

# Parallel corpora of new Catalan Civil Code: A terminology harmonization project

Mercè Lorente, Lluís de Yzaguirre,  
M. Amor Montané, Jorge Vivaldi  
Grup IULATERM  
Institut de Lingüística Aplicada (IULA)  
Universitat Pompeu Fabra (UPF)

# Outline

1. Introduction:
  - Catalan Civil Code project
  - Aims of the paper
2. The ALINEA tool
  - Tool specifications
  - Results
  - Application to the Catalan Civil Code project
  - Future improvements
3. The Catalan Civil Code project
4. Conclusions

# Catalan Civil Code project

- **Aims**

- Parallelization of the original text in Catalan and the respective translations into Spanish and English of the Catalan Civil Code
- Automatic term extraction of the Catalan Civil Code terminology (Catalan – Spanish – English)
- Terminology harmonization in the Spanish and English translations of the Catalan Civil Code

- **Partners**

- Parlament de Catalunya
- Departament de Justícia de la Generalitat de Catalunya
- Societat d'Estudis Jurídics de l'Institut d'Estudis Catalans
- Grup IULATERM de l'IULA, UPF

# Aims of the paper

- Description of features and operation of the ALINEA tool
- Description of parallel corpora processed in the Catalan Civil Code project
- Description of materials and results of the Catalan Civil Code project

# ALINEA tool: Description

- Author: Lluís de Yzaguirre
- Initial project: *Validador automàtic de Traduccions (para textos paralelos en lenguas románicas)*. (TS170/1999. Ministerio de Industria y Energía, ATYCA. Programa de fomento de tecnología industrial)
- PhD support: Ribas Bruguer, Marta (2006). *Alineació de textos jurídics paral·lels (català i castellà): alguns problemes*. UPF
- Description:
  - It uses the output of the lemmatization and disambiguation processes
  - It uses a massive calculation approach that allows the parallelization of translations which often are recreations of the original text
  - It allows refining the parallelized output with equivalence dictionaries automatically generated, with or without human supervision

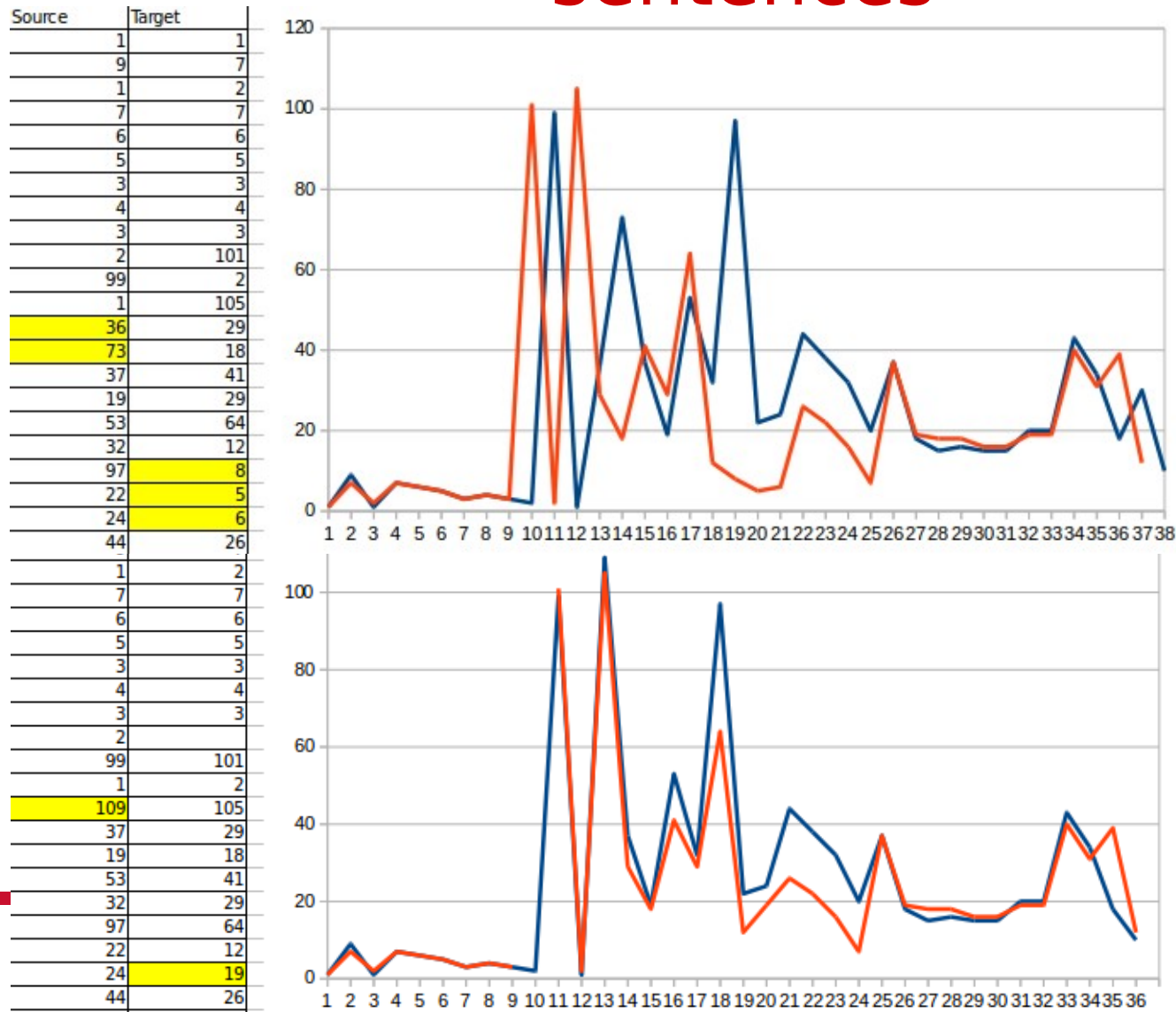
# ALINEA tool: Steps

- ALINEA steps (decision levels):
  - Splitting of the **text**; decision of which source sentences and target sentences will be compared
  - Comparison of two **sentences**
  - Comparison of two **words**

# ALINEA: Text splitting and pairing of sentences

- Discard stop words while maintaining punctuation marks
- Classify sentences by the number of retained units (lexical words, multiwords detected during preprocessing and punctuation marks)
- Apply sentence comparison strategy (to obtain a similarity coefficient):
  - to longest sentences,
  - to all other sentences, one to one,
  - to sentences with lower coefficient, one or more to two or to several.

# ALINEA: Text splitting and pairing of sentences





# Example of ALINEA parallelization

Parell 34 = 036:2 (000572-000619) vs. 036:1 (000763-000801)

Per provisió de data 25 de novembre de 1996 , es va assenyalar la vista per al dia 12 de desembre de 1996 , a les 11 hores, la qual va tenir lloc el dia i hora assenyalats.

En aquesta vista el lletrat de la part recurrent va sol·licitar que es cassés la sentència objecte de recurs d'acord amb la súplica de l'escrit d'interposició del recurs; i el lletrat de la part recorreguda, que es desestimés el recurs i es confirmés la sentència objecte de recurs amb expressa condemna a pagar les costes.

Por providencia de fecha 25 de noviembre de 1996 , se señaló para que tuviera lugar la vista el día 12 de diciembre de 1996 a las once horas, celebrándose el día y hora señalados, solicitando el letrado de la parte recurrente que se casara la sentencia recurrida de acuerdo con el súplico del escrito de interposición del recurso y el letrado de la parte recurrida que se desestimara el recurso y se confirmara la sentencia recurrida con expresa condena en costas.

# ALINEA: comparing sentences

- Each word in a sentence of source language is compared with all words of a given sentence in target language; this word comparison returns a coefficient for each word pair
- After tabulating this coefficients we obtain a Cartesian product (see Table 1)
- Only highest values are retained (see Table 2)
- The mean value of retained coefficients is the sentence pairing coefficient

# Example: ALINEA sentence comparison

**Table 1**

	congruència	:	concepte	modalitats	.
congruència	1125	80	392	297	20
:	80	375	80	60	240
concepte	392	80	825	337	60
modalitats	272	60	337	975	80
.	20	240	60	80	375

**Table 2**

	congruència	:	concepte	modalitats	.
congruència	1125	0	0	0	0
:	0	375	0	0	0
concepte	0	0	825	0	0
modalitats	0	0	0	975	0
.	0	0	0	0	375

# Example: ALINEA sentence comparison

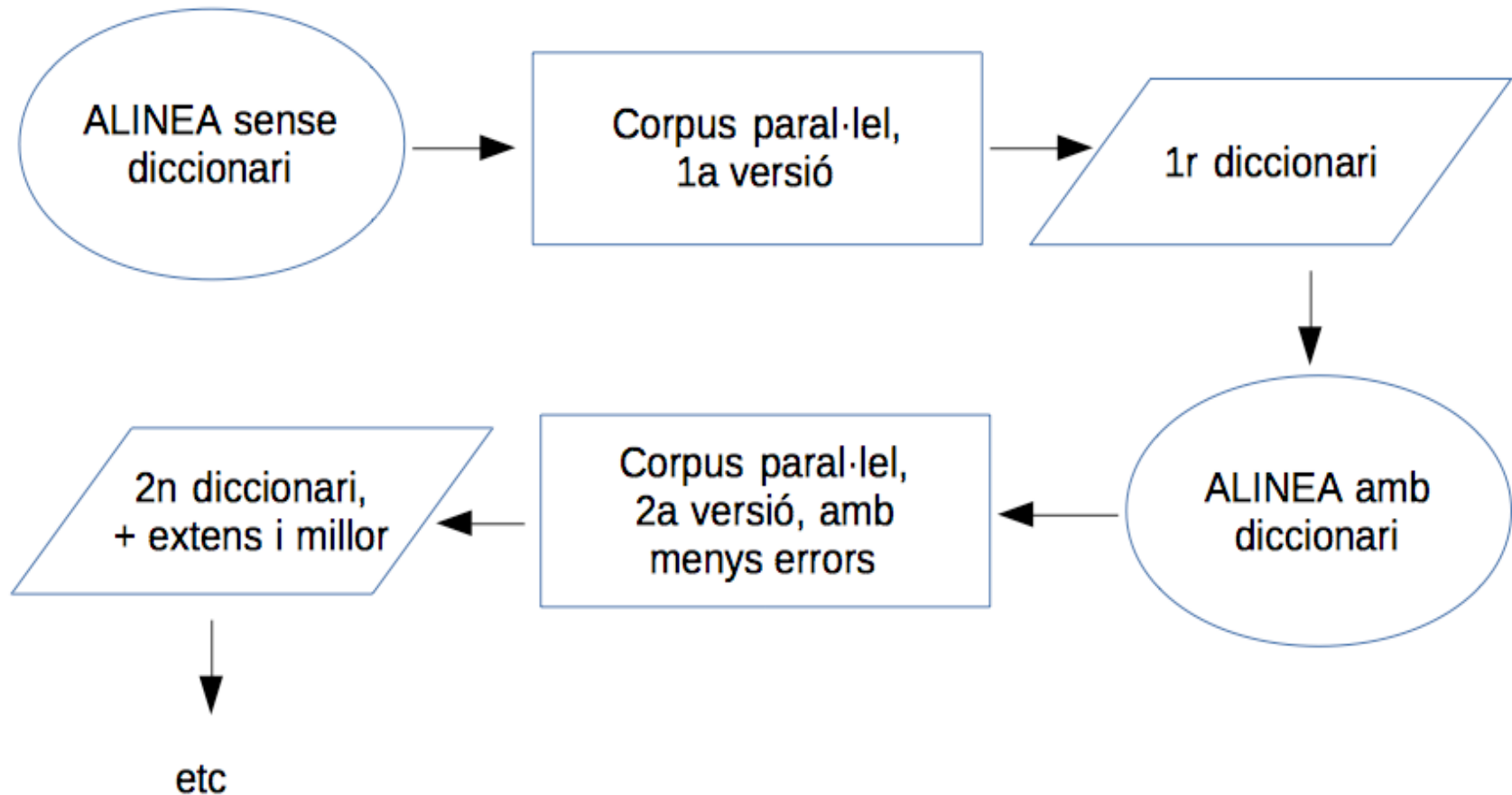
**Table 1 to 2**

	congruència	:	concepto	modalidades	.
congruència	1125	80	392	297	20
:	80	375	80	60	240
concepte	392	80	825	337	60
modalitats	272	60	337	975	80
.	20	240	60	80	375

# ALINEA: Comparison of words

- At the third level, word comparison extracts metrics from:
  - $\alpha$  comparative position of word A in sentence A (source) vs. position of word B in sentence B (target)
  - $\beta$  the highest (thru bigrams) of:
    - $\beta_1$  similarity of both tokens as they are (including case)
    - $\beta_2$  similarity of lemmata of both words
    - $\beta_3$  similarity of both SIMEX functions
  - $\gamma$  morphosyntactical labels similarity

# ALINEA: working with dictionaries



# Example of ALINEA word alignment

Parell 684. Frase 0689 (2) vs frase 0698 (1)	SSC = 457 (Traça=5)	Fitxer=2CPW1CQH
<p>Upon termination of guardianship, the guardian or, where applicable, the guardian and the administrator of property must submit final accounts of the guardianship to the judicial authority within <b>six</b> months after termination thereof .</p> <p>The judicial authority may extend this period, for just cause, for a further period of <b>three</b> months at most .</p>	<p>En acabar la tutela, el tutor o, si escau, el tutor i l'administrador patrimonial han de retre comptes finals de la tutela a l'autoritat judicial en el termini de <b>sis</b> mesos <b>a partir de</b> l'extinció d'aquella, prorrogables judicialment, per una causa justa, per un altre període de <b>tres</b> mesos <b>com a</b> màxim</p>	<ul style="list-style-type: none"> <li>• 999,months,N5P,mesos,N5-MP (11134 vs 11778)</li> <li>• 999,guardianship,N5S,tutela,N5-FS (11116 vs 11760)</li> <li>• 999,six,X,sis,X (11133 vs 11777)</li> <li>• 999,judicial,JA,judicial,JQ--6S (11131 vs 11775)</li> <li>• 999,final,JA,finals,JQ--6P (11128 vs 11772)</li> <li>• 998,months,N5P,mesos,N5-MP (11150 vs 11793)</li> <li>• 997,three,X,tres,X (11149 vs 11792)</li> <li>• 997,guardianship,N5S,tutela,N5-FS (11130 vs 11773)</li> <li>• 997,administrator,N5S,administrador,N5-MS (11124 vs 11767)</li> <li>• 995,accounts,N5P,comptes,N5-MP (11129 vs 11771)</li> <li>• 527,authority,N5S,autoritat,N5-FS (11132 vs 11774)</li> <li>• 484,period,N5S,període,N5-MS (11148 vs 11791)</li> <li>• 464,termination,N5S,termini,N5-MS (11135 vs 11776)</li> <li>• 398,judicial,JA,judicialment,D (11138 vs 11783)</li> <li>• 314,applicable,JA,patrimonial,JQ--6S (11121 vs 11768)</li> <li>• 312,authority,N5S,extinció,N5-FS (11139 vs 11779)</li> <li>• 282,further,JC,justa,JQ--FS (11147 vs 11787)</li> <li>• 274,most,DDS7,com a,D (11151 vs 11794)</li> <li>• 272,cause,VMI----,causa,N5-FS (11145 vs 11786)</li> <li>• 257,where,N5S,tutor,N5-MS (11120 vs 11762)</li> <li>• 252,extend,VMI----,retre,VI---- (11141 vs 11770)</li> <li>• 245,guardian,N5S,tutor,N5-MS (11123 vs 11766)</li> <li>• 238,period,N5S,màxim,N5-MS (11142 vs 11795)</li> <li>• 237,may,VOFDR66,han,VDR3P- (11140 vs 11769)</li> <li>• 231,must,VOFDR66,escau,VDR3S- (11126 vs 11764)</li> <li>• 215,submit,VMI----,acabar,VI---- (11127 vs 11759)</li> <li>• 185,property,N5S,prorrogables,JQ--6P (11125 vs 11782)</li> <li>• 148,thereof,DGA7,altre,EN--MS (11136 vs 11790)</li> <li>• 113,guardian,N5S,aquella,ED--FS (11118 vs 11780)</li> </ul>

# Example of ALINEA word alignment

Parell 684. Frase 0689 (2) vs frase 0698 (1)

SSC = 457 (Traça=5)

Upon termination of guardianship, the guardian or, where applicable, the guardian and the administrator of property must submit final accounts of the guardianship to the judicial authority within **six** months after termination thereof .

En acabar la tutela, el tutor o, si escau, el tutor i l'administrador patrimonial han de retre comptes finals de la tutela a l'autoritat judicial en el termini de **sis** mesos **a partir de** l'extinció d'aquella, prorrogables judicialment, per una causa justa, per un altre període de **tres** mesos **com a** màxim .

The judicial authority may extend this period, for just cause, for a further period of **three** months at most .



# Example of ALINEA word alignment

Fitxer=2CPW1CQH

- 999,months,N5P,mesos,N5-MP (11134 vs 11778)
- 999,guardianship,N5S,tutela,N5-FS (11116 vs 11760)
- 999,six,X,sis,X (11133 vs 11777)
- 999,judicial,JA,judicial,JQ--6S (11131 vs 11775)
- 999,final,JA,finals,JQ--6P (11128 vs 11772)
- 998,months,N5P,mesos,N5-MP (11150 vs 11793)
- 997,three,X,tres,X (11149 vs 11792)
- 997,guardianship,N5S,tutela,N5-FS (11130 vs 11773)
- 997,administrator,N5S,administrador,N5-MS (11124 vs 11767)
- 995,accounts,N5P,comptes,N5-MP (11129 vs 11771)
- 527,authority,N5S,autoritat,N5-FS (11132 vs 11774)
- 484,period,N5S,període,N5-MS (11148 vs 11791)
- 464,termination,N5S,termini,N5-MS (11135 vs 11776)
- 398,judicial,JA,judicialment,D (11138 vs 11783)
- 314,applicable,JA,patrimonial,JQ--6S (11121 vs 11768)
- 312,authority,N5S,extinció,N5-FS (11139 vs 11779)

- 274,most,DDS7,com a,D (11151 vs 11794)
- 272,cause,VMI----,causa,N5-FS (11145 vs 11786)
- 257,where,N5S,tutor,N5-MS (11120 vs 11762)
- 252,extend,VMI----,retre,VI---- (11141 vs 11770)
- 245,guardian,N5S,tutor,N5-MS (11123 vs 11766)
- 238,period,N5S,màxim,N5-MS (11142 vs 11795)
- 237,may,VOFDR66,han,VDR3P- (11140 vs 11769)
- 231,must,VOFDR66,escau,VDR3S- (11126 vs 11764)
- 215,submit,VMI----,acabar,VI---- (11127 vs 11759)
- 185,property,N5S,prorrogables,JQ--6P (11125 vs 11782)
- 148,thereof,DGA7,altre,EN--MS (11136 vs 11790)
- 113,guardian,N5S,aquella,ED--FS (11118 vs 11780)

# ALINEA future improvements

- The informations that ALINEA could incorporate in its decisions include (but are not limited to) the following:
  - Syntactic information, morphological ambiguity
  - Wordnet / MCR synsets
  - Information from Wikipedia and Dbpedia
  - Equivalences in a 3<sup>rd</sup> language (or in more languages)
  - Lexicographical and terminological thematic tags
  - Interaction with other tools (collocations extractor, monolingual term extractor)

# Catalan Civil Code project: Corpus

- Catalan Civil Code: 5 books (+ 1 non-published)
  - *First book. General provisions*
  - *Second book. Persons and family*
  - *Third book. Legal persons*
  - *Fourth book. Successions*
  - *Fifth book. Real rights*
- Description of the corpus
  - Catalan: original text
  - Spanish: automatic translations with post-editing
  - English: functional translations

# Catalan Civil Code project: Terminological harmonization

- Elaboration of a report with terminological problems detected in the Spanish and English translations:
  - equivalence
  - lack of sistematicity
  - precision
  - lack of equivalent
  - omission of equivalent

# Example of problems of equivalence

Catalan version	English translation
<p>Tot i mantenir la denominació tradicional, la <b>quarta vidual</b> ja no s'atribueix només al cònjuge vidu, sinó també al membre supervivent d'una unió estable de parella, i <b>no consisteix pròpiament</b> en una <b>quarta part</b> del <b>cabal relicte</b>, ja que la quarta part, com ja passava abans de l'entrada en vigor d'aquesta llei, actua només com a límit màxim.</p>	<p>Despite maintaining its traditional name, the <b>widow allowance</b> is not only attributed to the widowed spouse, but also to the surviving member of a stable relationship, and <b>it does not involve a fourth of the inherited estate</b>, since a fourth only acts as a maximum limit, as was the case before this law came into effect.</p>

# Example of problems of no systematicity

Catalan version	English translation
Llei 10/2008, del 10 de juliol, del llibre quart del Codi civil de Catalunya, relatiu a les <b>successions</b>	Law 10/2008, of 10 July, of the fourth book of the Civil Code of Catalonia, regarding <b>successions</b>
Aquesta llei, seguint el pla de codificació del dret civil català traçat per la Llei 29/2002, del 30 de desembre, primera llei del Codi civil de Catalunya, aprova el llibre quart del Codi civil, dedicat al dret de <b>successions</b> .	In line with the Catalan civil law coding scheme as set out by Law 29/2002, of 30 December, the first law of the Civil Code of Catalonia, this law gives approval to the fourth book of the Civil Code relating to <b>inheritances</b> .

# Example of problems of precision

Catalan version	English translation
<p>Si l'heretament és cumulatiu, l'hereu, quan mor l'heretant, adquireix els béns exceptuats de l'adquisició de present i els adquirits per l'heretant després de l'atorgament de l'<b>heretament</b>.</p>	<p>If the inheritance pact is cumulative, upon the death of the inheritance pact executor the heir acquires the assets not included in current acquisition as well as those acquired by the inheritance pact executor following the execution of the <b>marital inheritance pact</b>.</p>

# Example of problems of lack of equivalent

Catalan version	English translation
<p>Els dots, les <b>tenutes</b>, els <b>aixovars</b> i els <b>cabalatges</b>, els <b>esponsalícis</b> o <b>escreixos</b>, els tantumdem, els pactes d'igualtat de béns i guanys i els altres drets similars constituïts abans de l'entrada en vigor d'aquesta llei es continuen regint pel text refós de la Compilació del dret civil de Catalunya, aprovat pel Decret legislatiu 1/1984, del 19 de juliol.</p>	<p>The dowries, <b>tenutes</b>, <b>aixovars</b> and <b>cabalatges</b>, <b>esponsalícis</b> or <b>escreixos</b>, the tantundem, the agreements of equal property and profits and other similar rights<sup>2</sup> established prior to the entry into force of this Act remain governed by the recast text of the Compilation of the Civil Law of Catalonia, adopted by Legislative Decree 1/1984, of 19 July.</p>



# Example of problems of omission of equivalent

Catalan version	English translation
Un cop finit el termini, s'han de cancel·lar <b>d'ofici</b> els assentaments del Registre de Fundacions corresponents a la fundació dissolta.	Once this time-limit has expired, all records related to the foundation in the Registry of Foundations <b>must be cancelled</b> .

# Codi civil de Catalunya



Generalitat de Catalunya  
Departament  
de Justícia



Universitat  
Pompeu Fabra  
Barcelona



## Problemes

Inici

### 1 Idioma: CA-ES

1.1 Problemes  
terminològics  
(equivalència)

1.1.1 *resoldre*

1.1.2 *ròssec*

1.1.3 *adaptar*

1.1.4 *produir*

1.1.5 *rector*

1.1.6 *capítols  
matrimonials*

1.1.7 *laude*

1.1.8 *recórrer*

1.1.9 *rabassa morta*

1.2 Problemes  
terminològics  
(asistematicitat)

No n'hi ha

1.3 Problemes  
terminològics (precisió)

1.3.1 *pervenir*

1.4 Problemes  
terminològics (manca  
d'equivalent)

1.4.1 *revessejat*

1.5 Problemes

Cas: 009 *rabassa morta*

Llibre: cinquè

Problema: terminològic (equivalència)

Context:

Parell 67. Frase 0075 (1) vs frase 0073 (1)

SSC = 510 (Traça=2)

Es mantenen, també, normatives transitòries per als censos i les **rabasses mortes** constituïdes abans de l'entrada en vigor de la Llei 6/1990 i de la Llei 22/2001, amb el benentès que les normes que estableixen aquestes lleis per a facilitar l'extinció i la redempció d'aquests drets són llur actiu pràctic principal .

Se mantienen, también, normativas transitorias para los censos y las **rabasses mortes** constituidas antes de la entrada en vigor de la Ley 6/1990 y de la Ley 22/2001, bien entendido que las normas que establecían estas leyes para facilitar la extinción y redención de estos derechos son su activo práctico principal .

Recerca terminològica:

**rabassa morta** *c. nom. f.*

HISTÒRIA DEL DRET

Inespecífic

Contracte de conreu emfitèutic establert com a cessió d'un tros de terra per a conrear-hi vinya, amb la condició que el contracte restava dissolt en haver mort dos terços dels primers ceps plantats.

es primeras cepas

Font: *Diccionari Jurídic*


Comentari:


En la traducció a l'espanyol s'utilitza el manlleu. D'una banda, el *Diccionari Jurídic* dona com a equivalent *primeras cepas*. D'altra banda, caldria comprovar el tractament del manlleu pel que fa a la formació del plural i a la marca tipogràfica.


# Relevant links

- ALINEA (demo version)
- IULATERM Group
- [Handout](#)
- Animation

# Bibliography

Cabré, M. Teresa; De Yzaguirre, Lluís; Ribas, Marta; Vivaldi, Jordi (2000). "Some Technical Aspects about Aligning Near Languages". Dins Gavrilidou, M.; Carayannis, G.; Markantonatou, S.; Piperidis, S.; Stainhaouer, G. (ed.) Proceedings of Second International Conference on Language Resources and Evaluation. Athens: National Technical University of Athens II. Pàg. 545-548 

Cabré, M. Teresa; Tebé, Carles; De Yzaguirre, Lluís; Alonso, Araceli (2002). "Generación automática de diccionarios para la traducción especializada a partir de corpus alineados". Dins Chabás, Josep; Gaser, Ralph; Rey, Joelle (ed.) Proceedings ... International Conference on Specialized Translation. Barcelona: Universitat Pompeu Fabra. pàg. 187-199. ISBN 84-477-0820-9. 

Ribas Bruguer, Marta (2006). Alineació de textos jurídics paral·lels (català i castellà): alguns problemes. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra. ISBN 84-690-0319-4 

# Parallel corpora of new Catalan Civil Code: A terminology harmonization project

## Thank you

[iulaterm@upf.edu](mailto:iulaterm@upf.edu)

Grup IULATERM

Institut de Lingüística Aplicada (IULA)

Universitat Pompeu Fabra