This document is part of the distribution of the *Acquis Communautaire* Translation Memory in 22 languages, provided by the European Commission’s (EC) *Directorate General for Translation* (DGT) with support from the EC’s *Directorate General Joint Research Centre* (JRC). For information on the DGT Translation Memory, see http://langtech.jrc.it/DGT-TM.html. For the related distribution of the *JRC-Acquis* Multilingual Parallel corpus, see http://langtech.jrc.it/JRC-Acquis.html.

EUR-Lex pre-processing

The following describes the main operations carried out on EUR-Lex documents before they are submitted to Euramis Alignment. This description does not claim to be comprehensive. Note that Euramis alignments are not produced for documentary purposes: they are often incomplete as compared to the original documents. In addition, Euramis Alignment itself also manipulates text, e.g. sentence enumerators are deleted.

Deletions:

- start of document up to and including the string 'Document's text:' (or its translation)
- string 'TEXT CONTINUED UNDER DOC.NUM' until the end of the document
- replace multiple blanks by one
- blanks at beginning/end of line
- blank before: .;)?!, ^t (tab)
- blank after: (
- blanks inside dates of type '12. 10. 1997'
- dots at end of string
- returns preceding lines which consist exclusively of 'and' or 'or' (or their translations)
- graphics if enclosed by corresponding place holders (e.g. >START OF GRAPHIC> and its translations)
- upper case place holders for tables, graphics etc., if enclosed by angle brackets (e.g. >TABLE POSITION>; Swedish: also lower case and mixed
- the mention of duplicate in all languages ('DUPLIKAT', 'KAKSOISKAPPALE', ...)
- lines which consist exclusively of the string 'considérant ce qui suit:' and its equivalents in all languages
- lines which consist exclusively of numbers, parentheses, tabulators, commas, brackets of all kinds, quotes of all kinds, percentage signs, dots, slashes and blanks
- lines which consist exclusively of numbers, blanks and the string 'ex'
- footnotes and end notes which start with 'OJ', e.g. (1) OJ
- excess breaks in titles for all languages
- opening angle brackets which have no closing bracket (also after previous treatment)
- HTML codes

- excess closing angle brackets (also after previous treatment)
- as much garbage as possible at the end of the document

Modifications:

- convert table cells to paragraphs
- insert breaks between 'For the Commission', a name, and 'Member of the Commission' in all languages
- insert break after TITLE, PART and CHAPTER and the numberings used for them (Roman or Arabic) for all languages
- insert tabs in starting with, e.g., '1', '1.', '(a)', 'a)', '(1)', '1)'

Language-specific treatment:

- DA: heuristic treatment of diacritic characters: many EUR-Lex documents contain transcriptions, e.g. of 'å' to 'aa'. The program tries to change e.g. 'aa' to 'å' where necessary. Since it is impossible to cover all exceptions to that treatment (e.g. 'avokadoe' should not be changed to 'avokadø'), the results may contain errors.
- DE: similar treatment as for DA: many EUR-Lex documents contain transcriptions, e.g. of 'ä' to 'ae'. The program tries to change e.g. 'ae' to 'ä' where necessary. Since it is impossible to cover all exceptions to that treatment (e.g. 'Barbecue' should not be changed to 'Barbecü'), the results may contain errors.
- EN: introduce an additional return in those cases where sentences starting with 'whereas' (preceded by semicolon) are found inside paragraphs.
- ES: introduce an additional return in those cases where sentences starting with 'que' (preceded by semicolon) are found inside paragraphs.
- FI: delete line 'jotka katsovat' plus a break plus 'että'; delete excessive hard returns in titles.
- FR: introduce an additional return in those cases where sentences starting with 'que' or 'qu'' (preceded by semicolon) are found inside paragraphs.
- IT: introduce an additional return in those cases where sentences starting with 'considerando' or 'che' (preceded by semicolon) are found inside paragraphs.
- NL: introduce an additional return in those cases where sentences starting with 'dat' (preceded by semicolon) are found inside paragraphs; some modifications following the Dutch spelling reform, e.g. change 'produkt' to 'product'.
- PT: introduce an additional return in those cases where sentences starting with 'que' (preceded by semicolon) are found inside paragraphs