

# Multilingual Lexical Database Generation from parallel texts with endogenous resources

Emmanuel Giguët    Pierre-Sylvain Luquet

GREYC CNRS UMR 6072, Université de Caen, F14032 Caen Cedex, France,  
e-mail : {Emmanuel.Giguët,Pierre-Sylvain.Luquet}@info.unicaen.fr

## Abstract

This paper deals with multilingual database generation from parallel corpora. The idea is to contribute to the enrichment of lexical databases for languages with few linguistic resources. Our approach is endogenous: it relies on the raw texts only, it does not require external linguistic resources such as stemmers or taggers. The system produces alignments for the 20 European languages of the ‘Acquis Communautaire’ Corpus.

## 1 Introduction

### 1.1 Automatic processing of bilingual and multilingual corpora

Processing bilingual and multilingual corpora constitutes a major area of investigation in natural language processing. The linguistic and translational information that is available make them a valuable resource for translators, lexicographers as well as terminologists. They constitute the nucleus of example-based machine translation and translation memory systems.

Another field of interest is the constitution of multilingual lexical databases such as the project planned by the European Commission's Joint Research Centre (JRC) or the more established Papillon project. Multilingual lexical databases are databases for structured lexical data which can be used either by humans (e.g. to define their own dictionaries) or by natural language processing (NLP) applications.

Parallel corpora are freely available for research purposes and their increasing size demands the exploration of automatic methods. The ‘Acquis Communautaire’ (AC) Corpus is

such a corpus. Many research teams are involved in the JRC project for the enrichment of a multilingual lexical database. The aim of the project is to reach an automatic extraction of lexical tuples from the AC Corpus.

The AC document collection was constituted when ten new countries joined the European Union in 2004. They had to translate an existing collection of about ten thousand legal documents covering a large variety of subject areas. The ‘Acquis Communautaire’ Corpus exists as a parallel text in 20 languages. The JRC has collected large parts of this document collection, has converted it to XML, and provide sentence alignments for most language pairs (Erjavec et al., 2005).

### 1.2 Alignment approaches

Alignment becomes an important issue for research on bilingual and multilingual corpora. Existing alignment methods define a continuum going from purely statistical methods to linguistic ones. A major point of divergence is the granularity of the proposed alignments (entire texts, paragraphs, sentences, clauses, words) which often depends on the application.

In a coarse-grained alignment task, punctuation or formatting can be sufficient. At finer-grained levels, methods are more sophisticated and combine linguistic clues with statistical ones. Statistical alignment methods at sentence level have been thoroughly investigated (Gale & Church, 1991a/ 1991b ; Brown et al., 1991 ; Kay & Röscheisen, 1993). Others use various linguistic information (Simard et al., 1992 ; Papageorgiou et al., 1994). Purely statistical alignment methods are proposed at word level (Gale & Church, 1991a ; Kitamura & Matsumoto, 1995). (Tiedemann, 1993 ; Boutsis & Piperidis, 1996 ; Piperidis et al., 1997) combine statistical and linguistic information for the same task. Some methods make alignment suggestions at an intermediate

level between sentence and word (Smadja, 1992; Smadja et al., 1996; Kupiec, 1993; Kumano & Hirakawa, 1994; Boutsis & Piperidis, 1998).

A common problem is the delimitation and spotting of the units to be matched. This is not a real problem for methods aiming at alignments at a high level of granularity (paragraphs, sentences) where unit delimiters are clear. It becomes more difficult for lower levels of granularity (Simard, 2003), where correspondences between graphically delimited words are not always satisfactory.

## 2 General Instructions

The approach proposed here deals with the spotting of multi-grained translation equivalents. We do not adopt very rigid constraints concerning the size of linguistic units involved, in order to account for the flexibility of language and translation divergences. Alignment links can then be established at various levels, from sentences to words and obeying no other constraints than the maximum size of candidate alignment sequences and their minimum frequency of occurrence.

The approach is *endogenous* since the input is used as the only used linguistic resource. It is the multilingual parallel AC corpus itself. It does not contain any syntactical annotation, and the texts have not been lemmatised. In this approach, no classical linguistic resources are required. The input texts have been segmented and aligned at sentence level by the JRC. Inflectional divergencies of isolated words are taken into account without external linguistic information (lexicon) and without linguistic parsers (stemmer or tagger). The morphology is learnt automatically using an endogenous parsing module integrated in the alignment tool based on (Déjean, 1998).

We adopt a *minimalist* approach, in the line of GREYC. In the JRC project, many languages do not have available linguistic resources for automatic processing, neither inflectional or syntactical annotation, nor surface syntactic analysis or lexical resources (machine-readable dictionaries etc.). Therefore we can not use a large amount of *a priori* knowledge on these languages.

## 3 Considerations on the Corpus

### 3.1 Corpus definition

Concretely, the texts constituting the AC corpus are legal documents translated in several languages and aligned at sentence level. Here is a description of the parallel corpus, in the 20 languages available:

- Czech: cs.tar.gz 7106 docs, 131.0Mo
- Danish: da.tar.gz 8223 docs, 144.6Mo
- German: de.tar.gz 8249 docs, 152.0Mo
- Greek: el.tar.gz 8003 docs, 242.7Mo
- English: en.tar.gz 8240 docs, 138.9Mo
- Spanish: es.tar.gz 8207 docs, 156.7Mo
- Estonian: et.tar.gz 7844 docs, 144.5Mo
- Finnish: fi.tar.gz 8189 docs, 136.6Mo
- French: fr.tar.gz 8254 docs, 155.4Mo
- Hungarian: hu.tar.gz 7535 docs, 153.7Mo
- Italian: it.tar.gz 8249 docs, 152.4Mo
- Lithuanian: lt.tar.gz 7520 docs, 140.8Mo
- Latvian: lv.tar.gz 7867 docs, 152.5Mo
- Maltese: mt.tar.gz 6136 docs, 105.2Mo
- Dutch: nl.tar.gz 8247 docs, 153.8Mo
- Polish: pl.tar.gz 7768 docs, 154.0Mo
- Portuguese: pt.tar.gz 8210 docs, 153.0Mo
- Slovakian: sk.tar.gz 6963 docs, 131.3Mo
- Slovene: sl.tar.gz 7821 docs, 133.3Mo
- Swedish: sv.tar.gz 8233 docs, 133.7Mo

The documents contained in the archives are XML files, UTF-8 encoding, containing information on “sentence” tokenization. Each file is stamped with a unique identifier (the *celex* identifier). It refers to a unique document. Figure 1 is an excerpt of the document 31967R0741, in Czech.

Sentence alignments files are also provided with the corpus for 111 language pairs. The XML files encoded in UTF-8 are about 2M packed and 10M unpacked. Figure 2 is an excerpt of the alignment file of the document 31967R0741, for the language pair Czech-Danish.

In this file, the *xtargets* “ids” refer to the <P sid=“...”> of the Czech and Danish translations of the document 31967R0741.

```

<document celex="31967R0741" lang="cs" ver="1.0">
  <title>
    <P sid="1">NAŘÍZENÍ RADY č. 741/67/EHS ze dne 24. října 1967 o příspěvcích ze
      záruční sekce Evropského orientačního a záručního fondu</P>
  </title>
  <text>
    <P sid="2">NAŘÍZENÍ RADY č. 741/67/EHS</P>
    <P sid="3">ze dne 24. října 1967</P>
    <P sid="4">o příspěvcích ze záruční sekce Evropského orientačního a záručního
      fondu</P>
    <P sid="5">RADA EVROPSKÝCH SPOLEČENSTVÍ,</P>
    <P sid="6">s ohledem na Smlouvu o založení Evropského hospodářského společenství,
      a zejména na článek 43 této smlouvy,</P>
    <P sid="7">s ohledem na návrh Komise,</P>
    <P sid="8">s ohledem na stanovisko Shromáždění1,</P>
    <P sid="9">vzhledem k tomu, že zavedením režimu jednotných a povinných náhrad při
      vývozu do třetích zemí od zavedení jednotné organizace trhu pro zemědělské
      produkty, jež ve značné míře existuje od 1. července 1967, vyšlo kritérium nejnižší
      průměrné náhrady stanovené pro financování náhrad podle čl. 3 odst. 1 písm. a)
      nařízení č. 25 o financování společné zemědělské politiky2 z používání;</P>
  </text>
  [...]

```

Fig 1: Excerpt of the document 31967R0741 in Czech

```

<document celexid="31967R0741">
  <title1>NAŘÍZENÍ RADY č. 741/67/EHS ze dne 24. října 1967 o příspěvcích ze záruční sekce
    Evropského orientačního a záručního fondu</title1>
  <title2>Raadets forordning nr. 741/67/EOEF af 24. oktober 1967 om stoette fra Den
    europaeiske Udviklings- og Garantifond for Landbruget, garantisektionen</title2>
  <link type="1-2" xtargets="2;2 3" />
  <link type="1-1" xtargets="3;4" />
  <link type="1-1" xtargets="4;5" />
  <link type="1-1" xtargets="5;6" />
  [...]
  <link type="1-1" xtargets="49;53" />
  <link type="2-1" xtargets="50 51;54" />
  <link type="1-1" xtargets="52;55" />
</document>

```

Fig 2: Excerpt of the alignment file of the document 31967R0741 for the language pair Czech-Danish

The current version of our alignment system deals with one language pair at a time, whatever the languages are. The algorithm takes as input a corpus of bitexts aligned at sentence level. Usually, the alignment at this level outputs aligned windows containing from 0 to 2 segments. One-to-one mapping corresponds to a standard output (see link types “1-1” above). An empty window corresponds to a case of addition in the source language or to a case of omission in the target language. One-to-two mapping corresponds to split sentences (see link types “1-2” and “2-1” above).

Formally, each bitext is a quadruple  $\langle T1, T2, Fs, C \rangle$  where T1 and T2 are the two texts, Fs is the function that reduces T1 to an element set  $Fs(T1)$  and also reduces T2 to an element set

$Fs(T2)$ , and C is a subset of the Cartesian product of  $Fs(T1) \times Fs(T2)$  (Harris, 1988).

Different standards define the encoding of parallel text alignments. Our system natively handles TMX and XCES format, with UTF-8 or UTF-16 encoding.

## 4 The Resolution Method

The resolution method is composed of two stages, based on two underlying hypotheses. The first stage handles the document grain. The second stage handles the corpus grain.

### 4.1 Hypothesis

**hypothesis 1** : let's consider a bitext composed of the texts  $T_1$  and  $T_2$ . If a sequence  $S_1$  is repeated several times in  $T_1$  and in well-defined sentences<sup>1</sup>, there are many chances that a repeated sequence  $S_2$  corresponding to the translation of  $S_1$  occurs in the corresponding aligned sentences in  $T_2$ .

**hypothesis 2** : let's consider a corpus of bitexts, composed of two languages  $L_1$  and  $L_2$ . There is no guarantee for a sequence  $S_1$  which is repeated in many texts of language  $L_1$  to have a unique translation in the corresponding texts of language  $L_2$ .

## 4.2 Stage 1: Bitext analysis

The first stage handles the document scale. Thus it is applied on each document, individually. There is no interaction at corpus level.

### Determining the multi-grained sequences to be aligned

First, we consider the two languages of the document independently, the source language  $L_1$  and the target language  $L_2$ . For each language, we compute the repeated sequences as well as their frequency. The settings of the underlying algorithm are the minimum and maximum number of words forming the sequences, as well as the minimum frequency of the sequences that must be kept.

We use a greedy algorithm, similar to the algorithm used by (Vergne, 2005) for term extraction. The idea is to keep sequences of 1, 2, 3, ... words, while the sequence frequency in the document is greater than a particular threshold (1 for instance).

The algorithm does not retain the sub-sequences of a repeated sequence if they are as frequent as the sequence itself. For instance, if "subjects" appears with the same frequency than "healthy subjects" we retain only the second sequence. On the contrary, if "disease" occurs more frequently than "thyroid disease" we retain both.

When computing the frequency of a repeated sequence, the offset of each occurrence is memorized. So the output of this processing

stage is a list of sequences with their frequency and the offset list in the document.

*"thyroid cancer": list of segments where the sequence appears*  
45, 46, 46, 48, 51, 51, ...

### Handling inflections

Inflectional divergencies of isolated words are taken into account without external linguistic information (lexicon) and without linguistic parsers (stemmer or tagger). The morphology is learnt automatically using an endogenous approach derived from (Déjean, 1998). The algorithm is reversible: it allows to compute prefixes the same way, with reversed word list as input.

The basic idea is to approximate the border between the nucleus and the suffixes. The border matches the position where the number of distinct letters preceding a suffix of length  $n$  is greater than the number of distinct letters preceding a suffix of length  $n-1$ .

For instance, in the first English document of our corpus, "g" is preceded by 4 distinct letters, "ng" by 2 and "ing" by 10: "ing" is probably a suffix. In the first Greek document, "á" is preceded by 5 letters, "ká" by 1 and "iká" by 10. "iká" is probably a suffix.

The algorithm can generate some wrong morphemes, from a strictly linguistic point of view. But at this stage, no filtering is done in order to check their validity. We let the alignment algorithm do the job with the help of contextual information.

### Vectorial representation of the sequences

An *orthonormal space* is then considered in order to explore the existence of possible translation relations between the sequences, and in order to define translation couples. The existence of translation relations between sequences is approximated by the cosine of vectors associated to them, in this space.

The links in the alignment file allow the construction of this orthonormal space. This space has  $n_o$  dimensions, where  $n_o$  is the number of non-empty links. Alignment links with empty sets (`type="0-?"` or `type="?-0"`) corresponds to cases of omission or addition in one language.

---

<sup>1</sup> Here, « sentences » can be generalized as « textual segments »

Every repeated sequence is seen as a vector in this space. For the construction of this vector, we first pick up the segment offset in the document for each repeated sequence.

*“thyroid cancer”*: list of segments where the sequence appears  
45, 46, 46, 48, 51, 51

Then we convert this list in a  $n_L$ -dimension vector  $v_L$ , where  $n_L$  is the number of textual segments of the document of language  $L$ . Each dimension contains the number of occurrences present in the segment.

*“thyroid cancer”* : associated with a vector of  $n_L$  dimensions.

1	2	...	<b>45</b>	<b>46</b>	47	48	49	50	51	...	$n_L$
0	0		1	2	0	1	0	0	2		0

With the help of the alignment file, we can now make the projection of the vector  $v_L$  in the  $n_o$ -dimension vector  $v_o$ . For instance, if the link `<link type="2-1" xtargets="45 46;45" />` is located at rank  $r=40$  in the alignment file and if English is the first language ( $L=en$ ), then  $v_o[40] = v_{en}[45] + v_{en}[46]$ .

### Sequence alignment

For each sequence of  $L_1$  to be aligned, we look for the existence of a translation relation between it and every  $L_2$  sequence to be aligned. The existence of a translation relation between two sequences is approximated by the cosine of the vectors associated to them.

The cosine is a mathematical tool used in in Natural Language Processing for various purposes, e.g. (Roy & Beust, 2004) uses the cosine for thematic categorisation of texts. The cosine is obtained by dividing the scalar product of two vectors with the product of their norms.

$$\cos(x_i, y_i) = \frac{\sum x_i \cdot y_i}{\sqrt{\sum x_i^2} \times \sqrt{\sum y_i^2}}$$

We note that the cosine is never negative as vectors coordinates are always positive. The sequences proposed for the alignment are those that obtain the largest cosine. We do not propose

an alignment if the best cosine is inferior to a certain threshold.

### 4.3 Stage 2: Corpus management

The second stage handles the corpus grain and merges the information found at document grain, in the first stage.

#### Handling the Corpus Dimension

The bitext corpus is not a bag of aligned sentences and is not considered as if it were. It is a bag of bitexts, each bitext containing a bag of aligned sentences.

Considering the bitext level (or document grain) is useful for several reasons. First, for operational sake. The greedy algorithm for repeated sequence extraction has a cubic complexity. It is better to apply it on the document unit rather than on the corpus unit. But this is not the main reason.

Second, the alignment algorithm between sequences relies on the principle of translation coherence: a repeated sequence in  $L_1$  has many chances to be translated by the same sequence in  $L_2$  in the same text. This hypothesis holds inside the document but not in the corpus: a polysemic term can be translated in different ways according to the document genre or domain.

Third, the confidence in the generated alignments is improved if the results obtained by the execution of the process on several documents share compatible alignments.

#### Alignment Filtering and Ranking

The filtering process accepts terms which have been produced (1) by the execution on at least two documents, (2) by the execution on solely one document if the aligned terms correspond to the same character string or if the frequency of the terms is greater than an empirical threshold function. This threshold is proportional to the inverse term length since there are fewer complex repeated terms than simple terms.

The ranking process sorts candidates using the product of the term frequency by the number of output agreements.

## 5 Results

The results concern an alignment task between English and French, on 20 bitexts of the AC Corpus (see Annex for examples).

A previous evaluation (Giguet & Apidianaki, 2005) was achieved for alignment between a rich inflectional language (Greek) and a weak inflectional language (English), and carried on scientific papers.

Among the correct alignments, we find domain dependant lexical terms :

- legal terms of the EEC (*EEC initial verification / vérification primitive CEE, Regulation (EEC) No/règlement (CEE) n°*),
- specialty terms (*rear-view mirrors / rétroviseurs, poultry/volaille*).

We also find invariant terms (*km/h/km/h, kg/kg, mortem/mortem*).

We encounter alignments at different grain:

- *territory/territoire*
- *Member States/États membres,*
- *Whereas/Considérant que,*
- *fresh poultrymeat/viandes fraîches de volaille*
- *Having regard to the Opinion of the/vu l'avis.*

The wrong alignments mainly come from candidates that have not been confirmed by running on several documents (column ndoc=1): *on/la commercialisation des*.

### 5.1 Discussion

First, the results are similar to those obtained on the Greek/English scientific corpus.

Second, it is sometimes difficult to choose between distinct proposals for a same term when the grain vary : *Member/membre~ Member State~/membre~ Member States/États membres State/membre State~/membre~*. There is a problem both in the definition of terms and in the ability of an automatic process to choose between the components of the terms.

Third, thematic terms of the corpus are not always aligned, since they are not repeated. Coreference is used instead, thanks to nominal anaphora, acronyms, and also lexical reductions. Accuracy depends on the document domain. In

the medical domain, acronyms are aligned but not their expansion. However, we consider that this problem has to be solved by an anaphora resolution system, not by this alignment algorithm.

## 6 Conclusion

We showed that it is possible to contribute to the processing of languages for which few linguistic resources are available. We propose a solution to the spotting of multi-grained translation from parallel corpora. The results are surprisingly good and encourage us to improve the method, in order to reach a semi-automatic construction of a multilingual lexical database.

The endogenous approach allows to handle inflectional variations. We also show the importance of using the proper knowledge at the proper level (sentence grain, document grain and corpus grain). An improvement would be to calculate inflectional variations at corpus grain rather than at document grain.

The size of this “massive compilation” implies the design of specific strategies in order to handle it properly and quite efficiently. Special efforts have been done in order to manage the AC Corpus from our document management platform, WIMS.

The next improvement is to properly handle translations in multiple languages. An effort should be made to reduce silence. Another perspective is to integrate an endogenous coreference solver (Giguet & Lucas, 2004).

## References

- Altenberg B. & Granger, S. 2002. *Recent trends in cross-linguistic lexical studies*. In *Lexis in Contrast*, Altenberg & Granger (eds.), 2002.
- Boutsis, S., & Piperidis, S. 1998. *Aligning clauses in parallel texts*. In *Third Conference on Empirical Methods in Natural Language Processing*, 2 June, Granada, Spain, p. 17-26.
- Brown P., Lai J. & Mercer R. 1991. *Aligning sentences in parallel corpora*. In *Proc. 29<sup>th</sup> Annual Meeting of the Association for Computational Linguistics*, p. 169-176, 18-21 June, Berkley, California, 1991.
- Déjean H. 1998. *Morphemes as Necessary Concept for Structures Discovery from Untagged Corpora*.

- In Workshop on Paradigms and Grounding in Natural Language Learning, pages 295-299, PaGNLL Adelaide.
- (Erjavec et al., 2005). *Massive multilingual corpus compilation; Acquis Communautaire and totale*. In: 2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (L&T'05). Poznań, Poland, 21-23 April 2005.
- Gale W.A. & K.W. Church 1991a. *Identifying word correspondences in parallel texts*. In *Fourth DARPA Speech and Natural Language Workshop*, p. 152-157. San Mateo, California: Morgan Kaufmann.
- Gale W.A. & Church K. W. 1991b. *A Program for Aligning Sentences in Bilingual Corpora*. In *Proc. 29th Annual Meeting of the Association for Computational Linguistics*, p. 177-184, 18-21 June, Berkley, California.
- Giguet E. & Apidianaki M. 2005. *Alignement d'unités textuelles de taille variable*. Journée Internationales de la Linguistique de Corpus. Lorient.
- Giguet E. 2005. *Multi-grained alignment of parallel texts with endogenous resources*. RANLP'2005 Workshop "Crossing Barriers in Text Summarization Research". Borovets, Bulgaria.
- Giguet E. & Lucas N. 2004. *La détection automatique des citations et des locuteurs dans les textes informatifs*. In *Le discours rapporté dans tous ses états : Question de frontières*, J. M. López-Muñoz S. Marnette, L. Rosier, (eds.). Paris, l'Harmattan, pp. 410-418.
- Harris B. 1998. *Bi-text, a New Concept in Translation Theory*, Language Monthly (54), p. 8-10.
- Isabelle P. & Warwick-Armstrong S. 1993. *Les corpus bilingues: une nouvelle ressource pour le traducteur*. In Bouillon, P. & Clas A. (eds.), *La Traductique : études et recherches de traduction par ordinateur*. Montréal: Les Presses de l'Université de Montréal, p. 288-306.
- (Kay & Röscheisen, 1993) Kay M. & Röscheisen M. *Text-translation alignment*. Computational Linguistics, p.121-142, March 1993.
- (Kitamura & Matsumoto, 1996) Kitamura M. & Matsumoto Y. *Automatic extraction of word sequence correspondences in parallel corpora*. In *Proc. 4<sup>th</sup> Workshop on Very Large Corpora*, p. 79-87. Copenhagen, Denmark, 4 August 1996.
- Kupiec J. 1993. *An algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora*, *Proceedings of the 31<sup>st</sup> Annual Meeting of the Association of Computational Linguistics*, p. 23-30.
- Papageorgiou H., Cranias L. & Piperidis S. 1994. *Automatic alignment in parallel corpora*. In *Proceed. 32<sup>nd</sup> Annual Meeting of the Association for Computational Linguistics*, p. 334-336, 27-30 June, Las Cruces, New Mexico, 1994.
- Salkie R. 2002. *How can linguists profit from parallel corpora?*, In *Parallel Corpora, Parallel Worlds: selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22-23 April, 1999*, Lars Borin (ed.), Amsterdam, New York: Rodopi, p. 93-109.
- Simard M., Foster G., & Isabelle P. 1992. *Using cognates to align sentences in bilingual corpora*. In *Proceedings of TMI-92*, Montréal, Québec.
- Simard M. 2003. *Mémoires de Traduction sous-phrastiques*. Thèse de l'Université de Montréal.
- Smadja F. 1992. *How to compile a bilingual collocational lexicon automatically*. In *Proceedings of the AAAI-92 Workshop on Statistically-based NLP Techniques*.
- Smadja F., McKeown K.R. & Hatzivassiloglou V. 1996. *Translating Collocations for Bilingual Lexicons: A Statistical Approach*, Computational Linguistics. March, p. 1-38.
- Tiedemann J. 1993. *Combining clues for word alignment*. In *Proceedings of the 10<sup>th</sup> Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, p. 339-346, Budapest, Hungary, April 2003.
- Vergne J. 2005. *Une méthode indépendante des langues pour indexer les documents de l'Internet par extraction de termes de structure contrôlée*. In *Conférence Internationale sur le Document Electronique*. Beyrouth, Liban. May, 2005

## Appendix

source	ndoc	freq	cos	target
and	12	[336]	-	et
Member	10	[206]	-	membre~
Member State~	10	[201]	-	membre~
Member States	13	[143]	-	États membres
the	4	[392]	-	d~
of	5	[313]	-	de~

EEC	9	[118]	-	CEE
Directive	11	[95]	-	directive
or	8	[108]	-	ou
State~	5	[165]	-	membre~
Council	14	[47]	-	Conseil
and~	4	[121]	-	et
the~	3	[155]	-	l~
DIRECTIVE	15	[30]	-	DIRECTIVE
2	7	[60]	-	2
1971	15	[24]	-	1971
COUNCIL	13	[26]	-	CONSEIL
4	10	[33]	-	4
3	8	[41]	-	3
Annex	7	[42]	-	l'annexe
State	4	[71]	-	membre
provisions	8	[35]	-	dispositions
<b>Whereas</b>	10	[28]	-	<b>considérant que </b>
Member State	4	[63]	-	membre
2~	5	[43]	-	2
the Council	10	[19]	-	Conseil
approv~	3	[62]	-	modèle
requirements	6	[30]	-	prescriptions
mark~	6	[30]	-	marque~
5	8	[22]	-	5
Article	5	[35]	-	Article
Of~	2	[84]	-	de~
Economic	9	[18]	-	économique
control~	5	[31]	-	contrôl~
EEC pattern approval	4	[35]	-	CEE de modèle
verification	4	[34]	-	vérification
Council Directive	9	[15]	-	Conseil
EEC initial verification	5	[27]	-	vérification primitive CEE
<b>Having regard to the Opinion of the</b>	8	[16]	-	<b>vu l'avis </b>
THE	8	[16]	-	DES
-	4	[32]	-	-
3~	5	[24]	-	3~
II	6	[20]	-	II
;	5	[20]	-	;
Regulation	4	[24]	-	règlement
Commission	3	[32]	-	Commission
I	4	[24]	-	I
measures	3	[32]	-	mesures

Having regard to the	5	[19]	-	vu
pattern	2	[46]	-	modèle
6	7	[13]	-	6
in particular	6	[15]	-	notamment
meter~	3	[29]	-	compteur~
shall	4	[21]	-	à la
meters	3	[27]	-	compteurs
of Annex I	4	[18]	-	l'annexe I
5~	5	[13]	-	5
10	5	[13]	-	10
maximum	3	[21]	-	maximal~
15	5	[12]	-	15
item~	4	[15]	-	point~
Article~	3	[20]	-	Article
weight~	2	[28]	-	poids
1	2	[28]	-	1~
instruments	4	[14]	-	instruments
instrument~	2	[27]	-	instrument~
I~	3	[18]	-	I
separate	2	[26]	-	particulier~
method~	3	[17]	-	méthode~
may	5	[10]	-	peuvent
seed~	1	[48]	-	semences
se~	1	[46]	-	semences
methods	3	[15]	-	méthodes
seed	1	[45]	-	semences
pattern~	1	[44]	-	modèl~
by	4	[11]	-	par
Of Annex	3	[14]	-	point~
9	4	[10]	-	9
14	3	[13]	-	14
provision~	3	[13]	-	dispositions
Tests	2	[19]	-	essais
Committee	2	[18]	-	Comité
poultrymeat	1	[36]	-	de volaille
and methods of	4	[9]	-	de contrôle
Regulation (EEC) No	3	[12]	-	règlement (CEE) n°
OF	6	[6]	-	DES
control	4	[9]	-	de contrôle
6~	4	[9]	-	6
territory	2	[17]	-	territoire
liquid~	2	[17]	-	liquide~
particular	3	[11]	-	notamment
19~	3	[11]	-	19~



certain	3	[11]	- certain~
marks	3	[11]	- marques
mark	4	[8]	- la marque
directive	2	[16]	- directive particulière
trade	2	[16]	- échanges
pattern approval	1	[31]	- de modèle
pattern approval~	1	[31]	- de modèle
4~	5	[6]	- 4
12	3	[10]	- 12
approximat~	3	[10]	- rapprochement
certificate	3	[10]	- certificat
device~	3	[10]	- dispositif~
other	3	[10]	- autres que
for liquid~	2	[15]	- de liquides
measure~	3	[10]	- mesur~
measuring instruments	3	[10]	- instruments de mesurage
separate directive	2	[15]	- directive particulière
EEC pattern	1	[29]	- CEE de modèle
equipment	2	[14]	- dispositif~
EEC pattern approval~	1	[28]	- CEE de modèle
must	4	[7]	- doivent
fresh	1	[28]	- fraîche~
service	2	[14]	- service
1970	2	[14]	- 1970
of the	1	[28]	- d~
each	3	[9]	- chaque
July	3	[9]	- juillet
competent	2	[13]	- compétent~
this Directive	2	[13]	- la présente directive
relat~	3	[8]	- relatif~
26 July 1971	4	[6]	- du 26 juillet 1971
procedure	2	[12]	- procédure
on	1	[23]	- la commercialisation des
<b>fresh poultrymeat</b>	1	[23]	- <b>viandes fraîches de volaille</b>
EEC) No	2	[11]	- CEE
inform~	2	[11]	- inform~
the competent	2	[11]	- compétent~
into force	3	[7]	- en vigueur
symbol~	3	[7]	- marque~
the word~	1	[21]	- mot~
p~	1	[21]	- masse

subject to	3	[7]	- font l'objet
initial verification	1	[20]	- vérification primitive CEE
Directive~	1	[20]	- directiv~
two	4	[5]	- deux
material	1	[19]	- de multiplication
mass~	1	[19]	- à l'hectolitre
type-approv~	1	[19]	- CEE
than	2	[9]	- autres que
weight	1	[18]	- poids
amendments to	2	[9]	- les modifications