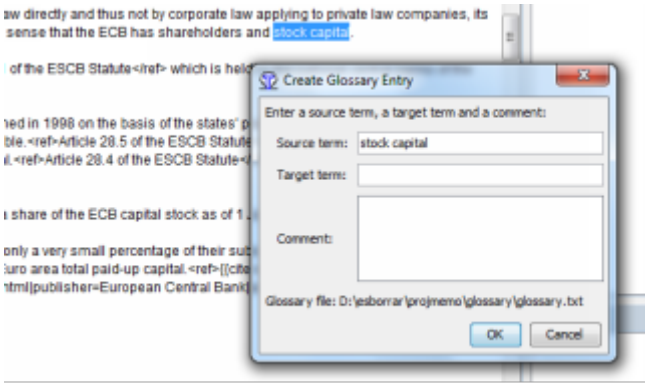
```
17 β-hydroxysteroid dehydrogenase   17β ヒドロキシステロイドデヒドロゲナ□ゼ    noun
17-hydroxycorticoid     17-ヒドロキシコルチコイド   noun
17-hydroxydesoxycorticosterone      17-ヒドロキシデゾキシコルチコステロン     noun
17-hydroxyprogesterone  17-ヒドロキシプロゲステロン noun
17-ketosteroid    17-ケトステロイド    noun
17-ks 17-ケトステロイド    noun
…
```

Este formato sólo permite representar glosarios bilingües. A esta versión se la conoce como UTX simple. Está prevista la aparición de un UTX-XML más complejo que permita la representación de glosarios multilingües.

3.6. Creación de bases de datos terminológicas

3.6.a. A medida que se traduce

Los traductores, a medida que van avanzando en una traducción van consultando diversas fuentes para resolver sus dudas terminológicas. En este momento es importante almacenar el resultado de las consultas en algún formato que sea consultable de manera fácil y rápida. La mejor opción, evidentemente, es almacenar las consultas en la misma base de datos terminológica que utiliza la propia herramienta de traducción asistida. Todas las herramientas de traducción asistida disponen de alguna funcionalidad que permite introducir nuevos términos en la bases de datos terminológica del proyecto. En la siguiente imagen podemos observar la pantalla de creación de entradas terminológicas de OmegaT. La herramienta permite seleccionar un término del original y cuando se abre la pantalla de creación de entradas el término seleccionado aparece automáticamente en el campo *Source term*. El traductor podrá completar el resto de campos y el término se almacenará automáticamente en la base de datos terminológica. A partir de este momento la información sobre el término aparecerá automáticamente en este proyecto de traducción. También podremos exportar estas entradas e importarlas a otras bases de datos.

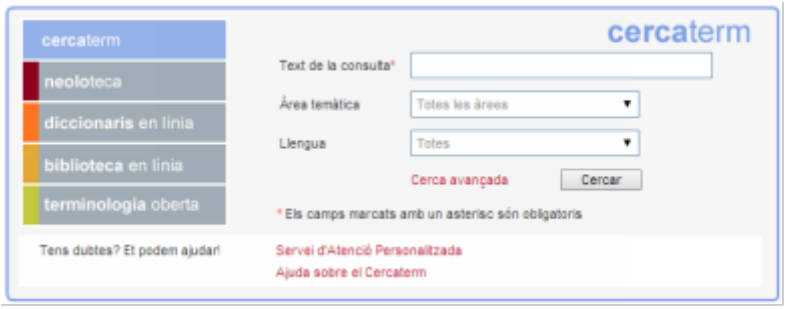


3.6.b. A partir de recursos terminológicos en Internet

En Internet podemos encontrar algunas páginas que distribuyen recursos terminológicos o bien que permiten realizar búsquedas terminológicas. En este apartado comentaremos algunos de estos recursos.

TermCat

Empezaremos hablando del TermCat (http://www.termcat.cat/), que es el centro de terminología de la lengua catalana, creado en 1985 por la Generalitat de Catalunya y el Institut d'Estudis Catalans. Aunque está creado para el catalán la mayoría de entradas terminológicas que recogen están también en inglés y castellano, y algunas también en francés o alemán. El TermCat por un lado ofrece el Cercaterm, que permite hacer consultas terminológicas a partir del término a buscar, la lengua y la posibilidad de indicar el área temática.



Además de la interfaz gráfica el TermCat ofrece un servicio de consultas terminológicas, para los casos en que no encontramos lo que buscamos en su interfaz.

El TermCat libera sus bases de datos terminológicas y las publica en la sección Terminologia Oberta (http://www.termcat.cat/ca/TerminologiaOberta/). Estas bases de datos terminológicas están en un formato XML no estándar pero que se puede convertir en TBX o texto tabulado mediante la herramienta TO2TBX (http://lpg.uoc.edu/TO2TBX/).

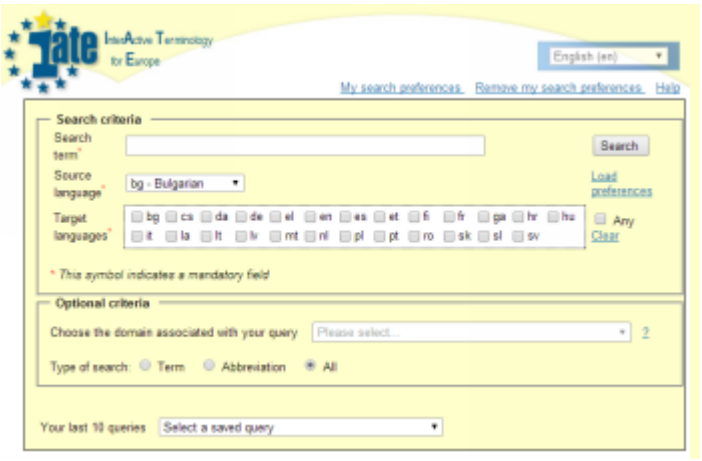
Además el TermCat ha desarrollado un programa de gestión de la terminología, el GesTerm, que se distribuye bajo una licencia libre (http://www.termcat.cat/ca/GesTerm/).

IATE

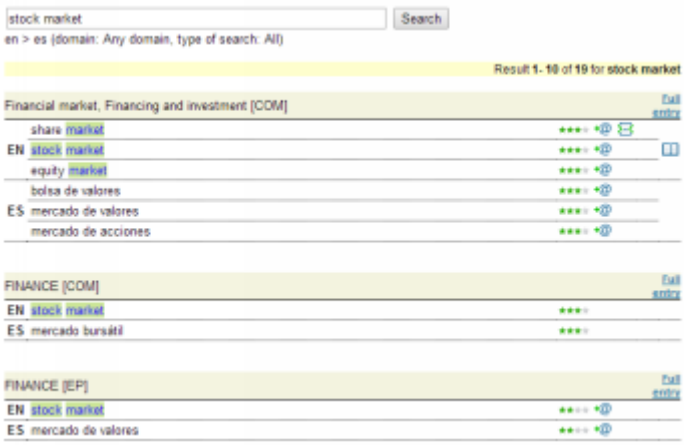
IATE (= "Inter-Active Terminology for Europe") (http://iate.europa.eu/) es la base de datos terminológica inter-institucional de la Unión Europea. Actualmente contiene aproximadamente 1.4 millones de entradas multilingües. Se han importado las siguientes bases de datos terminológicas:

- Eurodicautom (Commission)
- TIS (Council)
- Euterpe (EP)
- EUROTERM (Translation Centro)
- CDCTERM (Court of Auditors)

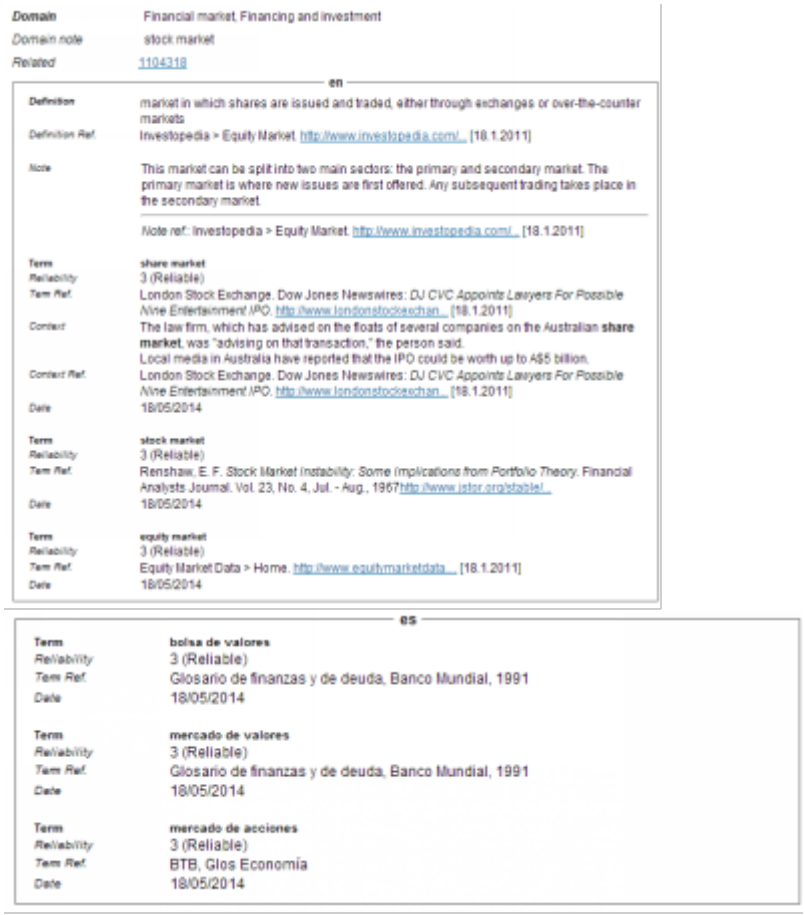
IATE permite hacer consultas mediante una interfaz de búsqueda:



Y obtener una serie de resultados:



De cada una de las entradas podemos obtener la entrada terminológica completa:



Recientemente, la base de datos IATE se ha liberado y se ha publicado como un archivo TBX de gran tamaño que contiene sus entradas. A continuación podemos observar una de estas entradas:

```
<termEntry id="IATE-84">
        <descripGrp>
          <descrip type="subjectField">1011</descrip>
        </descripGrp>
        <langSet xml:lang="bg">
          <tig>
            <term>компетенции на държави членове</term>
            <termNote type="termType">fullForm</termNote>
```

```
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="cs">
    <tig>
      <term>příslušnost členských států</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="da">
    <tig>
      <term>medlemsstatskompetence</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="de">
    <tig>
      <term>Zuständigkeit der Mitgliedstaaten</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="el">
    <tig>
      <term>αρμοδιότητα των κρατών μελών</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="en">
    <tig>
      <term>competence of the Member States</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="es">
    <tig>
      <term>competencias de los Estados miembros</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="et">
    <tig>
      <term>liikmesriikide pädevus</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="fi">
    <tig>
      <term>jäsenvaltioiden toimivalta</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="fr">
    <tig>
      <term>compétence des États membres</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="ga">
    <tig>
      <term>inniúlacht na mBallstát</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="hu">
```

```
    <tig>
      <term>tagállami hatáskör</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="it">
    <tig>
      <term>competenza degli Stati membri</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="lt">
    <tig>
      <term>valstybių narių kompetencija</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">2</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="lv">
    <tig>
      <term>dalībvalstu kompetence</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="nl">
    <tig>
      <term>bevoegdheid van de lidstaten</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="pl">
    <tig>
      <term>kompetencje państw członkowskich</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="pt">
    <tig>
      <term>competência dos Estados-Membros</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="ro">
    <tig>
      <term>competența statelor membre ale Uniunii Europene</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="sk">
    <tig>
      <term>právomoci členských štátov</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="sl">
    <tig>
      <term>pristojnost držav članic</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
    </tig>
  </langSet>
  <langSet xml:lang="sv">
    <tig>
      <term>medlemsstaternas behörighet</term>
      <termNote type="termType">fullForm</termNote>
      <descrip type="reliabilityCode">3</descrip>
```



```
      </tig>
    </langSet>
  </termEntry>
```

Como podemos ver, se trata de una base de datos multilingüe de las lenguas oficiales de la Unión Europea. No todas las entradas están en todas las lenguas. Las entradas contienen también información de campo temático (subjectField) que en esta entrada es el 1011, que corresponde a *European Union Law.*

En la siguiente tabla podemos observar todos los códigos de campo temático del IATE:

```
04    POLITICS
0406  Political framework
0411  Political Parties
0416  Electoral procedure and voting
0421  Parliament
0426  Parliamentary proceedings
0431  Politics and public safety
0436  Executive power and public service
08    INTERNATIONAL RELATIONS
0806  International affairs
0811  Cooperation policy
0816  International balance
0821  Defence
10    EUROPEAN UNION
1006  EU institutions and European civil service
1011  European Union law
1016  European construction
1021  EU finance
12    LAW
1206  Sources and branches of the law
1211  Civil law
1216  Criminal law
1221  Justice
1226  Organisation of the legal system
1231  International law
1236  Rights and freedoms
16    ECONOMICS
1606  Economic policy
1611  Economic growth
1616  Regions and regional policy
1621  Economic structure
1626  National accounts
1631  Economic analysis
20    TRADE
2006  Trade policy
2011  Tariff policy
2016  Trade
2021  International trade
2026  Consumption
2031  Marketing
2036  Distributive trades
24    FINANCE
2406  Monetary relations
2411  Monetary economics
2416  Financial institutions and credit
2421  Free movement of capital
2426  Financing and investment
2431  Insurance
2436  Public finance and budget policy
```



```
2441  Budget
2446  Taxation
2451  Prices
28    SOCIAL QUESTIONS
2806  Family
2811  Migration
2816  Demography and population
2821  Social framework
2826  Social affairs
2831  Culture and religion
2836  Social protection
2841  Health
2846  Construction and town planning
32    EDUCATION AND COMMUNICATIONS
3206  Education
3211  Teaching
3216  Organisation of teaching
3221  Documentation
3226  Communications
3231  Information and information processing
3236  Information technology and data processing
36    SCIENCE
3606  Natural and applied sciences
3611  Humanities
40    BUSINESS AND COMPETITION
4006  Business organisation
4011  Business classification
4016  Legal form of organisations
4021  Management
4026  Accounting
4031  Competition
44    EMPLOYMENT AND WORKING CONDITIONS
4406  Employment
4411  Labour market
4416  Organisation of work and working conditions
4421  Personnel management and staff remuneration
4426  Labour law and labour relations
48    TRANSPORT
4806  Transport policy
4811  Organisation of transport
4816  Land transport
4821  Maritime and inland waterway transport
4826  Air and space transport
52    ENVIRONMENT
5206  Environmental policy
5211  Natural environment
5216  Deterioration of the environment
56    AGRICULTURE, FORESTRY AND FISHERIES
5606  Agricultural policy
5611  Agricultural structures and production
5616  Farming systems
5621  Cultivation of agricultural land
5626  Means of agricultural production
5631  Agricultural activity
5636  Forestry
5641  Fisheries
60    AGRI-FOODSTUFFS
6006  Plant product
6011  Animal product
```



```
6016  Processed agricultural produce
6021  Beverages and sugar
6026  Foodstuff
6031  Agri-foodstuffs
6036  Food technology
64    PRODUCTION, TECHNOLOGY AND RESEARCH
6406  Production
6411  Technology and technical regulations
6416  Research and intellectual property
66    ENERGY
6606  Energy policy
6611  Coal and mining industries
6616  Oil industry
6621  Electrical and nuclear industries
6626  Soft energy
68    INDUSTRY
6806  Industrial structures and policy
6811  Chemistry
6816  Iron, steel and other metal industries
6821  Mechanical engineering
6826  Electronics and electrical engineering
6831  Building and public works
6836  Wood industry
6841  Leather and textile industries
6846  Miscellaneous industries
72    GEOGRAPHY
7206  Europe
7211  Regions of EU Member States
7216  America
7221  Africa
7226  Asia and Oceania
7231  Economic geography
7236  Political geography
7241  Overseas countries and territories
76    INTERNATIONAL ORGANISATIONS
7606  United Nations
7611  European organisations
7616  Extra-European organisations
7621  World organisations
7626  Non-governmental organisations
```

Se puede acceder a una lista mucho más detallada de los códigos de campo temático en: http://iate.europa.eu/tbx/IATE%20domain%20codes.csv

Cada entrada tiene también una información de fiabilidad (*reliabilityCode*) que puede tener tres niveles:

- 1: Fiabilidad no verificada
- 2: Fiabilidad mínima
- 3: Fiable
- 4: Muy fiable

Se puede encontrar una descripción completa de la información sobre los temas de la IATE en: http://iate.europa.eu/tbx/IATE%20Data%20Fields%20Explaind.htm

El fichero TBX que se puede descargar es un archivo muy grande y que es difícil de tratar con las herramientas estándar. En http://lpg.uoc.edu/IATE se puede acceder a ficheros de texto tabulado que contienen los términos clasificados por pares de lengua y por especialidades. También se puede descargar

una sencilla herramienta (IATE2tabtxt.py) que permite hacer la conversión del TBX en formatos de texto tabulados para pares de lenguas. En la misma web está disponible otra herramienta (IATE2TBX.py) que crea archivos TBX que contienen sólo la información de los pares de lengua y las especialidades deseados. De esta manera se pueden crear archivos mucho más fáciles de manipular con las herramientas estándar.

Eurovoc

Eurovoc (http://eurovoc.europa.eu/) es un tesauro multilingüe y multidisciplinar que incluye la terminología de los ámbitos de actividad de la Unión Europea, con especial énfasis en las tareas parlamentarias. Eurovoc está disponible en 23 lenguas oficiales de la Unión Europea (alemán, búlgaro, checo, croata, danés, eslovaco, esloveno, español, estonio, finés, francés, griego, húngaro, inglés, italiano, letón, lituano, maltés, neerlandés, polaco, portugués, rumano y sueco), además de la lengua de un tercer país (serbio). Eurovoc también está disponible en catalán[2] y euskera[3].

Eurovoc se puede descargar en varios formatos que pueden ser incorporados a bases de datos terminológicas.

Unterm

Unterm (http://unterm.un.org/) es la base de datos terminológica de las Naciones Unidas, que contiene términos técnicos y nomenclaturas en las seis lenguas oficiales de esta institución: árabe, chino, inglés, ruso y español.

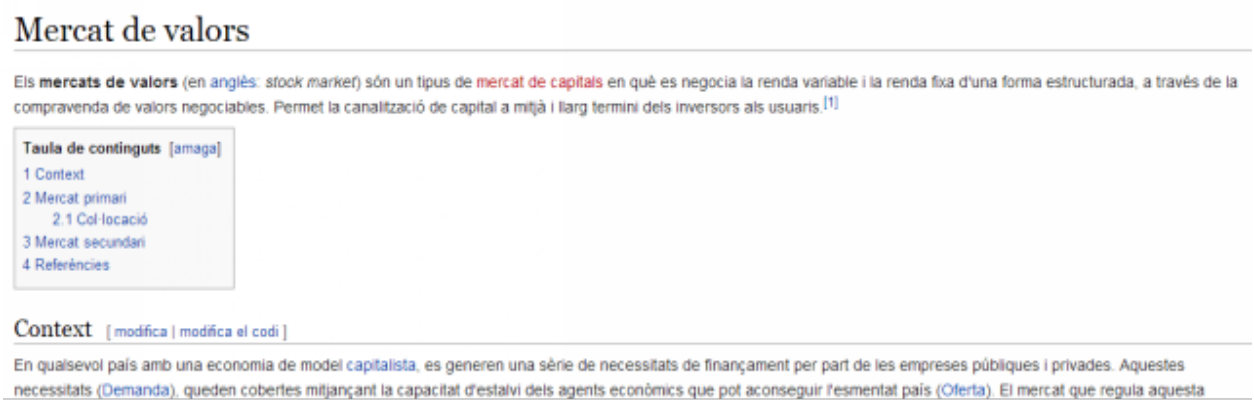
3.6.c. A partir de la Wikipedia

La Wikipedia (http://www.wikipedia.org/) es una enciclopedia multilingüe libre que se ha construido (y se sigue construyendo) de manera colaborativa. Como se trata de un recurso multilingüe puede resultar también una buena fuente de consulta para un traductor. Imaginemos que nos aparece el término *stock market* y que no sabemos cómo traducirlo ni lo hemos encontrado en nuestras bases de datos terminológicas. Podemos ver si en la Wikipedia inglesa existe una entrada para este término:

Cada artículo está relacionado con los artículos equivalentes en otras lenguas. Si nos fijamos. en la parte izquierda aparecen los enlaces interlingüísticos:

2 http://www.parlament.cat/web/documentacio/recursos-documentals/tesaurus#&cl=en
3 http://www.bizkaia.net/kultura/eurovoc/index.asp#&cl=en

Como podemos observar esta entrada está disponible en castellano y podemos ir directamente a la página castellana haciendo clic en el enlace:



Mercat de valors

Els **mercats de valors** (en anglès: *stock market*) són un tipus de mercat de capitals en què es negocia la renda variable i la renda fixa d'una forma estructurada, a través de la compravenda de valors negociables. Permet la canalització de capital a mitjà i llarg termini dels inversors als usuaris.[1]

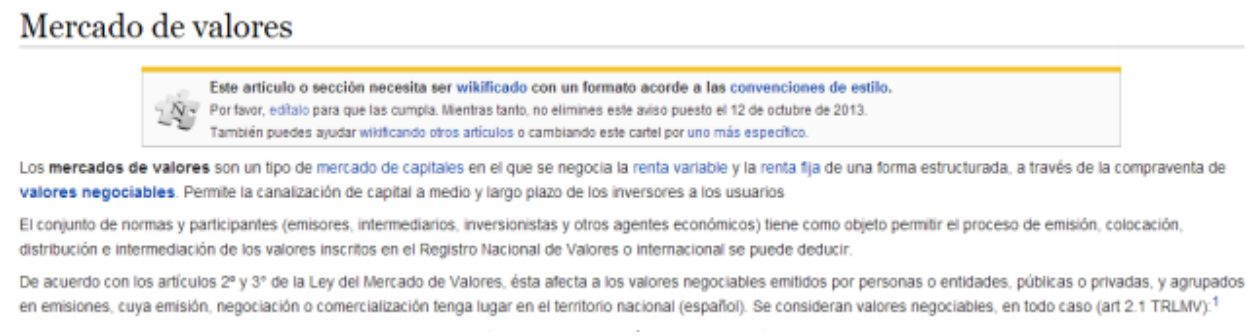
Taula de continguts [amaga]

1 Context
2 Mercat primari
  2.1 Col·locació
3 Mercat secundari
4 Referències

Context [modifica | modifica el codi]

En qualsevol país amb una economia de model capitalista, es generen una sèrie de necessitats de finançament per part de les empreses públiques i privades. Aquestes necessitats (Demanda), queden cobertes mitjançant la capacitat d'estalvi dels agents econòmics que pot aconseguir l'esmentat país (Oferta). El mercat que regula aquesta

[Para la versión castellana e inglesa]



Mercado de valores

Este artículo o sección necesita ser wikificado con un formato acorde a las convenciones de estilo.
Por favor, editalo para que las cumpla. Mientras tanto, no elimines este aviso puesto el 12 de octubre de 2013.
También puedes ayudar wikificando otros artículos o cambiando este cartel por uno más específico.

Los **mercados de valores** son un tipo de mercado de capitales en el que se negocia la renta variable y la renta fija de una forma estructurada, a través de la compraventa de **valores negociables**. Permite la canalización de capital a medio y largo plazo de los inversores a los usuarios

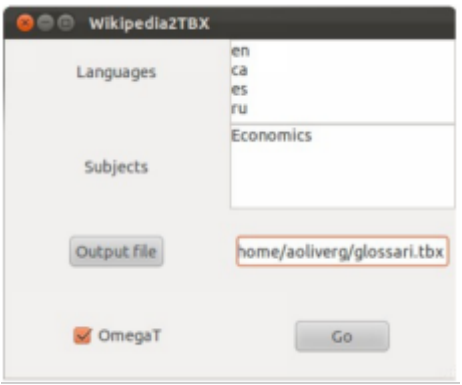
El conjunto de normas y participantes (emisores, intermediarios, inversionistas y otros agentes económicos) tiene como objeto permitir el proceso de emisión, colocación, distribución e intermediación de los valores inscritos en el Registro Nacional de Valores o internacional se puede deducir.

De acuerdo con los artículos 2º y 3º de la Ley del Mercado de Valores, ésta afecta a los valores negociables emitidos por personas o entidades, públicas o privadas, y agrupados en emisiones, cuya emisión, negociación o comercialización tenga lugar en el territorio nacional (español). Se consideran valores negociables, en todo caso (art 2.1 TRLMV):[1]

Y de esta manera determinar que la traducción de *stock market* en castellano puede ser *mercado de valores.*

Esta tarea puede resultar algo pesada, pero existe una sencilla aplicación, Wikipedia2TBX (http://lpg.uoc.edu/Wikipedia2TBX/) que permite la creación de bases de datos terminológicas a partir de la Wikipedia de una manera automática.

Este programa dispone de una interfaz muy sencilla:



A partir de las lenguas de interés (*Languages)* y de una o más áreas de especialidad (*Subjects*) permite crear un glosario terminológico:

```
<?xml version="1.0" ?>
<martif type="TBX" xml:lang="en">
  <text>
    <body>
      <termEntry id="1">
        <descrip type="subjectField">
          Economics
        </descrip>
    <langSet xml:lang="en">
      <tig>
        <term>
          Game theory
        </term>
      </tig>
    </langSet>
    <langSet xml:lang="ca">
      <tig>
        <term>
          Teoria dels jocs
        </term>
      </tig>
    </langSet>
    <langSet xml:lang="es">
      <tig>
        <term>
          Teoría de juegos
         </term>
      </tig>
    </langSet>
    <langSet xml:lang="ru">
      <tig>
```



```
        <term>
             Теория игр
        </term>
      </tig>
    </langSet>
  </termEntry>
```

y en formato tabulado para OmegaT:

```
Game theory Teoria dels jocs Economics
Human rights Drets Humans Economics
Smuggling Contraban Economics
Means of production Mitjans de producció Economics
Poverty Pobresa Economics
Innovation Innovació Economics
Optimism Optimisme Economics
Millionaire Milionari Economics
Liquidity trap Trampa de liquiditat Economics
Break-even Punt mort (economia) Economics
```

Los códigos de lengua que se utilizan son los correspondientes a los códigos ISO de dos letras (ISO 639-1). Estos códigos se pueden consultar en el siguiente enlace: http://es.wikipedia.org/wiki/ISO_639-1

Las áreas de especialidad son las propias de la Wikipedia y se tienen que expresar en inglés. Recordemos, sin embargo, que las áreas de especialidad son libres, es decir, cualquier usuario puede crear nuevas áreas. Para poder conocer qué áreas de especialidad hay es útil consultar el siguiente enlace: http://en.wikipedia.org/wiki/List_of_academic_disciplines

3.6.d. Extracción automática de terminología

A esta técnica de creación de bases de datos terminológicas le dedicaremos un apartado entero. Las técnicas de extracción automática (o quizás mejor dicho semiautomática) de terminología intentan detectar las unidades terminológicas presentes en un texto o conjunto de textos, sin un conocimiento previo de estas unidades. Como veremos en el siguiente apartado, aunque requieren una revisión manual exhaustiva, las técnicas de extracción automática de terminología son muy productivas y permiten la creación rápida de bases de datos terminológicas.

3.7. Extracción automática de terminología

3.7.1. Definición

La extracción automática de terminología es un proceso por el que se detectan una serie de candidatos a unidades terminológicas a partir de un texto o un conjunto de textos. Aunque suele recibir el nombre de extracción automática de terminología el proceso no es totalmente automático, ya que requiere de una revisión manual. Por este motivo muchos autores prefieren hablar de extracción *semiautomática* de terminología. En este libro mantendremos la denominación de automática ya que el proceso pretende ser automático, pero hoy por hoy, los resultados que se obtienen no son lo suficientemente precisos como para poder aceptar directamente el resultado de la extracción.

No hay que confundir el proceso de extracción automática de terminología con el proceso de detección automática de términos. En la detección automática de términos, los términos son conocidos *a priori*, y el sistema intenta determinar qué términos de una base de datos están presentes en el texto que estamos traduciendo. En este caso, la dificultad principal es la detección de formas flexionadas de los términos.

La extracción automática de terminología permite tratar de una manera rápida una gran cantidad de textos, y aunque el proceso de revisión manual es laborioso, se pueden construir bases de datos terminológicas de una manera muy rápida.

3.7.2. Clasificación de métodos para la extracción automática de terminología

En Pazienza (2005) se puede encontrar una explicación detallada de los métodos para la extracción automática de terminología. Los métodos se pueden clasificar en dos grandes grupos:

- *Métodos estadísticos:* la extracción de términos se realiza a partir de sus propiedades estadísticas (la característica más habitual es simplemente la frecuencia de aparición)
- *Métodos lingüísticos:* la extracción de términos se realiza a partir de sus propiedades lingüísticas (habitualmente sus propiedades morfosintácticas)

A menudo también se habla de *métodos híbridos*, que combinan métodos estadísticos y métodos lingüísticos. De hecho, estrictamente todos los métodos para la extracción automática de terminología son híbridos, ya que como veremos los métodos estadísticos hacen uso de una lista de palabras vacías (*stop-words*) y esto es en cierto modo un conocimiento lingüísticos; y los métodos lingüísticos también usan la frecuencia de aparición para ordenar los candidatos, y esta es una propiedad estadística.

La extracción de terminología puede ser *monolingüe,* en la que se extraen términos en una sola lengua; o *bilingüe* (y por extensión *multilingüe*) en la que se extraen términos en una lengua y sus equivalentes de traducción en otra (o en más de una) lengua. En la extracción bilingüe habitualmente se usan corpus paralelos (o memorias de traducción) para buscar de manera automática los equivalentes de traducción.

3.7.3. Métodos estadísticos para la extracción automática de terminología

Los *métodos estadísticos* para la extracción automática de terminología son aquellos que utilizan principalmente información estadística para detectar candidatos a términos.

Los métodos estadísticos se basan en el cálculo de *n-gramas*. Un *n-grama* es una combinación de *n* elementos. En el caso de la extracción automática de terminología estos elementos son palabras (o *tokens*). Observemos el siguiente ejemplo:

> Thus, maintaining stable prices is the only feasible objective for the single **monetary policy** over the medium term.

En esta oración tenemos (como mínimo) un término: *monetary policy*. Si calculamos los *bigramas* (*n-gramas* de orden 2), es decir, las combinaciones de dos palabras (o mejor dicho *tokens,* ya que algunas combinaciones incluyen elementos que no son palabras, como por ejemplo signos de puntuación), obtenemos:

bigramas:

```
[('Thus', ','), (',', 'maintaining'), ('maintaining', 'stable'), ('stable' prices '), (' prices
',' is ') , ('is', 'the'), ('the', 'only'), ('only', 'feasible'), ('feasible', 'objective'),
('objective', 'for') , ('for', 'the'), ('the', 'single'), ('single', 'monetary'), ('monetary',
'policy'), ('policy', 'over') , ('over', 'the'), ('the', 'medio'), ('medium', 'term'), ('term',
'.')]
```

Este cálculo se puede hacer de una manera muy sencilla con el lenguaje de programación Python y la librería NLTK *(Natural Language Toolkit).Vemos a continuación el código:*

```
import nltk

sentence=["Thus",",","maintaining","stable","prices","is", "the","only",
"feasible","objective","for",
"the" ,"single","monetary","policy","over","the","medium","term","."]

ngramsfrase=nltk.util.ngrams(sentence, 2, pad_left=False, pad_right=False, pad_symbol=None)

print ngramsfrase
```

En este caso la oración la tenemos ya *tokenizada* dentro del código. En otros ejemplos veremos cómo podemos hacer esta *tokenización.*

Es evidente que si el término presente en la oración está formado por dos palabras y calculamos los *bigramas* de esta oración, nuestro término estará presente en la lista de bigramas. Lo que pasa es que también obtenemos muchas otras combinaciones que no son terminológicas, cosas del estilo: *is the, the only,* etc. Como en este caso sólo hemos hecho la extracción a partir de una única oración, la información estadística como por ejemplo la frecuencia de aparición no es relevante (ya que todos los bigramas aparecen una única vez).

Si ahora hacemos lo mismo pero calculamos además de *bigramas* también *trigramas (n-gramas* de orden 3) de un corpus más grande (en el ejemplo un subconjunto de 10.000 oraciones del corpus ECB (European Central Bank) del inglés[4] (Tiedemann 2009), y ordenamos los candidatos por frecuencia, obtendremos los siguientes resultados (se muestran los 25 más frecuentes). Veamos en primer lugar el código en Python:

```
import nltk

reader = nltk.corpus.reader.plaintext.PlaintextCorpusReader (".", 'ecb-10K-en.txt',
encoding="utf-8")

words=reader.words()

#bi-grams

bigramsfrase=nltk.util.ngrams(words, 2, pad_left=False, pad_right=False, pad_symbol=None)
```

4 http://opus.lingfil.uu.se/ECB.php


```
#trigrams
trigramsfrase=nltk.util.ngrams(words, 3, pad_left=False, pad_right=False, pad_symbol=None)

fdist = nltk.probability.FreqDist()
for bigram in bigramsfrase:
    fdist.inc(" ".join(bigram))
for trigram in bigramsfrase:
    fdist.inc(" ".join(trigram))

#show the 25 more frequent bigrams and trigrams
cont=0
for ngram in fdist.keys():
    print fdist[ngram],ngram
    cont+=1
    if cont==25:
        break
```

Y ahora los resultados que se obtienen:

```
7198 of the
2586 CON /
2474 amp;
2470 to the
2462 & amp
2410 on the
2388 gt;
2382; gt
2086, pdf
2082 kB,
1948 in the
1886. The
1848, the
1748 Opinion donde
1742 the ECB
1680 (CON
1610 & apos
1570 apos;
1402 by the
1398; s
1,210 the European
1204 for the
1.152 and the
1138, en
1074 ECB /
```


Como podemos observar, aunque hemos calculado bigramas y trigramas, entre los 25 primeros "candidatos" no obtenemos trigramas, ya que las unidades más cortas son las más frecuentes. Todo lo que obtenemos son combinaciones de palabras funcionales, puntuaciones y menudo elementos que provienen de errores de conversión de formatos. Por tanto, lo que hemos obtenido no se parece ni mucho menos a una extracción de terminología.

Filtrado por palabras vacías (*stop-words*)

Para poder obtener esta lista en una lista de candidatos a términos será necesario filtrarla con una lista de *palabras vacías* (o *stop-words).*Estas listas contienen palabras funcionales y otros que no suelen aparecer ni en primera ni en última posición de un término.

Una lista de palabras vacías del inglés contendría palabras tales como: *a, a's, able, about, above, according, accordingly, across, actually, after, afterwards, again, against, ain't, all, allow, allows, almost, alone, along, already…*

El filtrado consistirá en eliminar de la lista de candidatos aquellos que empiezan o terminan con una palabra de la lista de palabras vacías. A continuación podemos observar el código y el resultado para el caso de la frase simple del primer ejemplo:

```
import nltk
import codecs
sentence=["Thus",",","maintaining","stable","prices","is", "the","only",
"feasible","objective","for",
"the" ,"single","monetary","policy","over","the","medium","term","."]
stopfile=codecs.open("stop-eng.txt","r",encoding="utf-8")
stopwords=[]
for line in stopfile.readlines():
    line=line.rstrip()
    stopwords.append(line)
stopwords.extend([".",",",";",":","?","!","'","\"","(",")","-","&","/"])
ngramsfrase=nltk.util.ngrams(sentence, 2, pad_left=False, pad_right=False, pad_symbol=None)
for ngram in ngramsfrase:
    if not ngram[0].lower() in stopwords and not ngram[1].lower() in stopwords:
        print ngram
```

Y ahora obtenemos los siguientes resultados:

```
('maintaining', 'stable')
('stable', 'prices')
('feasible', 'objective')
('single', 'monetary')
('monetary', 'policy')
('medium', 'term')
```

donde ya si que aparece nuestro término junto con otras unidades no terminológicas, pero donde el número de candidatos se ha reducido notablemente.

Si aplicamos el mismo filtrado al corpus de 10.000 oraciones del ECB corpus obtenemos:

```
import nltk
import codecs
```



```
stopfile=codecs.open("stop-eng.txt","r",encoding="utf-8")
stopwords=[]
for line in stopfile.readlines():
    line=line.rstrip()
    stopwords.append(line)
stopwords.extend([".",",","?","!","'","\"","(",")","-","&","/","[","]"])
reader = nltk.corpus.reader.plaintext.PlaintextCorpusReader(".", 'ecb-10K-en.txt', encoding="utf-8")
words=reader.words()
#bi-grams
bigramsfrase=nltk.util.ngrams(words, 2, pad_left=False, pad_right=False, pad_symbol=None)
#trigrams
trigramsfrase=nltk.util.ngrams(words, 3, pad_left=False, pad_right=False, pad_symbol=None)
fdist = nltk.probability.FreqDist()
for bigram in bigramsfrase:
    if not bigram[0] in stopwords and not bigram[1] in stopwords:
        fdist.inc(" ".join(bigram))
for trigram in bigramsfrase:
    if not trigram[0] in stopwords and not trigram[-1] in stopwords:
        fdist.inc(" ".join(trigram))
#show the 25 more frequent bigrams and trigrams
cont=0
for ngram in fdist.keys():
    print fdist[ngram],ngram
    cont+=1
    if cont==25:
        break
```

Y los siguientes candidatos:

```
960 Central Bank
838 European Central
662 euro area
630 Governing Council
508 Navigation Path
480 OJ L
478 The European
444 Member States
412 Legal framework
406 AL group
384 payment orders
362 OJ C
362 monetary policy
318 --- «
304 en Opinion
290 Opinion CON
```



```
280 European Union
268 EN ---
266 --- EN
266 001 ---
264 ▼ B
252 PM account
252 central banks
250 --- 22
244 payment order
```

Que dista aún mucho de ser perfecta pero que ha mejorado mucho: aparecen ya algunos términos, como *monetary policy* y *payment order* (que aparece tanto en plural como en singular), o *central banks* (sólo en plural) y *euro area.* Aparecen todavía algunas combinaciones de símbolos no deseadas (que se pueden eliminar añadiendo este símbolos como palabras de la lista de palabras vacías). También aparecen términos partidos, como sería *European Central Bank* que aparece como *Central Bank* y *European Central*. Un poco más adelante veremos estrategias para intentar arreglar todos estos resultados incorrectos.

TBXTools

El resto de ejemplos de esta sección los haremos haciendo referencia a la herramienta TBXTools, que es una clase escrita en Python que implementa muchos métodos de extracción automática de terminología y otras utilidades relacionadas con la gestión de la terminología. Esta herramienta se puede descargar de http://lpg.uoc.edu/TBXTools. Esta clase facilita enormemente la implementación de programas para la extracción automática de terminología. A continuación vemos el código correspondiente al ejemplo anterior pero escrito utilizando esta clase:

```
from TBXTools import *
import codecs

ecbterms=TBXTools()
ecbterms.load_sl_corpus("ecb-10K-en.txt")
ecbterms.load_stop_l1("stop-eng.txt")
ecbterms.statistical_term_extraction()
ecbterms.show_term_candidates(fmin=2)
ecbterms.save_term_candidates("termcandidates-5.txt",fmin=2)
```

Palabras vacías internas

Si analizamos a fondo los candidatos a términos obtenidos con el programa anterior veremos algunos ejemplos como:

```
53     banknotes and coins
40     Capital and reserves
29     clearing and settlement
9      approval or accession
7      directly or indirectly
6      suspension or termination
5      State or Government
```

que corresponden a candidatos erróneos no filtrados por palabras vacías, ya que habitualmente se descartan los candidatos que empiezan o terminan por palabras de la lista de palabras vacías. En estos ejemplos vemos que son trigramas en el que la palabra del medio es una conjunción. TBXTools permite definir una lista de palabras vacía interna, que utilizará para eliminar aquellos candidatos que tengan en su

interior (cualquier posición excepto la primera y la última) un palabra de esta lista. Una posible lista para el inglés estaría compuesta por las siguientes palabras:

```
and
but
or
nor
so
```

A continuación podemos ver un ejemplo de código que utiliza este filtrado:

```
ecbterms=TBXTools()
ecbterms.load_sl_corpus("ecb-10K-en.txt")
ecbterms.load_stop_l1("stop-eng.txt")
ecbterms.load_inner_stop_l1("stop-inner-eng.txt")
ecbterms.statistical_term_extraction()
ecbterms.show_term_candidates(fmin=2)
ecbterms.save_term_candidates("termcandidates-5.txt",fmin=2)
```

Normalización de mayúsculas y minúsculas

Si miramos a fondo el resultado de una extracción puramente estadística a menudo nos encontramos con resultados como:

```
181 monetary policy
34 Monetary policy
6 Monetary Policy
1 MONETARY POLICY
```

es decir, con el mismo término escrito con diferentes versiones según las mayúsculas y minúsculas. En realidad, desearíamos que nuestro sistema juntara todos estos candidatos bajo un único candidato y que sumara las frecuencias:

```
222 monetary policy
```

TBXTools (y muchas otras herramientas de extracción automática de terminología) es capaz de llevar a cabo esta normalización. Para ello, simplemente verifica si hay varias realizaciones de un mismo término que se diferencien sólo por el hecho de estar escritas con mayúsculas o minúsculas. A continuación vemos un ejemplo de código que implementa esta función:

```
from TBXTools import *
import codecs

ecbterms=TBXTools()
ecbterms.load_sl_corpus("ecb-10K-en.txt")
ecbterms.load_stop_l1("stop-eng.txt")
ecbterms.statistical_term_extraction()
ecbterms.case_normalization()
ecbterms.show_term_candidates(fmin=2)
ecbterms.save_term_candidates("termcandidates-6.txt",fmin=2)
```

Normalización morfológica

Entre los candidatos a términos nos encontraremos variantes morfológicas de los diferentes términos. Por ejemplo:

```
213 payment orders
122 payment order
```

cuando quisiéramos obtener la forma base y las frecuencias sumadas:

```
335 payment order
```

o bien:

```
3 economic policies
2 economic policy
```

cuando en realidad quisiéramos obtener:

```
5 economic policy
```

Dado que las técnicas estadísticas no utilizan demasiada información lingüística, el sistema no es capaz de saber que estas realizaciones son en realidad el mismo término. TBXTools permite definir una serie de reglas morfológicas sencillas que permiten al sistema juntar diferentes formas del mismo término. Para funcionar necesita una serie de reglas del estilo:

```
:s:L
:es:L
y:ies:L
```

La "L" de la regla significa que es el último elemento (Last) del n-grama. El primer campo (nulo para las dos primeras reglas e "y" para la tercera) es la terminación de lema; y el segundo elemento ("s", "es" i "ies") es la terminación de forma.

Y el código que llama a esta función:

```
from TBXTools import *
import codecs

ecbterms=TBXTools()
ecbterms.load_sl_corpus("ecb-10K-en.txt")
ecbterms.load_stop_l1("stop-eng.txt")
ecbterms.load_morphopatterns("morphopatterns-eng.txt")
ecbterms.statistical_term_extraction()
ecbterms.case_normalization()
ecbterms.morpho_normalization()
ecbterms.save_term_candidates("termcandidates-6.txt",fmin=2)
```

Estas reglas pueden ser útiles para lenguas con morfología simple, como por ejemplo el inglés. Para otras lenguas, la cantidad de reglas puede ser realmente importante y su definición muy compleja, por lo que es más recomendable utilizar las técnicas lingüísticas que explicaremos más adelante.

Detección de candidatos anidados

Cuando obtenemos candidatos a término de un orden *n* a menudo sucede que en realidad son parte de un término de algún orden superior (*n + 1* o incluso de un orden más alto). En el ejemplo que estamos trabajando, el candidato bigrama

```
419 European Central
```

es en realidad un fragmento del candidato trigrama

```
417 European Central Bank
```

TBXTools implementa la detección de términos anidados, añadiendo al código del programa la siguiente línea:

```
ecbterms.nest_normalization(verbose=True,percent=10)
```


entonces la herramienta verifica los posibles anidados. El porcentaje (que en el ejemplo está fijado a 10) se refiere a la diferencia máxima de frecuencias entre el candidato de orden *n* y el superior (en este caso el 10%). El sistema permite detectar anidados como los siguientes:

```
419 European Central --> 417 European Central Bank
134 group member --> 132 AL group member
119 national central --> 117 national central bank
83 area residents --> 83 euro area residents
83 area residents --> 82 area residents denominated
83 area residents --> 82 euro area residents denominated
83 euro area residents --> 82 euro area residents denominated
82 area residents denominated --> 82 euro area residents denominated
82 residents denominated --> 82 euro area residents denominated
79 week due --> 76 week due to transactions
46 Related ECB --> 45 Related ECB opinions
```

Detección de términos monopalabra (1-gramas o unigramas)

Si nos hemos fijado, habremos visto que las técnicas estadísticas se basan en el cálculo de combinaciones de palabras (de dos o más palabras). Esto no permite detectar términos monopalabra (unigramas, es decir, términos formados por una única palabra). Si calculamos los 1-gramas obtendremos todas las palabras del texto y si filtramos esta lista con palabras vacías simplemente eliminaremos las palabras vacías de la lista de candidatos. En el ejemplo que estamos trabajando obtendríamos una lista como la siguiente:

1848 ECB	425 OJ	800 central bank
1293 CON	424 insert	419 European Central
1237 amp	401 CB	417 European Central Bank
1194 gt	382 participant	335 euro area
1048 Opinion	381 national	335 payment order
1046 pdf	365 EC	315 Governing Council
1042 kB	359 credit	254 Navigation Path
964 European	355 area	249 legal framework
852 apos	339 monetary	230 PM account
844 euro	334 Regulation	225 member states
798 Article	331 system	222 monetary policy
784 Council	323 Governing	203 AL group
763 payment	323 accounts	145 Opinion CON
622 en	323 information	140 European Union
604 Bank	320 Member	139 credit institution
559 Central	318 framework	134 group member
512 Eurosystem	309 banks	
485 financial	306 institutions	
468 SEPA	303 AL	
437 settlement	297 policy	
433 account	290 EUR	
	287 payments	

Como vemos, pues, la aproximación estadística que hemos descrito no sirve para extraer términos monopalabra de un corpus. Se pueden seguir dos estrategias:

- Para obtener los términos monopalabra, el revisor humano aislará los candidatos que considere interesantes mientras revisa candidatos de orden superior .Por ejemplo, si a su lista de candidatos encuentra: *payment order* puede decidir que *payment* también es un término relevante de la disciplina.
- Para ciertas especialidades, por ejemplo medicina, hay una gran cantidad de cultismos formados por sufijos específicos (por ejemplo *-itis* en medicina). El sistema puede detectar palabras terminadas en estos sufijos y incluirlas en la lista de candidatos.

Detección de equivalentes de traducción en corpus paralelos

Si disponemos de un corpus paralelo, además de extraer candidatos de traducción en una de las lenguas, podemos detectar de manera automática los candidatos traducidos (es decir, las traducciones que se han utilizado en el mismo corpus) . Veamos el siguiente ejemplo, donde podemos observar las oraciones donde aparece el término en inglés *legal framework* y las correspondientes oraciones traducidas con el correspondiente equivalente de traducción *marco jurídico.*

Opinion on amending the **legal framework** for clearing operations (CON / 2009/66)	Dictamen sobre la reforma del **marco jurídico** de las operaciones de compensación (CON / 2009/66)
The proposed directive is a very welcome initiative , as it establishes a comprehensive **legal framework** for payment services in the EU .	La propuesta de directiva se acoge con gran satisfacción , pues establece un **marco jurídico** integral de los servicios de pago en la UE .
The Directive will greatly facilitate the operational implementation of SEPA instruments by the banking industry , as well as their adoption by end-users , by harmonising the applicable **legal framework** . This will provide the foundation for a single « domestic » euro payments market .	La Directiva , al armonizar el **marco jurídico** aplicable , facilitará en gran medida la aplicación operativa de los instrumentos de la SEPA en el sector bancario , además de su adopción por parte de los usuarios finales , lo que sentará las bases para un mercado único « interno » de pagos en euros

El proceso estadístico para el cálculo del equivalente de traducción en un corpus paralelo es muy sencillo:

- Se buscan todos los pares de oraciones donde aparezca el término a buscar en la parte correspondiente a la lengua de partida y guardamos las oraciones correspondientes a la lengua de llegada.
- Con todas las oraciones de la lengua de llegada que hemos guardado llevamos a cabo un proceso de extracción de terminología y filtramos los resultados con la lista de palabras vacías correspondiente a la lengua de llegada.
- El candidato a término que aparezca con más frecuencia será el candidato a equivalente de traducción que buscamos.

En este ejemplo, si extraemos los candidatos de la parte castellana y filtramos por stopwords, obtenemos los siguientes resultados:

```
3       marco jurídico
1       Dictamen sobre la reforma
1       adopción por parte
1       aplicación operativa
1       armonizar el marco
1       armonizar el marco jurídico
```

Donde el candidato más frecuente, *marco jurídico*, es realmente el equivalente de traducción del término buscado. Para que esta técnica pueda funcionar bien deben darse las siguientes circunstancias:

- El término a buscar debe aparecer con cierta frecuencia (si aparece sólo una vez será difícil obtener el equivalente de traducción).
- En el corpus se debe haber usado una traducción más o menos estable del término (si cada oración se ha utilizado una traducción diferente el sistema no podrá detectar la traducción correcta)
- Las oraciones donde aparece el término deben ser diferentes (si un término aparece 100 veces en un corpus pero es una misma oración que se repite mucho, el efecto es exactamente el mismo de que la oración aparezca sólo una vez)

TBXTools implementa la función de búsqueda en corpus paralelos. A continuación podemos observar un ejemplo de código. En corpus muy grandes, conviene fijar un número de veces máximo de apariciones

del término, para poder parar su búsqueda antes de explorar todo el corpus. Este comportamiento se controla con el parámetro *limitsents* que por defecto tiene el valor de 100.

```
from TBXTools import *
import codecs

ecbterms=TBXTools()
ecbterms.load_tabtxt_corpus("ecb-eng-spa.txt")
ecbterms.load_stop_l2("stop-spa.txt")
candidats=codecs.open("termes-ecb-eng.txt","r",encoding="utf-8")
sortida=codecs.open("termes-ecb-traduits-brut-eng-spa.txt","w",encoding="utf-8")
for c in candidats.readlines():
    c=c.rstrip()
    print "Candidate:",c
    tr=ecbterms.get_statistical_translation_candidate(c, candidates=5, limitsents=50)
    cadena=c+"\t"+tr
    print cadena
    sortida.write(cadena+"\n")
```

El programa puede dar más de una opción, de modo que si no acierta el revisor pueda aceptar alguna de las otras opciones. En el siguiente ejemplo el número de opciones estaba fijado en 5:

legal framework -> marco jurídico:régimen jurídico:Bancos Centrales:Europeo de Bancos:Orientación BCE
payment order -> orden de pago:órdenes de pago:módulo de pagos:banco central:miembros del grupo
credit institution -> entidad de crédito:módulo de pagos:banco central:Bonos del Estado:participante directo
price stability -> estabilidad de precios:medio plazo:política monetaria:Consejo de Gobierno:precios a medio

3.7.4.Métodos lingüísticos para la extracción automática de terminología

Los métodos lingüísticos para la extracción automática de terminología utilizan características lingüísticas de los términos para llevar a cabo la detección. La característica más utilizada son los *patrones morfosintácticos*. Los patrones morfosintácticos son combinaciones de etiquetas morfosintácticas que definen combinaciones que son típicamente terminológicas. Así por ejemplo (para el inglés) podríamos definir una serie de patrones:

```
NN NN
JJ NN
NN /of/ NN
```

NN NN indica una combinación de Nombre y Nombre (ambos en singular). JJ NN indica una combinación de adjetivo y nombre, y NN / of / NN indica una combinación de nombre, la palabra *of* y otro nombre. Algunos términos que siguen el patrón NN NN podrían ser *payment order* y *interest rate;* ejemplos de JJ NN serían *monetary policy* y *direct debit* y ejemplos de NN / of / NN serían *economy of scale* y *point of sale.* Hay que tener en cuenta que no todas las combinaciones que cumplan estos patrones serán realmente términos, sino que se detectarán también muchas otras combinaciones no terminológicas.

Para poder llevar a cabo extracción de terminología siguiendo la metodología lingüística será imprescindible disponer de un etiquetador morfosintáctico para la lengua de trabajo. En el apartado *Para ampliar conocimientos* del capítulo 2 de este mismo libro hablamos de estas herramientas. A continuación vemos un ejemplo de texto etiquetado morfosintácticamente utilizando Freeling:

```
in|in|IN|0.985534 the|the|DT|1 underlying|underlie|VBG|1 transaction|transaction|NN|1 (|(|Fpa|1
s|s|NNS|0.639593 )|)|Fpt|1 or|or|CC|1 payment|payment|NN|1 order|order|NN|0.909224 (|(|Fpa|1 s|s|
NNS|0.639593 )|)|Fpt|1 arising|arise|VBG|1 from|from|IN|1 criminal|criminal|JJ|0.959559 offences|
offence|NNS|1 or|or|CC|1
```

Para cada palabra del texto tenemos la forma, el lema, una etiqueta morfosintáctica y la probabilidad de que esta etiqueta sea la correcta (recordemos que muchas palabras son ambiguas desde el punto de vista morfosintáctico). Las etiquetas morfosintácticas expresan mediante un etiquetario (*tagset*) que a menudo

es dependiente de la lengua. A continuación observamos la etiquetario del Penn Treebank (Santorini 1990) para el inglés, que es el que usa el analizador Freeling para el inglés.

Table 2
The Penn Treebank POS tagset.

1. CC	Coordinating conjunction	25. TO	*to*
2. CD	Cardinal number	26. UH	Interjection
3. DT	Determiner	27. VB	Verb, base form
4. EX	Existential *there*	28. VBD	Verb, past tense
5. FW	Foreign word	29. VBG	Verb, gerund/present participle
6. IN	Preposition/subordinating conjunction	30. VBN	Verb, past participle
7. JJ	Adjective	31. VBP	Verb, non-3rd ps. sing. present
8. JJR	Adjective, comparative	32. VBZ	Verb, 3rd ps. sing. present
9. JJS	Adjective, superlative	33. WDT	*wh*-determiner
10. LS	List item marker	34. WP	*wh*-pronoun
11. MD	Modal	35. WP$	Possessive *wh*-pronoun
12. NN	Noun, singular or mass	36. WRB	*wh*-adverb
13. NNS	Noun, plural	37. #	Pound sign
14. NNP	Proper noun, singular	38. $	Dollar sign
15. NNPS	Proper noun, plural	39. .	Sentence-final punctuation
16. PDT	Predeterminer	40. ,	Comma
17. POS	Possessive ending	41. :	Colon, semi-colon
18. PRP	Personal pronoun	42. (	Left bracket character
19. PP$	Possessive pronoun	43.)	Right bracket character
20. RB	Adverb	44. "	Straight double quote
21. RBR	Adverb, comparative	45. '	Left open single quote
22. RBS	Adverb, superlative	46. "	Left open double quote
23. RP	Particle	47. '	Right close single quote
24. SYM	Symbol (mathematical or scientific)	48. "	Right close double quote

Otras lenguas usan etiquetarios diferentes. El Castalla y catalán a Freeling, por ejemplo, utilizan las etiquetas EAGLES:

```
La|el|DA0FS0|0.972269 creación|creación|NCFS000|1 de|de|SPS00|0.999984 la|el|DA0FS0|0.972269
Zona_Única_de_Pagos|zona_única_de_pagos|NP00000|1 en|en|SPS00|1 Euros|euros|NP00000|1 (|(|Fpa|1
SEPA|sepa|NP00000|0.241915 )|)|Fpt|1 ,|,|Fc|1 cuyo|cuyo|PR0MS000|1 objetivo|objetivo|NCMS000|
0.809524 es|ser|VSIP3S0|1 elimi|elimi|RG|0.751649 nar|nar|VMN0000|0.917775 los|el|DA0MP0|0.976481
obstáculos|obstáculo|NCMP000|1 para|para|SPS00|0.999103 los|el|DA0MP0|0.976481 pagos|pago|
NCMP000|1 en|en|SPS00|1 euros|euro|NCMP000|1 en|en|SPS00|1 un|uno|DI0MS0|0.987295 área|área|
NCFS000|1 que|que|PR0CN000|0.562517 actualmente|actualmente|RG|1 comprende|comprender|VMIP3S0|
0.96875 31|31|Z|1 países|país|NCMP000|1 ,|,|Fc|1 sigue|seguir|VMIP3S0|0.996454 avanzando|avanzar|
VMG0000|1 .|.|Fp|1
```

Formalismo para los patrones en TBXTools

TBXTool utiliza un formalismo muy potente para la expresión de los patrones terminológicos. Para explicar este formalismo lo haremos a través de una serie de ejemplos en inglés, catalán y castellano. Empecemos por el inglés. Consideramos que tenemos un texto como el siguiente:

> The **interest rate** on the marginal lending facility will be reduced by 50 basis points to 4.75%, with immediate effect. This issue of the Monthly Bulletin was finalised before the Governing Council's decision to cut the key ECB **interest rates** and to change the tender procedure and the standing facilities corridor on 8 October 2008.

Para poder hacer la extracción lingüística necesitamos disponer de este texto etiquetado:

```
The|the|DT|1 interest|interest|NN|0.995923 rate|rate|NN|0.995511 on|on|IN|0.971769 the|the|DT|1
marginal|marginal|JJ|1 lending|lend|VBG|1 facility|facility|NN|1 will|will|MD|0.989422 be|be|VB|1
reduced|reduce|VBN|0.626689 by|by|IN|0.997664 50|50|Z|1 basis|basis|NN|1 points|point|NNS|
0.934153 to|to|TO|0.999991 4.75_%|4.75/100|Zp|1 ,|,|Fc|1 with|with|IN|0.999953 immediate|
immediate|JJ|1 effect|effect|NN|0.985594 .|.|Fp|1  This|this|DT|0.99991 issue|issue|NN|0.924989
of|of|IN|0.999898 the|the|DT|1 Monthly_Bulletin|monthly_bulletin|NP|1 was|be|VBD|1 finalised|
finalise|VBN|0.41625 before|before|IN|0.918909 the|the|DT|1 Governing_Council|governing_council|
NP|1 's|'s|POS|0.747266 decision|decision|NN|1 to|to|TO|0.999991 cut|cut|VB|0.435206 the|the|DT|1
key|key|JJ|0.693262 ECB|ecb|NP|1 interest|interest|NN|0.995923 rates|rate|NNS|0.995158 and|and|
CC|1 to|to|TO|0.999991 change|change|VB|0.379737 the|the|DT|1 tender|tender|NN|0.805556
procedure|procedure|NN|1 and|and|CC|1 the|the|DT|1 standing|standing|NN|0.641463 facilities|
```



```
facility|NNS|1 corridor|corridor|NN|1 on|on|IN|0.971769 8_October_2008|[??:8/10/2008:??.??:??]|W|
1
```

Los patrones terminológicos se representan como una secuencia de etiquetas morfosintácticas. Si como patrón ponemos **NN NN** (es decir *Noun Singular or mass* seguido de *Noun Singular or mass)* obtenemos los siguientes candidatos (todos con frecuencia 1):

```
1       interest rate
1       tender procedure
```

Nótese que como hemos fijado los patrones a NN sólo detecta los nombres en singular. Si queremos detecte también nombres en plural podemos hacer uso de las expresiones regulares e indicar el siguiente patrón: **NN NN.***. Ahora obtendremos los siguientes candidatos:

```
1       basis points
1       interest rate
1       interest rates
1       standing facilities
1       tender procedure
```

También nos interesaría agrupar las formas singulares y plurales de un mismo lema y que al término detectado expresara el lema. El formalismo nos lo permite hacer con los paréntesis cuadrados []: **NN [NN.*]** En uso de este patrón obtendríamos los siguientes candidatos:

```
2       interest rate
1       basis point
1       standing facility
1       tender procedure
```

Fijémonos que hemos juntado *interest rate* y *interest rates* en una única forma base, pero contando como frecuencia 2, ya que aparece dos veces.

Ahora, para explicar otra característica del formalismo pasaremos a ofrecer un ejemplo del castellano. Consideramos que tenemos el siguiente texto:

> El BCE seguirá reconduciendo la liquidez hacia una situación equilibrada , de forma coherente con el objetivo de mantener los **tipos de interés** a corto plazo en un nivel próximo al **tipo de interés** de la operación principal de financiación .

Y su correspondiente análisis morfosintáctico:

```
El|el|DA0MS0|1 BCE|bce|NP00000|1 seguirá|seguir|VMIF3S0|1 reconduciendo|reconducir|VMG0000|1 la|
el|DA0FS0|0.972269 liquidez|liquidez|NCFS000|1 hacia|hacia|SPS00|1 una|uno|DI0FS0|0.951575
situación|situación|NCFS000|1 equilibrada|equilibrar|VMP00SF|1 ,|,|Fc|1 de|de|SPS00|0.999984
forma|forma|NCFS000|0.970944 coherente|coherente|AQ0CS0|1 con|con|SPS00|1 el|el|DA0MS0|1
objetivo|objetivo|NCMS000|0.809524 de|de|SPS00|0.999984 mantener|mantener|VMN0000|1 los|el|
DA0MP0|0.976481 tipos|tipo|NCMP000|1 de|de|SPS00|0.999984 interés|interés|NCMS000|1 a|a|SPS00|
0.996023 corto|corto|AQ0MS0|0.97619 plazo|plazo|NCMS000|1 en|en|SPS00|1 un|uno|DI0MS0|0.987295
nivel|nivel|NCMS000|1 próximo|próximo|AQ0MS0|1 a|a|SPS00|1 el|el|DA0MS0|1 tipo|tipo|NCMS000|1 de|
de|SPS00|0.999984 interés|interés|NCMS000|1 de|de|SPS00|0.999984 la|el|DA0FS0|0.972269 operación|
operación|NCFS000|1 principal|principal|AQ0CS0|0.986111 de|de|SPS00|0.999984 financiación|
financiación|NCFS000|1 .|.|Fp|1
```

Los patrones pueden expresar palabras en vez de etiquetas utilizando / /. Por ejemplo, el patrón [**NC.*] /de/ NC.*** detectaría el siguiente candidato:

```
2       tipo de interés
```


con frecuencia 2, ya que aparece tanto en singular como en plural. Recordad que las etiquetas morfosintácticas son dependientes de la lengua y que en el caso del castellano son diferentes que para el inglés.

En una misma extracción habitualmente se utiliza un conjunto de patrones morfosintácticos. Por ejemplo, para el inglés se podrían utilizar estos:

```
NN [NN.*]
JJ [NN.*]
VBG NN.*
[NN.*] /for/ JJ NN
[NN.*] /of/ JN NN.*
[NN.*] /for/ VBG
[NN.*] TO NNS NN.*
[NN.*] TO NN.*
[NN.*] /of/ NN.*
```

Y obtendríamos unos candidatos como los siguientes:

331	euro area	55	foreign currency	36	available liquidity
314	payment order	54	lending facility	36	exchange rate
263	insert name	53	excessive deficit	36	payment instrument
181	monetary policy	51	payment institution	35	card scheme
178	central bank	50	policy operation	35	reverse operation
136	credit institution	50	settlement system	34	foreign exchange
134	group member	49	s website	32	border payment
106	interest rate	48	financial statement	32	deposit facility
102	minimum reserve	45	asset item	32	medium term
99	euro banknotes	45	financial market	31	securities settlement
88	settlement bank	45	group manager	30	Having regard
83	area resident	44	payment system	30	deficit procedure
81	payment instruction	43	legal framework	30	payment service
79	last week	43	reserve asset	30	policy decision
78	country reference	43	third party	29	network service
78	external auditor	42	maintenance period	29	securities account
72	refinancing operation	41	direct debit	29	system settlement
70	ancillary system	41	insert reference	29	territorial unit
67	component system	41	management service	27	euro coin
65	price stability	41	reserve management	27	financial instrument
63	financial institution	40	credit transfer	27	member of staff
61	Additional information	40	direct participant	27	national law
61	settlement procedure	40	reserve requirement	26	account agreement
60	liability item	40	service provider	26	financial stability
57	foreign reserve	39	financial sector	26	main refinancing
57	intraday credit	38	balance sheet	26	relevant intermediary

56	business day	38	statistical information	26	single currency

Detección de términos anidados

En el caso de la extracción lingüística nos podemos encontrar también en casos de términos anidados, como ocurría con las estrategias estadísticas. La ventaja es que la búsqueda de los anidamientos puede ir más guiada, ya que a partir de la observación de los patrones terminológicos se puede prever qué casos de anidamiento se producirán.

Detección de términos monopalabra

La detección de términos monopalabra (es decir, aquellos términos formados por una única palabra) también supone un problema para las técnicas lingüísticas. El patrón típico para el inglés sería **[NN. *]**, Pero este patrón detectaría todos los sustantivos del texto.

Detección de equivalentes de traducción usando estrategia estadística

La búsqueda de equivalentes de traducción se puede llevar a cabo con una estrategia lingüística. El principal problema se produce debido a que un mismo patrón terminológico en la lengua de partida puede corresponder a varios patrones de traducción para la lengua de llegada.

Patrón eng	**Ejemplo**	**Traducción spa**	**Patrón traducción spa**
NN [NN.*]	credit institution	institución de crédito	[NC.*] /de/ NC.*
NN [NN.*]	business day	día laborable	[NC.*] [AQ.*]
NN [NN.*]	euro zone	zona euro	[NC.*] NC.*

Por este motivo, aunque la extracción terminológica se lleve a cabo con una estrategia lingüística, a menudo la búsqueda de equivalentes de traducción se hace siguiendo la estrategia lingüística

3.7.5. Medidas estadísticas

Hasta ahora hemos ordenado los resultados de la extracción de terminología, tanto para la estrategia lingüística como para la estadística, por frecuencia de aparición. La frecuencia de aparición, a pesar de ser una medida estadística muy simple, ha demostrado ser muy efectiva en extracción de terminología.

Existen toda una serie de medidas estadísticas complementarias que se pueden emplear en la extracción automática de terminología. Estas medidas estadísticas se pueden clasificar en dos dimensiones: dimensión lingüística y dimensión estadística.

Atendiendo a la dimensión estadística las medidas se pueden dividir según expresan *unithood* o *termhood:*

- *Unithood:* expresa la fuerza o la estabilidad de las colocaciones sintagmáticas (Pazienza 2005)
- *Termhood:* hace referencia al grado en que una palabra puede ser considerada un término en un cierto dominio (Alcina 2009)

La *unithood,* aunque captura un aspecto importante de los términos, no es una característica exclusiva de los términos ya que también se aplica a muchas otras unidades lingüísticas complejas (formadas por más de una palabra). La *termhood,* en cambio, sí es una característica propias de los términos, tanto sean multipalabra como si están formados por una sola palabra.

Atendiendo a la dimensión estadísticas las medidas se pueden clasificar en:

- Medidas del grado de asociación
- Medidas de la significancia de la asociación
- Medidas heurísticas

En la tabla siguiente (reproducida de Piazencia 2005) se pueden observar algunas medidas estadísticas clasificadas según las dimensiones lingüística y estadística:

	grado de asociación	significancia de la asociación	heurística
unithood	MI Dice Factor	z-score T-score X^2 Log Lokelihood Ratio	MI^2 MI^3
termhood			Freqüència C-Value Co-Occurrence

No entraremos en mucho detalle sobre estas medidas y consideraremos únicamente el cálculo de las medidas aplicadas a bigramas. La explicación, de nuevo, la obtenemos de Piazencia (2005) y quien quiera profundizar en detalles podrá consultar esta fuente.

Las medidas de asociación se utilizan para estimar la *unithood* y, como hemos dicho, se utilizan no sólo en terminología, sino de manera general para estimar las colocaciones entre dos palabras *(u* y *v),* basándose en las evidencias estadísticas sobre la ocurrencia de estas palabras en el corpus. Estas evidencias se expresan mediante una *tabla de contingencia* de frecuencias. A continuación presentamos estos cálculos para el caso de los bigramas. Llamaremos U y V en la primera y segunda palabra de la colocación. La coocurrencia de (u, v) se expresa con la frecuencia O11 y N es el número total de coocurrencias en el corpus (N = O11 + U12 + U21 + U22).

	V=v	V≠v
U=u	O11	O12
U≠u	O21	O22

También podemos definir las *frecuencias marginales* como:

R1 = O11 + U12
R2 = U21 + U22

C1 = O11 + U21
C2 = U12 + U22

En la tabla siguiente (Piazencia 2005) se pueden observar las fórmulas de cálculo de diversas medidas estadísticas utilizadas en extracción de terminología:

MEASURE	ADOPTED FORMULA
Frequency	$f = O_{11}/N$
Church Mutual Information	$MI = \log_2(O_{11}/E_{11})$
Mutual Information variants	$MI^2 = \log_2(O_{11}^2/E_{11})$ $MI^3 = \log_2(O_{11}^3/E_{11})$
Dice Factor	$DF = 2\frac{O_{11}}{R_1 + C_1}$
T-score	$TS = (O_{11} - E_{11})/\sqrt{O_{11}}$
Log Likelihood Ratio	$LLR = -2\log\frac{L(O_{11}, C_1, r)\cdot L(O_{12}, C_2, r)}{L(O_{11}, C_1, r_1)\cdot L(O_{12}, C_2, r_2)}$ where: $L(k,n,r) = r^k(1-r)^{n-k}$ $r = R_1/N$ $r_1 = O_{11}/C_1$ $r_2 = O_{12}/C_2$
C-value	$CV = (len - 1)\cdot\left(f - \frac{f(t)}{\lvert t\rvert}\right)$
Co-Occurrence	$CO = -\frac{\sum_N\sum_M O_{11i}}{\lvert N\rvert}$

Hay mucha bibliografía sobre cuál de estas medidas es la que funciona mejor en la tarea de extracción automática de terminología. Piacencia (2005) llega a la conclusión de que la mejor medida es la frecuencia (conclusión a la que también llegan Daile (1994) y Evert (2001), seguida por *T-score* y *C-value.* En cambio encuentra, a diferencia que en otros estudios, que *Log Likelihood Ratio* no funciona tan bien, aunque mejor que MI, MI ^{3} y *Dice Factor.*

Más recientemente (Vázquez, 2014) llega a la conclusión de que la frecuencia y *T-Score* (también llamada *T-student)* son las medidas estadísticas que funcionan mejor.

Como conclusión podemos decir que la frecuencia es una de las principales medidas estadísticas para aplicar la extracción automática de terminología. Esta medida es muy simple de calcular (simplemente mediante recuentos) y por tanto su interpretación también es muy sencilla.

3.8.Conclusiones

En este capítulo hemos ofrecido una visión general de los conceptos relacionados con la terminología con una visión práctica orientada a la traducción. Hemos aprendido a crear bases de datos terminológicas utilizando diferentes recursos y hemos presentado con detalle las principales técnicas para la extracción automática de terminiología. También se ha presentado el formato estándar para el intercambio de bases de datos terminológicas: el TBX *(Term* Base eXchange).

Bibliografía

Alcina, Amparo, Valero, Esperanza and Rambla, Elena (eds.) (2009) *Terminología y sociedad del conocimiento.* Peter Lang. ISBN 978-3-03911-593-8

Daille, B. (1994) *Approach mixte pour l'extraction de termilogie: statistique lexicale et filters linguistiques*. PhD Thesis, C2V, TALANA, Universitè Paris VII
Evert S., Krenn B. (2001) *Methods for the qualitative evaluation of lexical association measures*. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France. pp. 188-195

Localization Industry Standards Association (LISA) (2008) *Systems to manage terminology, knowledge, and content - TermBase eXchange (TBX)* (http://www.gala-global.org/oscarStandards/tbx/tbx_oscar.pdf)

Pazienza, Maria Teresa, Marco Pennacchiotti, and Fabio Massimo Zanzotto.(2005) *Terminology extraction: an analysis of linguistic and statistical approaches*. Knowledge Mining (2005): 255-279.

Sánchez-Gijón, Pilar (2004) *L'ús de corpus en la traducció especialitzada: compilació de corpus ad hoc i extracció de recursos terminològics*- IULA, Grup Tradumàtica, Departament de Traducció i d'Interpretació, Barcelona, 353p.

Santorini, B. (1990). *Part-of-speech tagging guidelines for the Penn treebank project*. 3rd Revision, 2nd printing, Feb. 1995. University of Pennsylvania.

Jörg Tiedemann (2009) *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*. In N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing (vol V), pages 237-248, John Benjamins, Amsterdam/Philadelphia

Vàzquez, Mercè (2014) Estratègies estadístiques aplicades a l'extracció automàtica de terminologia. Tesi Doctoral. Universitat Pompeu Fabra