3. Les bases de dades terminològiques

Índex

3.1. Introducció

En el capítol anterior hem vist un dels principals recursos per a la traducció: les memòries de traducció. En aquest capítol veure un altre recurs de gran importància: les bases de dades terminològiques. Començarem el capítol amb una possible definició de terme i presentarem els principals conceptes relacionats amb el treball terminològic. Veurem també la importància de la terminologia per a la traducció.

Bona part del capítol el dedicarem a la revisar les tècniques de creació de bases de dades terminològiques i li donarem una gran importància a les tècniques d'extracció automàtica de terminologia..

Presentarem també a fons el format TBX per a l'intercanvi de bases de dades terminològiques i veurem alguns recursos terminològics lliures i de lliure accés.

3.2. Terminologia i traducció

Abans de començar el treball terminològic cal establir una definició de terme que ens ajudi a decidir si una determinada unitat és o no és un terme. Començarem per la següent definició (Pazienza 2005):

> *Un terme és una representació superficial d'un concepte d'un domini específic*[1]

Entenem com a *representació superficial* la denominació, és a dir, la paraula o conjunt de paraules que s'utilitza per referir-se a aquest concepte. Per *domini específic* entenem el *camp d'especialitat*. Aquest aspecte és molt important, ja que la terminologia treballa amb unitats de coneixement especialitzat que es fan servir en camps d'especialitat concrets. No considerem, doncs, el llenguatge general, tot i que alguns termes referents a conceptes d'ús comú estan també presents en el llenguatge no especialitzat.

Els termes es poden veure com a unitats formades per un *concepte* i la seva *denominació*. Aquesta és la base de la Teoria General de la Terminologia de Wüster. Tal i com explica Sánchez-Gijón (2004), la Teoria General de la Terminologia estructura el coneixement en sistemes conceptuals i dota a cada concepte d'una denominació amb l'objectiu d'establir una comunicació inequívoca entre els experts.

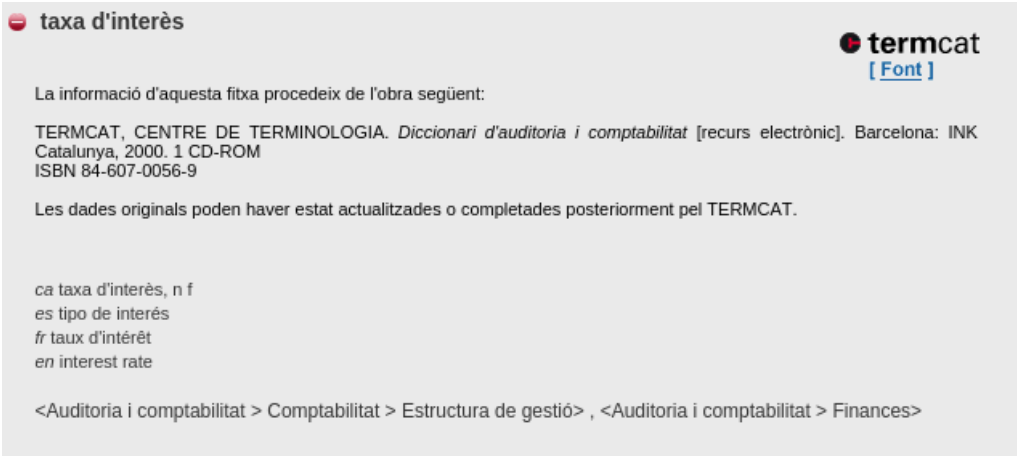
Hi ha moltes altres teories sobre la terminologia, però no entrarem en detalls en aquest capítol. Qui vulgui aprofundir en aquest tema pot llegir la tesi doctoral de Mercè Vàzquez (2014).

Sovint el traductor oblida les teories sobre la terminologia i les definicions de terme i inclou en els seus reculls terminològics unitats que no poden ser considerades pròpiament termes, però que precisen també d'una gran consistència en la seva traducció. En aquest capítol veurem com construir i gestionar bases de dades terminològiques i tindrem en ment també aquesta aproximació més pràctica.

1 *A surface representation of specific domain concept*

3.3. Bases de dades terminològiques, diccionaris generals i coneixement enciclopèdic

Las *bases de dades terminològiques* són reculls terminològics sistematitzats i generalment (avui dia, sempre) en un format informàtic i per tant utilitzables mitjançant un ordinador. Les bases de dades terminològiques, doncs, recullen termes, que són unitats pròpies del llenguatge especialitzat.



taxa d'interès

termcat

[Font]

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

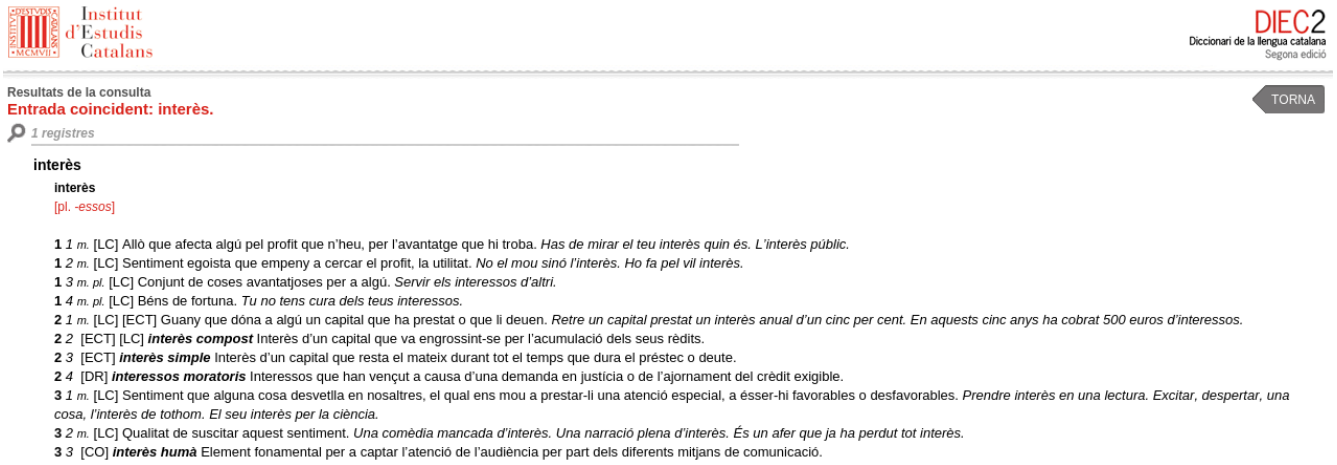
TERMCAT, CENTRE DE TERMINOLOGIA. *Diccionari d'auditoria i comptabilitat* [recurs electrònic]. Barcelona: INK Catalunya, 2000. 1 CD-ROM
ISBN 84-607-0056-9

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>

Els diccionaris generals inclouen paraules pròpies del llenguatge general. Alguns termes propis de camps especialitzats, que són d'ús comú, han entrat en el llenguatge general i per tant poden estar inclosos en els diccionaris.



Institut d'Estudis Catalans

DIEC2 Diccionari de la llengua catalana Segona edició

Resultats de la consulta
Entrada coincident: interès.
1 registres
TORNA

interès

interès
[pl. -essos]

1 *1 m.* [LC] Allò que afecta algú pel profit que n'heu, per l'avantatge que hi troba. *Has de mirar el teu interès quin és. L'interès públic.*
1 *2 m.* [LC] Sentiment egoista que empeny a cercar el profit, la utilitat. *No el mou sinó l'interès. Ho fa pel vil interès.*
1 *3 m. pl.* [LC] Conjunt de coses avantatjoses per a algú. *Servir els interessos d'altri.*
1 *4 m. pl.* [LC] Béns de fortuna. *Tu no tens cura dels teus interessos.*
2 *1 m.* [LC] [ECT] Guany que dóna a algú un capital que ha prestat o que li deuen. *Retre un capital prestat un interès anual d'un cinc per cent. En aquests cinc anys ha cobrat 500 euros d'interessos.*
2 *2* [ECT] [LC] ***interès compost*** Interès d'un capital que va engrossint-se per l'acumulació dels seus rèdits.
2 *3* [ECT] ***interès simple*** Interès d'un capital que resta el mateix durant tot el temps que dura el préstec o deute.
2 *4* [DR] ***interessos moratoris*** Interessos que han vençut a causa d'una demanda en justícia o de l'ajornament del crèdit exigible.
3 *1 m.* [LC] Sentiment que alguna cosa desvetlla en nosaltres, el qual ens mou a prestar-li una atenció especial, a ésser-hi favorables o desfavorables. *Prendre interès en una lectura. Excitar, despertar, una cosa, l'interès de tothom. El seu interès per la ciència.*
3 *2 m.* [LC] Qualitat de suscitar aquest sentiment. *Una comèdia mancada d'interès. Una narració plena d'interès. És un afer que ja ha perdut tot interès.*
3 *3* [CO] ***interès humà*** Element fonamental per a captar l'atenció de l'audiència per part dels diferents mitjans de comunicació.

El *coneixement enciclopèdic* és el coneixement del món acumulat per la humanitat. Una enciclopèdia és un compendi d'aquest coneixement. Les entrades de les enciclopèdies es refereixen a elements culturals de tipologia molt diversa però que, a diferència dels diccionaris, no constitueixen material estrictament lexicogràfic [font Vikipèdia]. La enciclopèdia, a més d'oferir una definició dels mots o termes, ofereixen informació més àmplia i aprofundida.

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]

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:

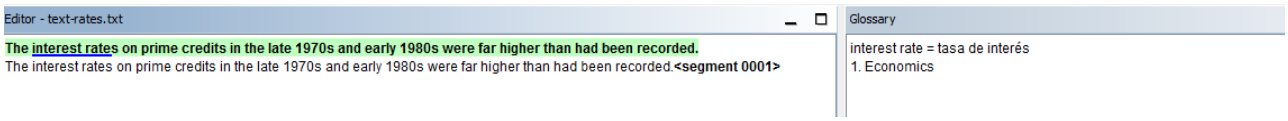
$$K_f = K_0(1 + i_N)$$

Donde:

Aquestes tres fonts: bases de dades terminològiques, diccionaris generals i enciclopèdies; són eines de consulta molt habituals dels traductors.

3.4. Cerca automàtica a bases de dades terminològiques

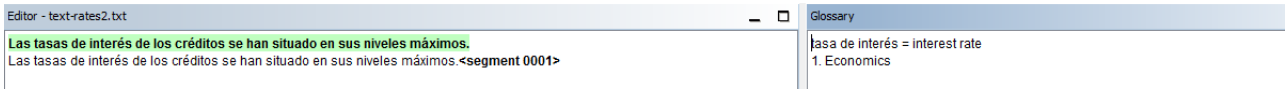
Les eines de traducció assistida permeten una consulta automàtica a bases de dades terminològiques. Si al segment que estem traduint apareix un terme de la base de dades terminològica, el programa ressaltarà aquest terme i ens mostrarà la informació relativa (com per exemple, la traducció) en una de les pantalles de l'eina. A continuació veiem un exemple:


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

Fixem-nos que l'eina ha de ser capaç de reconèixer el terme de la base de dades terminològica tot i que aquest aparegui en una altra forma. Si ens fixem, en la base de dades terminològica tenim recollida la forma base (singular): *interest rate* però en el text apareix en plural *interest rates*.

Les eines de traducció assistida han de ser capaces de trobar termes en altres formes sense tenir un coneixement massa profund ni específic de la llengua. Per a llengües amb una morfologia complexa els sistemes genèrics de reconeixement de termes poden fallar.

En OmegaT el reconeixement de termes es du a terme mitjançant *tokenitzadors* que son capaços de fer *stemming* (és a dir, eliminar els afixos morfològics de les paraules). D'aquesta manera, eliminant aquests afixos tant en el text del segment a traduir com en les entrades de la base de dades terminològica, el programa és capaç de trobar les entrades tot i que no coincideixin plenament les formes. Veiem ara el mateix exemple per al castellà:

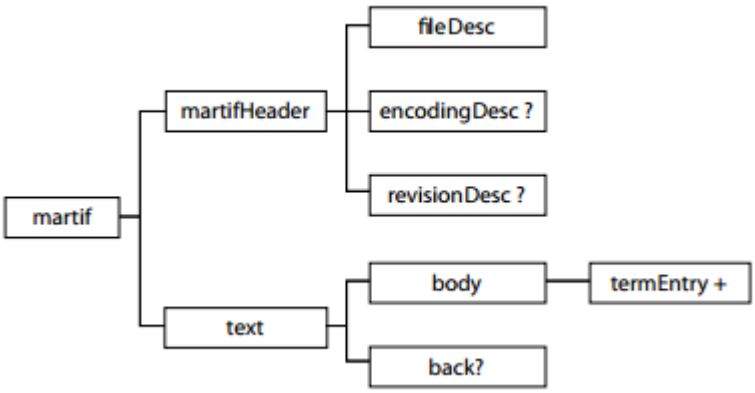

Editor - text-rates2.txt
Las tasas de interés de los créditos se han situado en sus niveles máximos.
Las tasas de interés de los créditos se han situado en sus niveles máximos.<segment 0001>
Glossary
tasa de interés = interest rate
1. Economics

Tot i que reconegui el terme en una forma diferent, si recuperem la traducció de la base de dades terminològica, per regla general els sistemes de traducció assistida no seran capaços de inserir en terme traduït en la forma correcta (en plural en aquests exemples). Per poder assolir això el sistema hauria de disposar de més informació lingüística.

3.5. Format d'intercanvi de bases de dades terminològiques: TBX

En el capítol anterior, dedicat a les memòries de traducció, vam veure el format d'intercanvi TMX (*Translation Memory eXchange*). En el cas de les bases de dades terminològiques hi ha un format similar, també basat en XML, anomenat TBX (*Term Base eXchange*). La idea és exactament la mateixa: tot i que cada gestor de bases de dades terminològiques i cada eina de traducció assistida pugui treballar amb un format intern diferent per representar les bases de dades terminològiques, podrem compartir les dades terminològiques amb altres eines fent servir aquest format d'intercanvi.

El TBX és un estàndard internacional (ISO 30042:2008) per a la representació de dades terminològiques, publicat conjuntament per la ISO (*International Standard Organisation*) i LISA (*Localization Industry Standard Association*). Es poden trobar les especificacions completes a LISA (2008). Aquí presentarem un petit resum de les característiques més destacades d'aquest format. Al següent esquema podem observar l'estructura d'un document MARTIF (*Machine-Readable Terminology Interchange Format*):

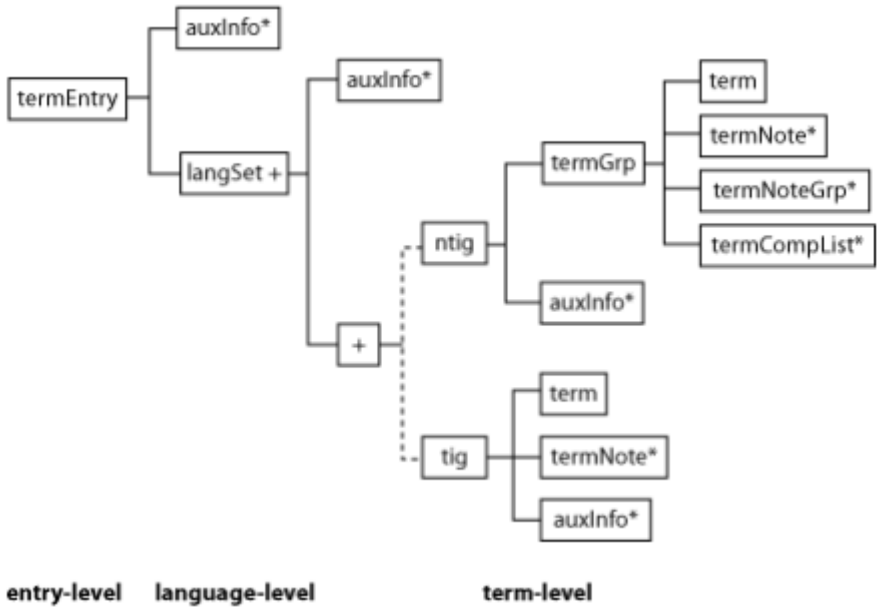


Com podem veure en l'esquema, el nivell més alt del document XML és l'element `<martif>`, que consisteix en un element `<martifHeader>` i un element `<text>`. Els noms d'aquests elements s'han agafat de la norma ISO 12200 i té les seves arrels en la *Text Encoding Initiative*. L'element `<text>` consisteix en les entrades terminològiques, que estan englobades en un element `<body>` element i informació complementària. En TBX, la informació complementària es troba en l'element `<back>`.

La informació sobre la codificació de caràcters s'ha d'incloure en la capçalera només quan la codificació sigui diferent d'Unicode.

Components d'una entrada terminològica

Cada entrada terminològica dins de l'element `<body>` s'anomena `<termEntry>` i segueix l'estructura del metamodel TMF. En la següent figura podem observar els nivells d'una entrada terminològica:



El requadre auxInfo correspon a informació que es pot associar a qualsevol dels tres nivells: el nivell d'Entrada Terminològica (<termEntry>, és a dir, el nivell de concepte), el nivell de Llengua (<langSet>) i el nivell de Terme (és a dir, la denominació, <ntig> o la seva versió simplificada <tig>). Els elements **<termNote>** i **<termNoteGrp>** només poden aparèixer en el nivell de Terme o per sota.

Exemple d'arxiu TBX

A continuació podem observar un exemple d'arxiu 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>
```

Dialectes del TBX

El TBX és un format d'intercanvi molt ben dissenyat i que s'adapta perfectament a les tasques terminològiques més complexes. Les especificacions completes del TBX tenen més de 90 planes i implementar aplicacions que siguin totalment compatibles amb l'estàndard sencer resulta, sinó complicat, sí molt laboriós.

Moltes de les tasques relacionades amb la terminologia, i especialment les tasques pràctiques relacionades amb la feina del traductor, no requereixen un format d'intercanvi tan complet. Per aquest motiu s'han creat *dialectes* del propi TBX amb l'objectiu de simplificar-lo. En aquesta secció presentarem dos d'aquests dialectes: el TBX-Basic i el TBX-Min. Un aspecte molt important a tenir en compte d'aquests dialectes és que són en sí TBX vàlids. Per tant, una aplicació capaç de llegir i processar els TBX complets, serà també capaç de llegir aquests dialectes simplificats.

TBX-Basic

El TBX-Basic està dissenyat per proporcionar les categories de dades que es fan servir de manera habitual en les tasques de traducció i localització. Les principals diferències entre el TBX complet i el TBX-Basic són les següents:

- El TBX disposa dels element s `<tig>` i `<ntig>` per als grups d'informació sobre el terme. En canvi, el TBX-Basic només disposa de l'element `<tig>`.
- El TBX-Basic no permet documentar els components d'un terme (és a dir, les parts individuals dels termes). Per tant els següents elements no estan presents en TBX-Basic: `<termComp>`, `<termCompList>`, `<termCompGrp>`, i `<termGrp>`.
- El TBX-Basic no admet els següents elements d'agrupació i dels seus elements fills: `<adminGrp>`,`<termNoteGrp>`, `<itemSet>` i `<itemGrp>`. En TBX-Basic només s'admeten els següents elements d'agrupació: `<descripGrp>` i `<transacGrp>`.
- En TBX-Basic, l'element `<descripGrp>` es fa servir només per associar una font a una definició o a un context. Per tant, els següents elements fills no s'admeten en TBX-Basic: `<descripNote>`, `<admin>`, `<adminGrp>`, `<note>`,`<ref>`, `<xref>`.
- En TBX-Basic, els valors dels atributs "DCSName" i "XCSContent" no són compatibles amb l'etiqueta de paràgraf en el element `<encodingDesc>`.

TBX-Min

El TBX-Min està dissenyat per facilitar l'emmagatzematge de glossaris monolingües i bilingües i per permetre una conversió fàcil amb TBX-Basic i amb altres formats per a l'emmagatzematge de glossaris, com per exemple l'UTX (*Universal Terminology eXchange* - http://www.aamt.info/english/utx/). Una de les seves utilitats més clares és la utilització d'aquest format per enviar un glossari a un traductor que participa en un projecte i que no necessita de tota la informació que potencialment està inclosa en el glossari complet en format TBX-Basic (o TBX complet). El traductor pot rebre aquest TBX-Min, treballar amb ell introduint canvis o nova informació. Un cop acabada la feina pot enviar aquest TBX-Min i els canvis es poden incloure en la TBX-Basic (o complet).

La capçalera d'un TBX-Min conté la següent informació:

- Un ID únic per identificar el document
- El nom del creador
- Una descripció de la base de dades terminològica
- La direccionalitat: monodireccional o bidireccional
- La llengua de partida i d'arribada (només es poden fer servir dues llengües)
- La llicència de la base de dades terminològica
- La data en la que es va crear la base de dades terminològica (en formato ISO 8601)

El cos (<body>) conté les entrades aniuades usuals <langGroup> i <termEntry>.

L'element <termGroup> pot contenir:

- el text del terme (denominació). Aquest és l'únic element obligatori
- una nota
- la categoria gramatical
- el nom del client
- l'estat del terme

El format UTX (*Universal Terminology eXchange*)

El format UTF és un format estàndard per a glossaris molt simple i que pot ser fàcilment compartit i reutilitzat entre diferents eines. L'UTX és un format molt senzill, ja que és simplement text tabulat, i es pot crear i manipular fàcilment amb qualsevol editor de textos o full de càlcul. Els termes poden contenir un camp de Term Status (que pot ser *approved* o *forbidden*).

A continuació podem observar un exemple de glossari en aquest format:

```
#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 ▪ □□□□□▪ □□□▪
```



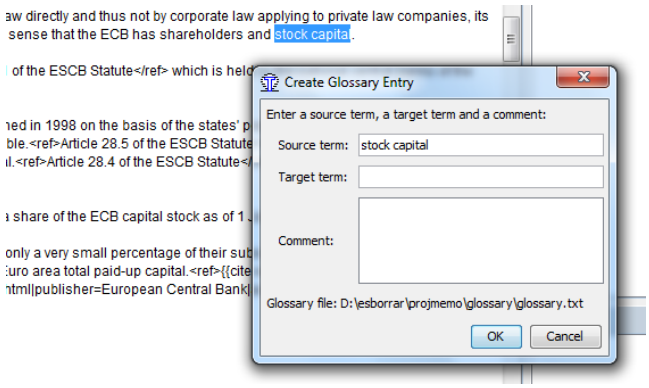
```
▪ □□▪ □▪ □□□▪ □□□▪ □ □▪ □ □▪ □□□▪ □□□▪ □ □ □ □□▪ □□▪ □□□▪ □□□▪ □ □ □□□▪ □□□□▪ □ □ □ □ □ □ □ □ □ □ □ □□▪ □
#src  tgt   src:pos     comment
1/2 T vector      1/2T ベクトル noun
1/2FF 1/2 拡張分画  noun
1/3 ER mean 駆出早期 1/3 ▪ ▪ □□□□□▪ □  noun
1/3EF 1/3 駆出分画  noun
1/3ER mean  駆出早期 1/3 ▪ ▪ □□□□□▪ □  noun
1/3FF 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
…
```

Aquest format només permet representar glossaris bilingües. A aquesta versió se la coneix com a UTX simple. Està prevista l'aparició d'un UTX-XML més complex que permeti la representació de glossaris multilingües.

3.6. Creació de bases de dades terminològiques

3.6.a. A mesura que es tradueix

Els traductors, a mesura que van avançant en una traducció van consultant diverses fonts per resoldre els seus dubtes terminològics. En aquest moment és important emmagatzemar el resultat de les consultes en algun format que sigui consultable de manera fàcil i ràpida. La millor opció, evidentment, és emmagatzemar les consultes en la mateixa base de dades terminològica que fa servir la pròpia eina de traducció assistida. Totes les eines de traducció assistida disposen d'alguna funcionalitat que permet introduir nous termes en la bases de dades terminològica del projecte. En la següent imatge podem observar la pantalla de creació d'entrades terminològiques d'OmegaT. L'eina permet seleccionar un terme de l'original i quan s'obre la pantalla de creació d'entrades el terme seleccionat apareix automàticament al camp *Source term*. El traductor podrà completar la resta de camps i el terme s'emmagatzemarà automàticament a la base de dades terminològica. A partir d'aquest moment la informació sobre el terme apareixerà automàtica en aquest projecte de traducció. També podrem exportar aquestes entrades i importar-les a altres bases de dades.

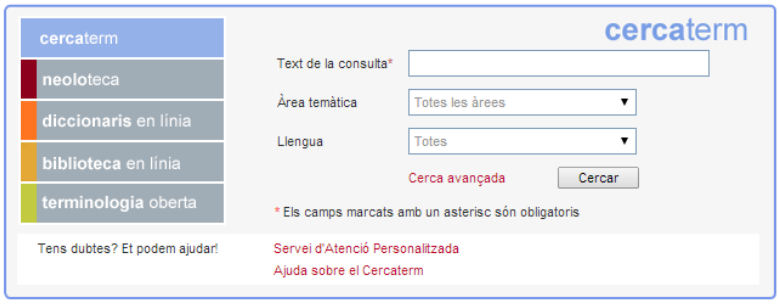


3.6.b. A partir de recursos terminològics a Internet

A Internet podem trobar algunes planes que distribueixen recursos terminològics o bé que permeten fer cerques terminològiques. En aquest apartat comentarem alguns d'aquests recursos.

TermCat

Començarem parlant del TermCat (http://www.termcat.cat/), que és el centre de terminologia de la llengua catalana, creat el 1985 per la Generalitat de Catalunya i l'Institut d'Estudis Catalans. Tot i que està creat per al català la majoria d'entrades terminològiques que recullen estan també en anglès i castellà, i algunes també en francès o alemany. El TermCat per una banda ofereix el CercaTerm, que permet fer consultes terminològiques a partir del terme a cercar, la llengua i la possibilitat d'indicar l'àrea temàtica.



A més de la interfície gràfica el TermCat ofereix un servei de consultes terminològiques, per als casos que no trobem el que cerquem en la seva interfície.

El TermCat allibera les seves bases de dades terminològiques i les publica en la secció Terminologia Oberta (http://www.termcat.cat/ca/TerminologiaOberta/). Aquestes bases de dades terminològiques estan en un format XML no estàndard però que es pot convertir a TBX o a text tabulat mitjançant l'eina TO2TBX (http://lpg.uoc.edu/TO2TBX/).

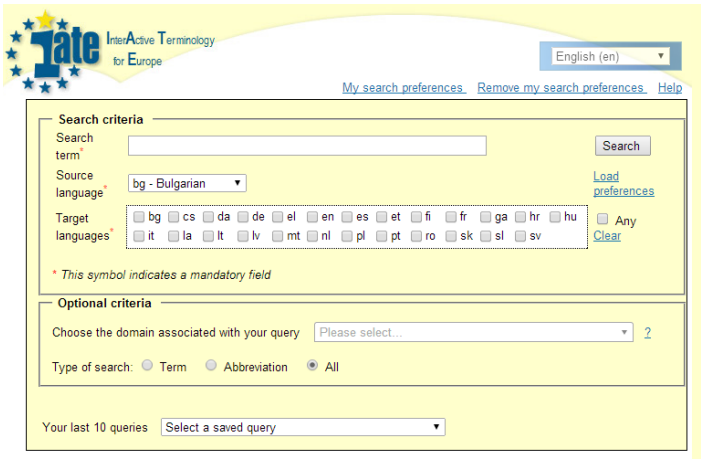
A més el TermCat ha desenvolupat un programa de gestió de la terminologia, el GesTerm, que es distribueix sota una llicència lliure (http://www.termcat.cat/ca/GesTerm/).

IATE

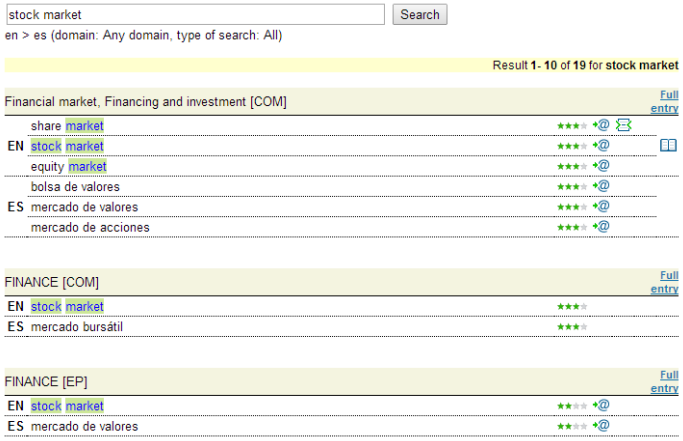
IATE (= "Inter-Active Terminology for Europe") (http://iate.europa.eu/) és la base de dades terminològica inter-institucional de la Unió Europea. Actualment conté aproximadament 1.4 milions d'entrades multilingües.S'han importat les següents bases de dades terminològiques:

- Eurodicautom (Commission)
- TIS (Council)
- Euterpe (EP)
- Euroterms (Translation Centre)
- CDCTERM (Court of Auditors)

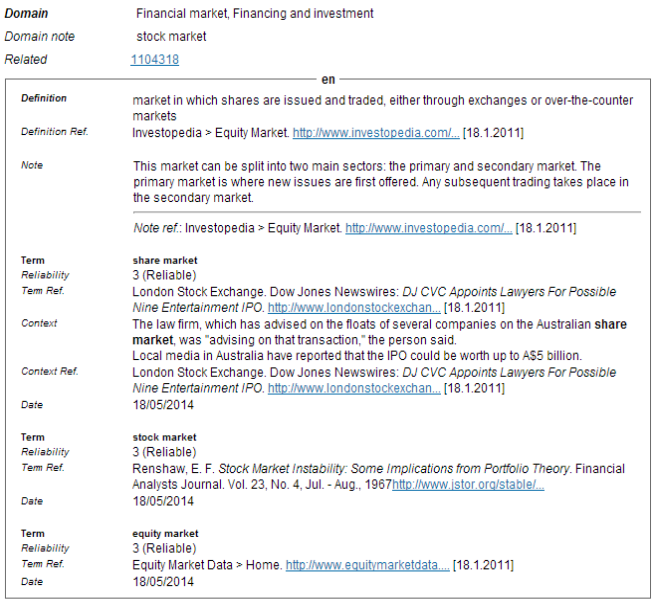
IATE permet fer consultes mitjançant una interfície de cerca:



I obtenir una sèrie de resultats:



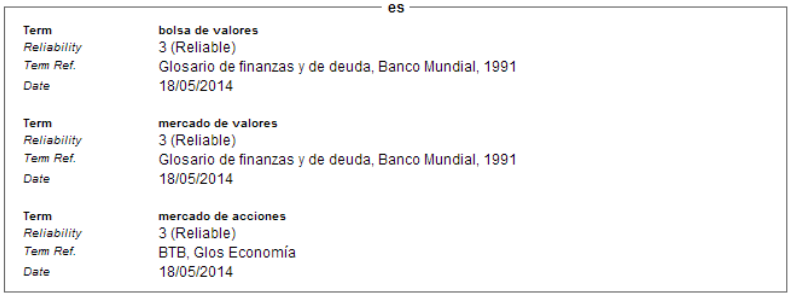
De cada una de les entrades podem obtenir l'entrada terminològica completa:



Domain	Financial market, Financing and investment
Domain note	stock market
Related	1104318

en

Definition	market in which shares are issued and traded, either through exchanges or over-the-counter markets
Definition Ref.	Investopedia > Equity Market. http://www.investopedia.com/... [18.1.2011]
Note	This market can be split into two main sectors: the primary and secondary market. The primary market is where new issues are first offered. Any subsequent trading takes place in the secondary market.
	Note ref.: Investopedia > Equity Market. http://www.investopedia.com/... [18.1.2011]
Term	**share market**
Reliability	3 (Reliable)
Term Ref.	London Stock Exchange. Dow Jones Newswires: *DJ CVC Appoints Lawyers For Possible Nine Entertainment IPO.* http://www.londonstockexchan... [18.1.2011]
Context	The law firm, which has advised on the floats of several companies on the Australian **share market**, was "advising on that transaction," the person said. Local media in Australia have reported that the IPO could be worth up to A$5 billion.
Context Ref.	London Stock Exchange. Dow Jones Newswires: *DJ CVC Appoints Lawyers For Possible Nine Entertainment IPO.* http://www.londonstockexchan... [18.1.2011]
Date	18/05/2014
Term	**stock market**
Reliability	3 (Reliable)
Term Ref.	Renshaw, E. F. *Stock Market Instability: Some Implications from Portfolio Theory.* Financial Analysts Journal. Vol. 23, No. 4, Jul. - Aug., 1967http://www.jstor.org/stable/...
Date	18/05/2014
Term	**equity market**
Reliability	3 (Reliable)
Term Ref.	Equity Market Data > Home. http://www.equitymarketdata.... [18.1.2011]
Date	18/05/2014



es

Term	**bolsa de valores**
Reliability	3 (Reliable)
Term Ref.	Glosario de finanzas y de deuda, Banco Mundial, 1991
Date	18/05/2014
Term	**mercado de valores**
Reliability	3 (Reliable)
Term Ref.	Glosario de finanzas y de deuda, Banco Mundial, 1991
Date	18/05/2014
Term	**mercado de acciones**
Reliability	3 (Reliable)
Term Ref.	BTB, Glos Economía
Date	18/05/2014

Recentment, la base de dades IATE s'ha alliberat i s'ha publicat com a un arxiu TBX de grans dimensions que conté les seves entrades. A continuació podem observar una d'aquestes entrades:

```
<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>
```


Com podem veure, es tracta d'una base de dades multilingüe de les llengües oficials de la Unió Europea. No totes les entrades estan en totes les llengües. Les entrades contenen també informació de camp temàtic (subjectField) que en aquesta entrada és el 1011, que correspón a *European Union Law.*

En la següent taula podem observar tots els codis de camp temàtic de l'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
```

Es pot accedir a una llista molt més detallada dels codis de camp temàtic a: http://iate.europa.eu/tbx/IATE%20domain%20codes.csv

Cada entrada té també una informació de fiabilitat (*reliabilityCode*) que pot tenir tres nivells:

- 1: Fiabilitat no verificada
- 2: Fiabilitat mínima

- 3: Fiable
- 4: Molt fiable

Es pot trobar una descripció completa de la informació sobre els temes de l'IATE a: http://iate.europa.eu/tbx/IATE%20Data%20Fields%20Explaind.htm

El fitxer TBX que es pot descarregar és un fitxer molt gran i que és difícil de tractar amb les eines estàndard. En http://lpg.uoc.edu/IATE es pot accedir a fitxers de text tabulat que contenen els termes classificats per parells de llengua i per especialitats. També es pot descarregar una senzilla eina (IATE2tabtxt.py) que permet fer la conversió del TBX en formats de text tabulats per parells de llengües. En la mateixa web està disponible una altra eina (IATE2TBX.py) que crea arxius TBX que contenen només la informació dels parells de llengua i les especialitats desitjats. D'aquesta manera es poden crear fitxers molt més fàcils de manipular amb les eines estàndard.

Eurovoc

Eurovoc (http://eurovoc.europa.eu/) és un tesaurus multilingüe i multidisciplinari que inclou la terminologia dels àmbits d'activitat de la Unió Europea, amb especial èmfasi en les tasques parlamentàries. Eurovoc està disponible en 23 llengües oficials de la Unió Europea (alemany, búlgar, txec, croat, danès, eslovac, eslovè, espanyol, estonià, finès, francès, grec, hongarès, anglès, italià, letó, lituà, maltès, neerlandès, polonès, portuguès, romanès i suec), a més de la llengua d'un tercer país (serbi). Eurovoc també està disponible en català[2] i eusquera[3].

Eurovoc es pot descarregar en diversos formats que poden ser incorporats a bases de dades terminològiques.

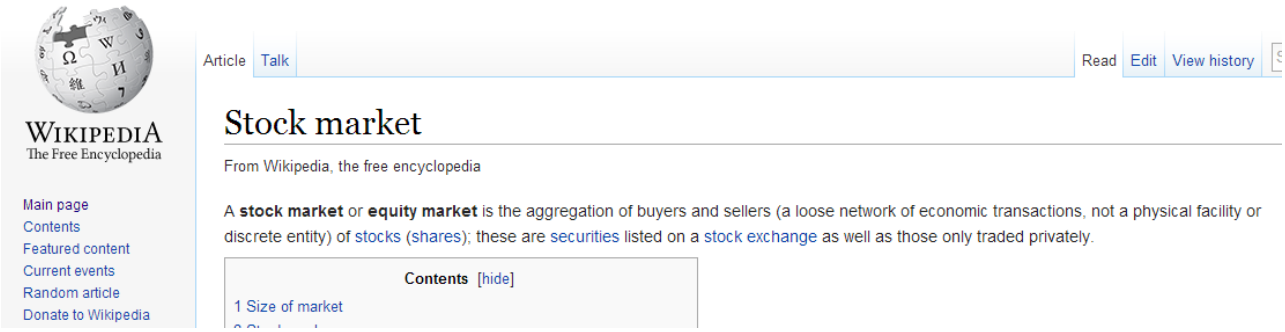
Unterm

Unterm (http://unterm.un.org/) és la base de dades terminològica de les Nacions Unides, que conté termes tècnics i nomenclatures en les sis llengües oficials d'aquesta institució: àrab, xinès, anglès, rus i espanyol.

3.6.c. A partir de la Vikipèdia

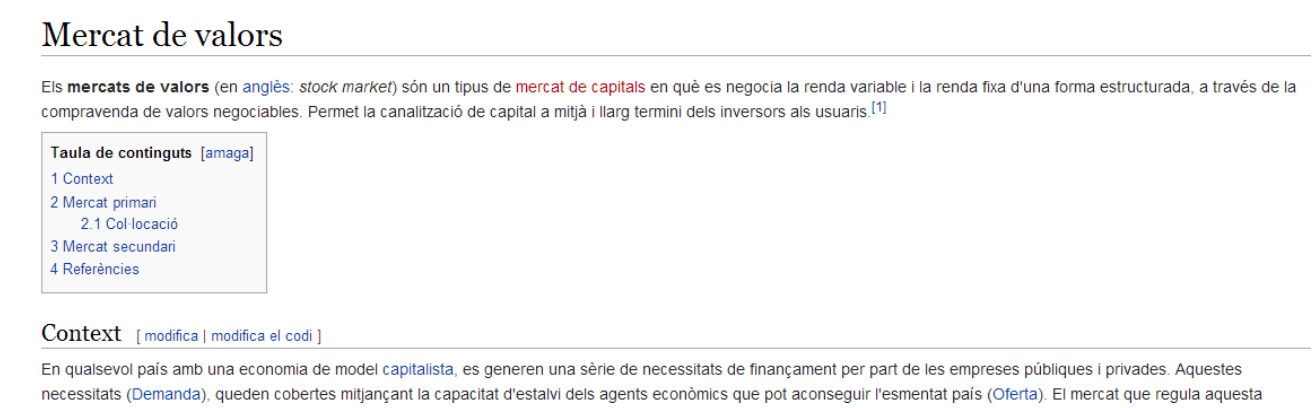
La Vikipèdia (http://www.wikipedia.org/) és una enciclopèdia multilingüe lliure que s'ha construït (i es continua construint) de manera col·laborativa. Com que es tracta d'un recurs multilingüe pot resultar també una bona font de consulta per a un traductor. Imaginem-nos que ens apareix el terme *stock market* i que no sabem com traduir-lo ni l'hem trobat en les nostres bases de dades terminològiques. Podem veure si en la Wikipedia anglesa existeix una entrada per a aquest terme:

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



Cada article està relacionat amb els articles equivalents en altres llengües. Si ens fixem. a la part esquerra apareixen els enllaços interlingüístics:

Com podem observar aquesta entrada està també disponible en català i podem anar directament a la plana catalana fent clic a l'enllaç:



Mercat de valors

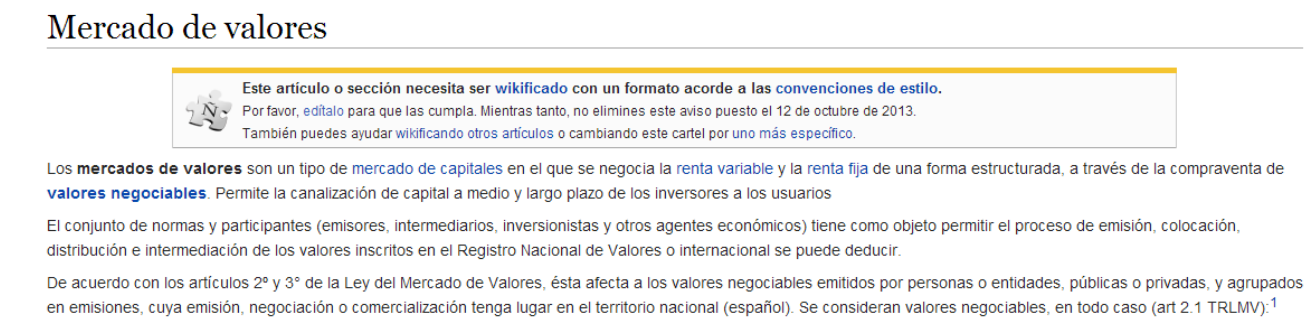
Els **mercats de valors** (en anglès: *stock market*) són un tipus de mercat de capitals en què es negocia la renda variable i la renda fixa d'una forma estructurada, a través de la compravenda de valors negociables. Permet la canalització de capital a mitjà i llarg termini dels inversors als usuaris.[1]

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

[per a la versió castellana i anglesa]



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

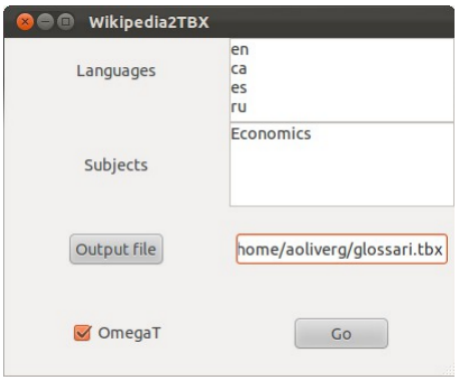
El conjunto de normas y participantes (emisores, intermediarios, inversionistas y otros agentes económicos) tiene como objeto permitir el proceso de emisión, colocación, distribución e intermediación de los valores inscritos en el Registro Nacional de Valores o internacional se puede deducir.

De acuerdo con los artículos 2º y 3° 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]

I d'aquesta manera determinar que la traducció de *stock market* en català pot ser *mercat de valors*.

Aquesta tasca pot resultar una mica pesada, però hi ha disponible una senzilla aplicació, Wikipedia2TBX (http://lpg.uoc.edu/Wikipedia2TBX/) que permet la creació de bases de dades terminològiques a partir de la Vikipèdia d'una manera automàtica.

Aquest programa disposa d'una interfície molt senzilla:



A partir de les llengües d'interès (*Languages*) i d'una o més àrees d'especialitat (*Subjects*) permet crear un glossari terminològic:

```
<?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>
```

i en format tabulat per 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
```


Els codis de llengua que es fan servir són els corresponents als codis ISO de dues lletres (ISO 639-1). Aquests codis es poden consultar al següent enllaç: http://ca.wikipedia.org/wiki/ISO_639-1

Les àrees d'especialitat són les pròpies de la Vikipèdia i s'han d'expressar en anglès. Recordem, però, que les àrees d'especialitat són lliures, és a dir, qualsevol usuari en pot crear. Per poder conèixer quines àrees d'especialitat hi ha és útil consultar el següent enllaç: http://en.wikipedia.org/wiki/List_of_academic_disciplines

3.6.d. Extracció automàtica de terminologia

A aquesta tècnica de creació de bases de dades terminològiques li dedicarem un apartat sencer. Les tècniques d'extracció automàtica (o potser millor dit semiautomàtica) de terminologia intenten detectar les unitats terminològiques presents en un text o conjunt de textos, sense un coneixement previ d'aquestes unitats. Com veurem al següent apartat, tot i requerir una revisió manual exhaustiva, les tècniques d'extracció automàtica de terminologia són molt productives i permeten la creació ràpida de bases de dades terminològiques.

3.7. Extracció automàtica de terminologia

3.7.1. Definició

L'extracció automàtica de terminologia és un procés pel qual es detecten una sèrie de candidats a unitats terminològiques a partir d'un text o un conjunt de textos. Tot i que acostuma a rebre el nom d'extracció automàtica de terminologia el procés no és totalment automàtic, ja que requereix d'una revisió manual. Per aquest motiu molts autors prefereixen parlar d'extracció *semiautomàtica* de terminologia. En aquest llibre mantindrem la denominació d'automàtica ja que el procés pretén ser automàtic, però ara per ara, els resultats que s'obtenen no són prou precisos com per poder acceptar directament el resultat de l'extracció.

Cal no confondre el procés d'extracció automàtica de terminologia amb el procés de detecció automàtica de termes. En la detecció automàtica de termes, els termes són coneguts *a priori*, i el sistema intenta determinar quins termes d'una base de dades estan presents en el text que estem traduint. En aquest cas, la dificultat principal és la detecció de formes flexionades dels termes.

L'extracció automàtica de terminologia permet tractar d'una manera ràpida una gran quantitat de textos, i tot i que el procés de revisió manual és laboriós, es poden construir bases de dades terminològiques d'una manera molt ràpida.

3.7.2. Classificació de mètodes per a l'extracció automàtica de terminologia

A Pazienza (2005) es pot trobar una explicació detallada dels mètodes per a l'extracció automàtica de terminologia. Els mètodes es poden classificar en dos grans grups:

- *Mètodes estadístics*: l'extracció de termes es fa a partir de les seves propietats estadístiques (la característica més habitual és simplement la freqüència d'aparició)
- *Mètodes lingüístics*: l'extracció de termes es fa a partir de les seves propietats lingüístiques (habitualment les seves propietats morfosintàctiques)

Sovint també es parla de *mètodes híbrids*, que combinen mètodes estadístics i mètodes lingüístics. De fet, estrictament tots els mètodes per a l'extracció automàtica de terminologia són híbrids, ja que com veurem els mètodes estadístics fan ús d'una llista de paraules buides (*stop-words*) i això és en certa manera un coneixement lingüístics; i els mètodes lingüístics també fan servir la freqüència d'aparició per a endreçar els candidats, i aquesta és una propietat estadística.

L'extracció de terminologia pot ser *monolingüe*, en la que s'extreuen termes en una sola llengua; o *bilingüe* (i per extensió *multilingüe*) en la que s'extreuen termes en una llengua i els seus equivalents de traducció en una altra (o en més d'una) llengua. En l'extracció bilingüe habitualment es fan servir corpus paral·lels (o memòries de traducció) per cercar de manera automàtica els equivalents de traducció.

3.7.3. Mètodes estadístics per a l'extracció automàtica de terminologia

Els *mètodes estadístics* per a l'extracció automàtica de terminologia són aquells que fan servir principalment informació estadística per detectar candidats a termes.

Els mètodes estadístics es basen en el càlcul d'*n-grames*. Un *n-grama* és una combinació d'*n* elements. En el cas de l'extracció automàtica de terminologia aquests elements són paraules (o *tokens*). Observem el següent exemple:

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

En aquesta oració tenim (com a mínim) un terme: *monetary policy*. Si calculem els *bigrames* (*n-grames* d'ordre 2), és a dir, les combinacions de dues paraules (o millor dit *tokens*, ja que algunes combinacions inclouen elements que no són paraules, com per exemple signes de puntuació), obtenim:

bigrams:

```
[('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', 'medium'), ('medium', 'term'), ('term', '.')]
```

Aquest càlcul es pot fer d'una manera molt senzilla amb el llenguatge de programació Python i la llibreria NLTK (*Natural Language Toolki)*. Veiem a continuació el codi:

```
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 aquest cas l'oració la tenim ja *tokenitzada* dins del codi. En altres exemples veurem com podem fer aquesta *tokenització*.

És evident que si el terme present en l'oració està format per dues paraules i calculem els *bigrams* d'aquesta oració, el nostre terme estarà present en la llista de bigrams. El que passa és que també obtenim moltes altres combinacions que no són terminològiques, coses de l'estil: *is the*, *the only*, etc. Com que en aquest cas només hem fet l'extracció a partir d'una única oració, la informació estadística com per exemple la freqüència d'aparició no és rellevant (ja que tots els bigrams apareixen una única vegada).

Si ara fem el mateix però calculem a més de *bigrams* també *trigrams* (*n-grames* d'ordre 3) d'un corpus més gran (en l'exemple el fem d'un subconjunt de 10.000 oracions del corpus ECB (European Central Bank) de l'anglès[4] (Tiedemann 2009), i endrecem els candidats per freqüència, obtindrem els següents resultats (es mostren els 25 més freqüents). Veiem en primer lloc el codi en Python:

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


```
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)
#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
```

I ara els resultats que s'obtenen:

```
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 on
1742 the ECB
1680 ( CON
1610 & apos
1570 apos ;
```

```
1402 by the
1398 ; s
1210 the European
1204 for the
1152 and the
1138 , en
1074 ECB /
```

Com podem observar, tot i que hem calculat bigrams i trigrams, entre els 25 primers "candidats" no obtenim cap trigram, ja que les unitats més curtes són les més freqüents. Tot el que obtenim són combinacions de paraules funcionals, puntuacions i sovint elements que provenen d'errors de conversió de formats. Per tant, el que hem obtingut no s'assembla ni de bon tros a una extracció de terminologia.

Filtratge per paraules buides (*stop-words*)

Per poder obtenir aquesta llista en una llista de candidats a termes caldrà filtrar-la amb una llista de *paraules buides* (o *stop-words*). Aquestes llistes contenen paraules funcionals i d'altres que no acostumen a aparèixer ni en primera ni en darrera posició d'un terme.

Una llista de paraules buides de l'anglès contindria paraules com ara: *a, a's, able, about, above, according, accordingly, across, actually, after, afterwards, again, against, ain't, all, allow, allows, almost, alone, along, already*...

El filtratge consistirà en eliminar de la llista de candidats aquells que comencen o acaben amb una paraula de la llista de paraules buides. A continuació podem observar el codi i el resultat per al cas de la frase simple del primer exemple:

```
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
```

I ara obtenim els següents resultats:

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

on ja si que apareix el nostre terme junt amb altres unitats no terminològiques, però on el nombre de candidats s'ha reduït notablement.

Si apliquem el mateix filtratge al corpus de 10.000 oracions de l'ECB corpus obtenim:

```
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
```

I els següents candidats:

```
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 encara molt de ser perfecta però que ha millorat molt: apareixen ja alguns termes, com *monetary policy* i *payment order* (que apareix tant en plural com en singular), o *central banks* (només en plural) i *euro area*. Apareixen encara algunes combinacions de símbols no desitjades (que es poden eliminar afegint aquest símbols com a paraules de la llista de paraules buides). També apareixen termes partits, com seria *European Central Bank* que apareix com a *Central Bank i European Central*. Una mica més endavant veurem estratègies per intentar arreglar tots aquests resultats incorrectes.

TBXTools

La resta d'exemples d'aquesta secció els farem fent referència a l'eina TBXTools, que és una classe escrita en Python que implementa molts mètodes d'extracció automàtica de terminologia i altres utilitats relacionades amb la gestió de la terminologia. Aquesta eina es pot descarregar de http://lpg.uoc.edu/TBXTools. Aquesta classe facilita enormement la implementació de programes per a l'extracció automàtica de terminologia.. A continuació veiem el codi corresponent a l'exemple anterior però escrit fent servir aquesta classe:

```
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)
```

Paraules buides internes

Si analitzem a fons els candidats a termes obtinguts amb el programa anterior veurem alguns exemples com:

```
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 corresponen a candidats erronis no filtrats per paraules buides, ja que habitualment es descarten els candidats que comencen o acaben per paraules de la llista de paraules buides. En aquests exemples veiem que són trigrams en el que la paraula del mig és una conjunció. TBXTools permet definir una llista de paraules buida interna, que farà servir per eliminar aquells candidats que tinguin en el seu interior (qualsevol posició excepte la primera i la darrera) un paraula d'aquesta llista. Una possible llista per a l'anglès estaria composada per les següents paraules:

```
and
but
or
nor
so
```

A continuació podem veure un exemple de codi que fa sevir aquest filtratge:

```
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)
```

Normalització de majúscules i minúscules

Si mirem a fons el resultat d'una extracció purament estadística sovint ens trobem amb resultats com:

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

és a dir, amb el mateix terme escrit amb diferents versions segons les majúscules i minúscules. En realitat, desitjaríem que el nostre sistema ajuntés tots aquests candidats sota un d'únic i que sumés les freqüències:

```
222 monetary policy
```

TBXTools (i moltes altres eines d'extracció automàtica de terminologia) és capaç de dur a terme aquesta normalització. Per fer-ho, simplement verifica si hi ha diverses realitzacions d'un mateix terme que es diferenciïn només pel fet d'estar escrites amb majúscules o minúscules. A continuació veiem un exemple de codi que implementa aquesta funció:

```
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)
```

Normalització morfològica

Entre els candidats a termes ens trobarem variants morfològiques dels diferents termes. Per exemple:

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

quan voldríem obtenir la forma base i les freqüències sumades:

```
335 payment order
```

o bé:

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

quan en realitat voldríem obtenir:

```
5 economic policy
```

Donat que les tècniques estadístiques no fan servir gaire informació lingüística, el sistema no és capaç de saber que aquestes realitzacions són en realitat el mateix terme. TBXTools permet definir una sèrie de regles morfològiques senzilles que permeten al sistema ajuntar diferents formes del mateix terme. Per funcionar necessita una sèrie de regles de l'estil:

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

La "L" de la regla significa que és l'últim element (*Last*) de l'n-grama. El primer camp de la regla (nul per a la primera i darrera regla i "y" per a la tercera és la terminació de lema; i el segon element ("s", "es" i "ies") és la terminació de forma.

I el codi que crida a aquesta funció:


```
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)
```

Aquestes regles poden ser útils per llengües amb morfologia simple, com per exemple l'anglès. Per altres llengües, la quantitat de regles pot ser realment important i la seva definició molt complexa, de manera que és més recomanable fer servir les tècniques lingüístiques que explicarem més endavant.

Detecció de candidats aniuats

Quan obtenim candidats a terme d'un ordre *n* sovint passa que en realitat són part d'un terme d'algun ordre superior (*n+1* o fins i tot d'un ordre més alt). En l'exemple que estem treballant, el candidat bigram

```
419 European Central
```

és en realitat un fragment del candidat trigram

```
417 European Central Bank
```

TBXTools implementa la detecció de termes aniuats, afegint al codi del programa la següent línia:

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

llavors l'eina verifica els possibles aniuaments. El percentatge (que en l'exemple està fixat a 10) significa quina és la diferència màxima de freqüències entre el candidat d'ordre *n* i el superior (en aquest cas el 10%). El sistema permet detectar aniuaments com els següents:

```
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
```

Detecció de termes monoparaula (1-grames o unigrames)

Si ens hem fixat, haurem vist que les tècniques estadístiques es basen en el càlcul de combinacions de paraules (de dues o més paraules). Això no permet detectar termes monoparaula (unigrames, és a dir, termes formats per una única paraula). Si calculem els 1-grames obtindrem totes les paraules del text i si filtrem aquesta llista amb paraules buides simplement eliminarem les paraules buides de la llista de candidats. En l'exemple que estem treballant obtindríem una llista com la següent:

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

Com veiem doncs, l'aproximació estadística que hem descrit no serveix per a extreure termes monoparaula d'un corpus. Es poden seguir dues estratègies:

- Per obtenir els termes monoparaula, el revisor humà aïllarà els candidats que consideri interessants mentre revisa candidats d'ordre superior. Per exemple, si a la seva llista de candidats troba: *payment order* pot decidir que *payment* també és un terme rellevant de la disciplina.
- Per a certes especialitats, per exemple medicina, hi ha una gran quantitat de cultismes formats per sufixes específics (per exemple *-itis* en medicina). El sistema pot detectar paraules acabades en aquests sufixos i incloure-les en la llista de candidats.

Detecció d'equivalents de traducció a corpus paral·lels

Si disposem d'un corpus paral·lel, a més d'extreure candidats de traducció en una de les llengües, podem detectar de manera automàtica els candidats traduïts (és a dir, les traduccions que s'han fet servir en el mateix corpus). Veiem el següent exemple, on podem observar les oracions on apareix el terme en anglès *legal framework* i les corresponents oracions traduïdes amb el corresponent equivalent de traducció *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 procés estadístic per al càlcul del equivalent de traducció en un corpus paral·lel és molt senzill:

- Es busquen tots els parells d'oracions on aparegui el terme a cercar en la part corresponent a la llengua de partida i guardem les oracions corresponents a la llengua d'arribada.
- Amb totes les oracions de la llengua d'arribada que hem guardat fem un procés d'extracció de terminologia i filtrem els resultats amb la llista de paraules buides corresponent a la llengua d'arribada.
- El candidat a terme que aparegui amb més freqüència serà el candidat a equivalent de traducció que cerquem.

En aquest exemple, si extraiem els candidats de la part castellana i filtrem per stopwords, obtenim els següents resultats:

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

On el candidat més freqüent, *marco jurídico*, és realment l'equivalent de traducció del terme cercat. Per a que aquesta tècnica pugui funcionar bé s'han de donar les següents circumstàncies:

- El terme a cercar ha d'aparèixer amb certa freqüència (si apareix només una vegada serà difícil obtenir l'equivalent de traducció).
- En el corpus s'ha d'haver fet servir una traducció més o menys estable del terme (si a cada oració s'ha fet servir una traducció diferent el sistema no podrà detectar la traducció correcta)
- Les oracions on apareix el terme han de ser diferents (si un terme apareix 100 vegades en un corpus però és una mateixa oració que es repeteix molt, l'efecte és exactament el mateix de que l'oració aparegui només una vegada)

TBXTools implementa la funció de cerca a corpus paral·lels. A continuació podem observar un exemple de codi. En corpus molt grans, convé fixar un nombre de vegades màxim d'aparicions del terme, per poder parar la cerca abans d'explorar tot el corpus. Aquest comportament es controla amb el paràmetre *limitsents* que per defecte té 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 pot donar més d'una opció, de manera que si no l'encerta el revisor pugui acceptar alguna de les altres opcions. En el següent exemple el nombre d'opcions estava fixat 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. Mètodes lingüístics per a l'extracció automàtica de terminologia

Els mètodes lingüístics per a l'extracció automàtica de terminologia fan servir característiques lingüístiques dels termes per dur a terme la detecció. La característica més emprada són els *patrons morfosintàctics*. Els patrons morfosintàctics són combinacions d'etiquetes morfosintàctiques que defineixen combinacions que són típicament terminològiques. Així per exemple (per a l'anglès) podríem definir una sèrie de patrons:

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

NN NN indica una combinació de Nom i Nom (tots dos en singular). JJ NN indica una combinació d'adjectiu i nom, i NN /of/ NN indica una combinació de nom, la paraula of i un altre nom. Alguns termes que segueixen el patró NN NN podrien ser *payment order* i *interest rate*; exemples de JJ NN serien *monetary policy* i *direct debit* i exemples de NN /of/ NN serien *economy of scale* i *point of sale*. Cal tenir en compte que no totes les combinacions que compleixin aquests patrons seran realment termes, sinó que es detectaran també moltes altres combinacions no terminològiques.

Per poder dur a terme extracció de terminologia seguint la metodologia lingüística serà imprescindible disposar d'un etiquetador morfosintàctica per a la llengua de treball. En l'apartat *Per ampliar coneixements* del capítol 2 d'aquest mateix llibre vam parlar d'aquestes eines. A continuació veiem un exemple de text etiquetat morfosintàcticament fent servir 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
```

Per a cada paraula del text tenim la forma, el lema, una etiqueta morfosintàctica i la probabilitat que aquesta etiqueta sigui la correcta (recordem que moltes paraules són ambigües des del punt de vista morfosintàctic). Les etiquetes morfosintàctiques s'expressen mitjançant un etiquetari (*tagset*) que sovint és depenent de la llengua. A continuació observem l'etiquetari del Penn Treebank (Santorini 1990) per a l'anglès, que és el que fa servir l'analitzador Freeling per a l'anglès.

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

Altres llengües fan servir etiquetaris diferents. El castellà i català a Freeling, per exemple, fan servir les etiquetes 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
```

Formalisme per als patrons en TBXTools

TBXTool fa servir un formalisme força potent per a l'expressió dels patrons terminològics. Per explicar aquest formalisme ho farem a través d'una sèrie d'exemples en anglès, català i castellà. Comencem per l'anglès. Considerem que tenim un text com el següent:

> 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.

Per poder fer l'extracció lingüística necessitem tenir aquest text etiquetat:

```
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
```

Els patrons terminològics es representen com a una seqüència d'etiquetes morfosintàctiques. Si com a patró posem **NN NN** (es a dir *Noun Singular or mass* seguit de *Noun Singular or mass*) obtenim els següents candidats (tots amb freqüència 1):

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

Fixem-nos que com hem fixat els patrons a NN només detecta els noms en singular. Si volem detecti també noms en plural podem fer ús de les expressions regulars i indicar el següent patró: **NN NN.***. Ara obtindrem els següents candidats:

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

També ens interessaria agrupar les formes singulars i plurals d'un mateix lema i que al terme detectat s'expressés el lema. El formalisme ens ho permet fer amb els parèntesis quadrats []: **NN [NN.*]** Fent ús d'aquest patró obtindríem els següents candidats:

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

Fixem-nos que hem ajuntat *interest rate* i *interest rates* en una única forma base, però comptant com a freqüència 2, ja que hi apareix dues vegades.

Ara, per explicar una altra característica del formalisme passarem a fer un exemple del castellà. Considerem que tenim el següent text:

> 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 .

I la seva corresponent anàlisi morfosintàctica:

```
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
```

Els patrons poden expressar paraules en comptes d'etiquetes fent servir / /. Per exemple, el patró [**NC.*] /de/ NC.*** detectaria el següent candidat:

```
2       tipo de interés
```

amb freqüència 2, ja que apareix tant en singular com en plural. Recordeu que les etiquetes morfosintàctiques són depenents de la llengua i que en el cas del castellà són diferents que per a l'anglès.

En una mateixa extracció habitualment es fa servir un conjunt de patrons morfosintàctics. Per exemple, per a l'anglès es podrien fer servir aquests:

```
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.*
```

I obtindríem uns candidats com els següents:

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

Detecció de termes aniuats

En el cas de l'extracció lingüística ens podem trobar també en casos de termes aniuats, com passava amb les estratègies estadístiques. L'avantatge és que la cerca dels aniuaments pot anar més guiada, ja que a partir de l'observació dels patrons terminològics es pot preveure quins casos d'aniuament es produiran.

Detecció de termes monoparaula

La detecció de termes monoparaula (és a dir, aquells termes formats per una única paraula) també suposa un problema per a les tècniques lingüístiques. El patró típic per a l'anglès seria **[NN.*]**, però aquest patró detectaria tots els substantius del text.

Detecció d'equivalents de traducció fent servir estratègia estadística

La cerca d'equivalents de traducció es pot dur a terme amb una estratègia lingüística. El principal problema es produeix pel fet que un mateix patró terminològic en la llengua de partida pot correspondre a diversos patrons de traducció per a la llengua d'arribada.

Patró eng	**Exemple**	**Traducció spa**	**Patró traducció 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.*

Per aquest motiu, tot i que l'extracció terminològica es dugui a terme amb una estratègia lingüística, sovint la cerca d'equivalents de traducció es fa seguint l'estratègia lingüística

3.7.5. Mesures estadístiques

Fins ara hem endreçat els resultats de l'extracció de terminologia, tant per a l'estratègia lingüística com per a l'estadística, per freqüència d'aparició. La freqüència d'aparició, tot i ser una mesura estadística molt simple, ha demostrat ser molt efectiva en extracció de terminologia.

Hi ha tot un seguit de mesures estadístiques complementàries que es poden emprar en l'extracció automàtica de teminologia. Aquestes mesures estadístiques es poden classificar en dues dimensions: dimensió lingüística i dimensió estadística.

Atenent a la dimensió estadística les mesures es poden dividir segons expressen *unithood* o *termhood*:

- *Unithood*: expressa la força o l'estabilitat de les col·locacions sintagmàtiques (Pazienza 2005)
- *Termhood*: fa referència al grau en què una paraula pot ser considerada un terme en un cert domini (Alcina 2009)

La *unithood*, tot i que captura un aspecte important dels termes, no és una característica exclusiva dels termes ja que també s'aplica a moltes altres unitats lingüístiques complexes (formades per més d'una paraula). La *termhood*, en canvi, sí que és una característica pròpies dels termes, tant siguin multiparaula com si estan formats per una sola paraula.

Atenent a la dimensió estadístiques les mesures es poden classificar en:

- Mesures del grau d'associació
- Mesures de la significància de l'associació
- Mesures heurístiques

A la taula següent (reproduïda de Piazencia 2005) es poden observar algunes mesures estadístiques classificades segons les dimensions lingüística i estadística:

	grau d'associació	significància de l'associació	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 entrarem en molt detall sobre aquestes mesures i considerarem únicament el càlcul de les mesures aplicades a bigrames. L'explicació, de nou, l'obtenim de Piazencia (2005) i qui vulgui aprofundir en detalls podrà consultar aquesta font.

Les mesures d'associació es fan servir per estimar la *unithood* i, com hem dit, es fan servir no només en terminologia, sinó de manera general per estimar les col·locacions entre dues paraules (*u* i *v*), basant-se en les evidències estadístiques sobre l'ocurrència d'aquestes paraules en el corpus. Aquestes evidències s'expressen mitjançant una *taula de contingència* de freqüències. A continuació presentem aquests càlculs pel cas dels bigrames. Anomenarem U i V a la primera i segona paraula de la col·locació. La coocurrència de (u,v) s'expressa amb la freqüència O11 i N és el nombre total de coocurrències en el corpus (N = O11+ O12+ O21+ O22).

	V=v	V≠v
U=u	O11	O12
U≠u	O21	O22

També podem definir les *freqüències marginals* com:

R1 = O11+ O12
R2 = O21+ O22
C1 = O11+ O21
C2 = O12+ O22

A la taula següent (Piazencia 2005) es poden observar les fórmules de càlcul de diverses mesures estadístiques emprades en extracció de terminologia:

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	$CO = -\frac{\sum_N\sum_M O_{11i}}{\lvert N\rvert}$

Hi ha molta bibliografia sobre quina d'aquestes mesures és la que funciona millor en la tasca d'extracció automàtica de terminologia. Piacencia (2005) arriba a la conclusió de que la millor mesura és la freqüència (conclusió a la que també arriben Daile (1994) i Evert (2001), seguida per *T-score* i *C-value*. En canvi troba, a diferència que en altres estudis, que *Log Likelihood Ratio* no funciona tan bé, tot i que millor que MI, MI^3 i *Dice Factor*.

Més recentment (Vázquez, 2014) arriba a la conclusió que la freqüència i *T-Score* (també anomenada *T-student*) són les mesures estadístiques que funcionen millor.

Com a conclusió podem dir que la freqüència és una de les principals mesures estadístiques per aplicar a l'extracció automàtica de terminologia. Aquesta mesura és molt simple de calcular (simplement mitjançant comptatges) i per tant la seva interpretació també és molt senzilla.

3.8. Conclusions

En aquest capítol hem fet una visió general dels conceptes relacionats amb la terminologia amb una visió pràctica orientada a la traducció. Hem après a crear bases de dades terminològiques fent servir diferents recursos i hem presentat amb detall les principals tècniques per a l'extracció automàtica de terminiologia. També s'ha presentat el format estàndard per a l'intercanvi de bases de dades terminològiques: el TBX (*Term Base eXchange*).

Bibliografia

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