

# Projecting POS tags and syntactic dependencies from English and French to Polish in aligned corpora

Sylvia Ozdowska

ERSS - CNRS & Université Toulouse-le Mirail

Maison de la Recherche

5 allées Antonio Machado

F-31058 Toulouse Cedex 9

ozdowska@univ-tlse2.fr

## Abstract

This paper presents the first step to project POS tags and dependencies from English and French to Polish in aligned corpora. Both the English and French parts of the corpus are analysed with a POS tagger and a robust parser. The English/Polish bi-text and the French/Polish bi-text are then aligned at the word level with the GIZA++ package. The intersection of IBM-4 Viterbi alignments for both translation directions is used to project the annotations from English and French to Polish. The results show that the precision of direct projection vary according to the type of induced annotations as well as the source language. Moreover, the performances are likely to be improved by defining regular conversion rules among POS tags and dependencies.

## 1 Introduction

A clear imbalance may be observed between languages, such as English or French, for which a number of NLP tools as well as different linguistic resources exist (Leech, 1997) and those for which they are sparse or even absent, such as Polish. One possible option to enrich resource-poor languages consists in taking advantage of resource-rich/resource-poor language aligned corpora to induce linguistic information for the resource-poor side from the resource-rich side (Yarowski et al., 2001; Borin, 2002; Hwa et al., 2002). For Polish, this has been made possible on account of its accessing to the European Union (EU) which has resulted in the construction of a large multilingual

corpus of EU legislative texts and a growing interest for new Member States languages.

This paper presents a direct projection of various morpho-syntactic informations from English and French to Polish. First, a short survey of related works is made in order to motivate the issues addressed in this study. Then, the principle of annotation projection is explained and the framework of the experiment is described (corpus, POS tagging and parsing, word alignment). The results of applying the annotation projection principle from two different source languages are finally presented and discussed.

## 2 Background

Yarowski, Ngai and Wicentowski (2001) have used annotation projection from English in order to induce statistical NLP tools for instance for Chinese, Czech, Spanish and French. Different kinds of analysis were produced: POS tagging, noun phrase bracketing, named entity tagging and inflectional morphological analysis, and relied on to train statistical tools for each task. The authors report that training allows to overcome the problem of erroneous and incomplete word alignment thus improving the accuracy as compared to direct projection: 96% for core POS tags in French.

The study proposed by Hwa, Resnik, Weinberg and Kolak (2002) aims at quantifying the degree to which syntactic dependencies are preserved in English/Chinese aligned corpora. Syntactic relationships are projected to Chinese either directly or using elementary transformation rules which leads to 68% precision and about 66% recall.

Finally, Borin (2002) has tested the projection of major POS tags and associated grammatical informations (number, case, person, etc.) from

Swedish to German. 95% precision has been obtained for major POS tags<sup>1</sup> whereas associated grammatical informations have turned out not to be applicable across the studied languages. A rough comparison has been made between Swedish, German and additional languages (Polish, English and Finnish). It tends to show that it should be possible to derive indirect yet regular POS correspondences, at least across fairly similar languages.

The projection from French and English to Polish presented in this paper is basically a direct one. It concerns different linguistic informations: POS tags and associated grammatical information as well as syntactic dependencies. Regarding the works mentioned above, uneven results are expected depending on the type of annotations induced. This is the first point this study considers. The second one is to identify regularity in rendering some French or English POS tags or dependencies with some Polish ones. Finally, the idea is to test if the results vary significantly with respect to the source language used for the induction.

### 3 Projecting morpho-syntactic annotations

We take as the starting point of annotation projection the direct correspondence assumption as formulated in (Hwa et al., 2002): “for two sentences in parallel translation, the syntactic relationships in one language directly map the syntactic relationships in the other”, and extend it to POS tags as well. The general principle of annotation projection in aligned corpora may be explained as follows:

if two words  $w1$  and  $w2$  are translation equivalents within aligned sentences, the morpho-syntactic informations associated to  $w1$  are assigned to  $w2$

In this study, the projected annotations are POS tags, with gender and number subcategories for nouns and adjectives, on one hand, and syntactic dependencies on the other hand.

Let us take the example of *Commission* and *Komisja*, respectively  $w1_i$  and  $w2_m$ , two aligned words (figure 1). In accordance with the annotation projection principle, *Komisja* is first assigned the POS *N* (noun) as well as the information on its number, *sg* (singular), and gender *f* (feminine).

<sup>1</sup>Assessed on correct alignments.

Furthermore, the dependencies connecting  $w1_i$  to other words  $w1_j$  are examined. Foreach  $w1_j$ , if there is an alignment linking  $w1_j$  and  $w2_n$ , the dependency identified between  $w1_i$  and  $w2_j$  is projected to  $w2_m$  and  $w2_n$ . For example, the noun *Commission* ( $w1_i$ ) is syntactically connected to the verb *adopte* ( $w1_j$ ) through the subject relation and *adopte* is aligned to *przyjmuje* ( $w2_n$ ). Therefore, it is possible to induce a dependency relation, namely a subject one, between *Komisja* ( $w2_m$ ) and *przyjmuje* ( $w2_n$ )<sup>2</sup>.

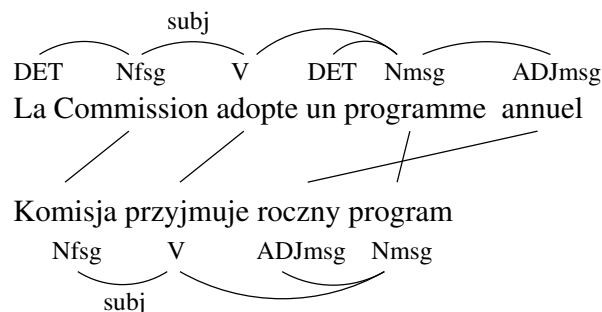


Figure 1: Projection of POS tags and dependencies from French to Polish

The induced dependencies are given the same label as the source dependencies that is to say that the noun *Komisja* and the verb *przyjmuje* are connected through the subject relation. Moreover, in this preliminary study, the projection is basically limited to cases where there is exactly one relation going from  $w1_i$  and  $w1_j$  on the one hand, and from  $w2_m$  and  $w2_n$  on the other hand. Thus, as shown in figure 2, the relation connecting *Komisja* and *przyjmuje* could not be induced from English since *Commission* and *adopt* are not linked directly but by means of the modal *shall*.

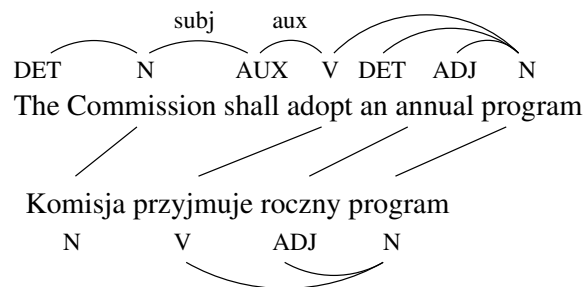


Figure 2: Projection of POS tags and dependencies from English to Polish

<sup>2</sup>The POS and the additional grammatical informations available are also projected from the verb *adopte* to *przyjmuje*.

The only exception concerns the complement and prepositional complement relations. Indeed, Polish is a highly inflected language which means that: 1) word order is less constrained than in French and English 2) syntactic relations between words are indicated by the case. This is the reason why, going back to figure 1, the projection from the nouns *programme* and *travail*, linked by the preposition *de*, results in the induction of a relation between the nouns *program* and *pracy*.

## 4 Experimental framework

### 4.1 Bi-texts

The countries wishing to join the EU have first to approve the *Acquis Communautaire*. The *Acquis communautaire* encompasses the core EU law, its resolutions and declarations as well as the common aims pursued since its creation in the 1950s. It comprises about 8,000 documents that have been translated and published by official institutions<sup>3</sup> thus ensuring a high quality of translation. Each language version of the *Acquis* is considered semantically equivalent to the others and legally binding. This collection of documents is made available on Europe’s website<sup>4</sup>.

The AC corpus is made of a part of the *Acquis* texts in 20 languages<sup>5</sup>, and in particular the languages of the new Member States<sup>6</sup>. It has been collected and aligned at the sentence level by the Language Technology team at the *Joint Research Centre* working for the European Commission<sup>7</sup> (Erjavec et al., 2005; Pouliquen and Steinberger, 2005). It is one of the largest parallel corpus regarding its size<sup>8</sup> and the number of different languages it covers. A portion of the English, French and Polish parts form the multilingual parallel corpus selected for this study. Table 1 gives the main features of each part of the corpus.

<sup>3</sup>Only European Community legislation printed in the paper edition of the *Official Journal of the European Union* is deemed authentic.

<sup>4</sup><http://europa.eu.int/eur-lex/lex>

<sup>5</sup>German, English, Danish, Spanish, Estonian, Finish, French, Greek, Hungarian, Italian, Latvian, Lithuanian, Maltese, Deutch, Polish, Portugese, Slovak, Slovene, Swedish and Czech.

<sup>6</sup>In 2004, the EU welcomed ten new Member States: Cyprus, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Czech Republic, Slovakia, Slovenia.

<sup>7</sup><http://www.jrc.cec.eu.int/langtech/index.html>

<sup>8</sup>The number of word forms goes from 6 up to 13 million according to the language. The parts corresponding to the languages of the new Member States range from 6 up to 10 million word forms as compared to 10 up to 13 million for

	English	French	Polish
<i>word forms</i>	562,458	809,036	764,684
<i>sentences</i>		52,432	

Table 1: AC – the English/French/Polish parallel corpus

## 4.2 Bi-text processing

### 4.2.1 POS tagging

Both the English and French parts of the corpus have been POS tagged and parsed. The POS tagging has been performed using the TreeTagger (Schmidt, 1994). Among the morpho-syntactic informations provided by the TreeTagger’s tagset, only the main distinctions are kept for further analysis: noun, verb, present participle, adjective, past participle, adverb, pronoun and conjunction (coordination and subordination). Nouns, adjectives and past participles are assigned data related to their number and gender and verbs are assigned information on voice, gender and form (infinitive or not), if available (table 2). The TreeTagger’s output is given as input to the parser after a post-processing stage which modifies the tokenization. Some multi-word units are conflated (for example complex prepositions such as *in accordance with*, *as well as* for English, *conformément à*, *sous forme de* for French, adverbs like *in particular*, *at least*, *en particulier*, *au moins*, or even verbs *prendre en considération*, *avoir recours*).

### 4.2.2 Parsing

Each post-processed POS-tagged corpus is analysed with a deep and robust dependency parser: SYNTAX (Fabre and Bourigault, 2001; Bourigault et al., forthcoming). For each sentence, SYNTAX identifies syntactic relations between words such as subject (SUBJ), object (OBJ), prepositional modifier (PMOD), prepositional complement (PCOMP), modifier (MOD), etc. Both versions of the parser are being developed according to the same procedure and architecture. The outputs are quite homogeneous in both languages since the dependencies are identified and represented in the same way, thus allowing the comparison of annotations induced from either French or English. Table 2 gives some examples of the basic relations taken into account as well as the tags assigned to the syntactically connected words. The

the languages of the “pre-enlargement” EU.

parts of speech are in upper case ( $N$  represents a noun,  $V$  a verb, etc.) and the grammatical information (number, gender) is in lower case ( $sg$  represents the singular,  $pl$  the plural,  $f$  the feminine and  $m$  the masculine).

(the) Regulation_Nsg	$\xleftarrow{\text{SUBJ}}$	establishes_Vsg		
(le) règlement_Nmsg	$\xleftarrow{\text{SUBJ}}$	détermine_Vsg		
covering_PPR	$\xrightarrow{\text{OBJ}}$	placing_PPR	$\xrightarrow{\text{PMOD}}$	on_PREP
$\xrightarrow{\text{PCOMP}}$ (the) market_Nsg				
(qui) régissent_Vpl	$\xrightarrow{\text{OBJ}}$	(la) mise_Nfsg	$\xrightarrow{\text{PMOD}}$	sur_PREP
$\xrightarrow{\text{PCOMP}}$ (le) marché_Nmsg				
further_ADJ	$\xleftarrow{\text{MOD}}$	calls_Npl		
appels_Nmpl	$\xrightarrow{\text{MOD}}$	supplémentaires_ADJpl		
(the) Member_Nsg	$\xleftarrow{\text{MOD}}$	States_Npl		
(les) États_Nmpl	$\xrightarrow{\text{MOD}}$	Membres_Nmpl		
(the debates) clearly_ADV	$\xleftarrow{\text{MOD}}$	illustrate_Vpl		
(les débats) montrent_Vpl	$\xrightarrow{\text{MOD}}$	clairement_ADV		
(placing on) the_DET	$\xleftarrow{\text{DET}}$	market_Nsg		
la_DET	$\xleftarrow{\text{DET}}$	mise (sur) le_DET	$\xleftarrow{\text{DET}}$	marché_Nmsg

Table 2: Syntactic dependencies identified with SYNTAX

### 4.2.3 Word alignment

The English/Polish parts of the corpus on the one hand, and the French/Polish parts on the other hand, have been aligned at the word level using the GIZA++ package<sup>9</sup> (Och and Ney, 2003). GIZA++ consists of a set of statistical translation models of different complexity, namely the IBM ones (Brown et al., 1993). For both corpora, the tokenization resulting from the post-processing stage prior to parsing was used in the alignment process for the English and Polish parts in order to keep the same segmentation especially to facilitate manual annotation for evaluation purposes. Moreover, each word being assigned a lemma at the POS tagging stage, the sentences given as input to GIZA++ were lemmatized, as lemmatization has proven to boost statistical word alignment performances. On the Polish side, a rough tokenization using blanks and punctuation was realised; no lemmatization was performed. The IBM-4 model has been trained on each bi-text in both translation directions and the intersection of Viterbi

<sup>9</sup>GIZA++ is available at <http://www.jfoch.com/GIZA++.html>.

alignments obtained has been used to project the morpho-syntactic annotations. In other words, our first goal was to test the extent to which the direct projection across English or French and Polish was accurate. Therefore, we relied only on one-to-one alignments, thus favouring precision to the detriment of recall for this preliminary study. Figure 3 shows an example of word alignment output. The intersection in both directions is represented with plain arrows; the dotted ones represent uni-directional alignments. It shows that the intersection results in an incomplete alignment which may differ depending on the pair of languages considered and the segmentation performed in each language<sup>10</sup>.

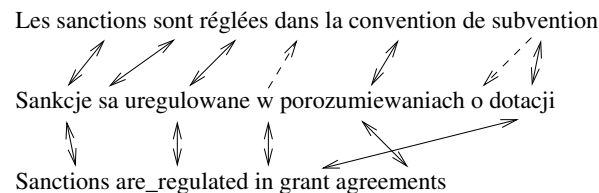


Figure 3: Intersection of IMB-4 model Viterbi alignments in both translation directions

## 5 Evaluation

### 5.1 Method

In order to evaluate the annotation projection, an *a posteriori* reference was constructed, which means that a sample of the output was selected randomly and annotated manually. There are some advantages to work with this kind of reference. First, it is less time-consuming than an *a priori* reference built independently from the output obtained. Second, it allows to skip the cases for which it is difficult to decide whether they are correct or not: syntactic analysis may be ambiguous and translation often makes it difficult to determine which source unit corresponds to which target one (Och and Ney, 2003). A better level of confidence may thus be ensured with an *a posteriori* reference in comparison with a human annotation task where a choice is to be made for each case. Finally, whatever strategy is adopted, there is always a part of subjectivity in human annotation. Thus, the results may vary from one person to another. The major drawback of an *a posteriori* reference is that it allows to assess only precision

<sup>10</sup>The underscore indicates token conflation.

and not recall since it precisely only contains data provided as output of the algorithm subjected to evaluation.

## 5.2 Parameters

The sample used in order to constitute the *a posteriori* reference is made of 50 French/Polish sentences and 50 English/Polish sentences. The same sentences in each language version were selected. Indeed, one of the goals of this study is to determine if the choice of the source language has an influence on annotation projection results. These 50 sentences correspond to 800 evaluated tags and 400 evaluated dependencies in the French/Polish bi-text, and 782 evaluated POS tags and 391 dependencies in the English/Polish bi-text.

Several parameters have been taken into account for each type of annotation projection by answering *yes* or *no* to the points listed below.

For POS tags:

- 1a. the projected POS is the correct one;
- 2a. the gender and number of nouns, adjectives and past participles are correct.

The gender parameter has been evaluated only for the projection from French to Polish as this information was not available in English.

For dependencies:

- 1b. there is a dependency relation between two given Polish words regardless of its label;
- 2b. the label of the dependency is correct.

Each time the answer to points 2a and 2b was *no*, the information about the correct annotation was added.

## 6 Results

### 6.1 Performances

Table 3 presents the number of projected POS tags and dependencies with respect to each source language. It gives the precision for each parameter, POS tag (1a), number and gender (2a), unlabeled dependencies (1b) and labeled dependencies (2b) assessed against the *a posteriori* reference.

It shows that the number of projected POS tags as well as syntactic relations is slightly lower when English is used as source language. A lower number of identified alignment links or dependencies may explain this difference. It also should be

		Fr/Pl	En/Pl
	<i>projected POS tags</i>	800	782
1a	<i>POS tags</i>	.87	.88
2a	<i>number</i>	.88	.91
2a	<i>gender</i>	.59	–
	<i>projected dependencies</i>	400	391
1b	<i>unlabeled dependencies</i>	.83	.82
2b	<i>labeled dependencies</i>	.62	.67

Table 3: Precision according to each evaluated parameter

noted that the evaluated projections are not necessarily the same in both corpora. As mentioned in section 5.1, the same sentences were chosen for evaluation. Nevertheless, since word alignment depends on the pair of languages involved, it has an impact on the projections obtained and the *a posteriori* reference built on their basis.

The precision rates vary according to the type of informations induced. No significant difference is observed whether the source language is French or English. The number subcategory achieves the highest score: 0.88 and 0.91 respectively for French/Polish and English/Polish. Dependencies rank second—0.83 and 0.82—but an important decrease in accuracy—about 20%—is observed when their labels are taken into account. Finally, for French, the gender category achieves the lowest score: 0.59. The main reasons for which annotation projection fails are investigated hereafter. The projection of the number and gender subcategories are not taken into account.

### 6.2 Result analysis

There are various reasons for the failure of the POS tags and dependencies’ projection: a) word alignment, b) lexical density, c) tokenization, d) POS tagging/parsing errors and e) insertion (for dependencies). In following examples, the word alignments are bold faced and in order to avoid confusion, the POS tags on the Polish side are the intended POS tags and not the induced POS tags.

a) The noun *countries* is aligned to *trzecich*<sup>11</sup> which is actually an adjective. On the other hand, *participation* and *udział* being aligned, the projected dependency is also erroneous.

*Participation*  $_N1$  of *third countries*  $_N2$   
*Udział*  $_N1$  państw *trzecich*  $_ADJ2$

<sup>11</sup>The correct alignment is *państw*.

b) *Under* is translated by the prepositional phrase *na podstawie* but is aligned only to *podstawie* which is a noun. Thus, the projected tag cannot be assigned just to *podstawie*, which is also the case with the PMOD dependency between *zawarte* and *podstawie*.

**concluded**\_PPA<sub>1</sub> **under**\_PREP<sub>2</sub> *the general framework*

**zawarte**\_PPA<sub>1</sub> **na podstawie**\_N<sub>2</sub> *ogólnych ram*

c) This case is similar to the previous but the difference in lexical density is partly caused by the conflation of *in accordance with*, which corresponds to the prepositional phrase *zgodnie z*, at the post-processing stage of the POS tagging.

*They must be constituted*  
**in accordance with**\_PREP<sub>1</sub> **the law**\_N<sub>2</sub>

*Muszą być ustanowione* **zgodnie**\_ADV<sub>1</sub> *z*  
**prawem**\_N<sub>2</sub>

d) The following example shows an error in PCOMP attachment resulting in an error in dependency projection: *with* is linked to *pursue* instead of *activities* and the same relation is assigned to *o* and *zajmować*.

*They must* **pursue**\_V<sub>1</sub> *activities* **with**\_PREP<sub>2</sub> *a European dimension*

*Muszą* **zajmować**\_V<sub>1</sub> *się działalnością* **o**\_PREP<sub>2</sub> *europejskim wymiarze*

e) On the Polish side, the inserted noun *postanowień* governs *traktatu*. Thus, the PCOMP dependency does not link *dla* and *traktatu* but *dla* and *postanowień*.

*Without prejudice* **for**\_PREP<sub>1</sub> *the* **Treaty**\_N<sub>2</sub>

*Bez uszczerbku* **dla**\_PREP<sub>1</sub> *postanowień*  
**Traktatu**\_N<sub>2</sub>

Considering the precision figures, in particular those accounting for the projection of dependencies which decrease significantly when labels are considered, we tried to determine if there are indirect yet regular French/Polish and English/Polish correspondences. By indirect correspondence we mean that a given source POS tag or dependency is usually rendered by a given Polish POS tag or dependency. The correspondences are calculated provided there is no error prior to projection (word alignment, tagging or parsing).

Table 4 shows the direct and indirect correspon-

dences among the POS tags which occur in the reference set. We can see that there is a direct correspondence among POS tags in 92% and 93% of the cases respectively for French/Polish and English/Polish projection. Moreover, the indirect correspondences, for example noun/adjective or verb/noun, are similar for both source languages. The following examples show occurrences of noun/adjective and verb/noun correspondences.

*the exercise of* **implementing**\_N *powers*

*l'exercice des compétences d'exécution*\_N

*wykonywania uprawnień wykonawczych*\_ADJ

*measures planned to* **ensure**\_V *dissemination*

*mesures prévues pour* **assurer**\_V *la diffusion*  
*środków zaplanowane dla* **zapewnienia**\_N  
*rozpowszechnienia*

Some indirect correspondences are more probable than others that seem unexpected. Most of the time the latter come from the differences in tokenization mentioned above.

Fr POS	PI POS	c
N_359	N_349; ADJ_6; PPA_3; V_1	.97
ADJ_74	ADJ_69; N_3; V_1; DET_1	.93
V_68	V_55; N_13	.80
PPA_67	PPA_59; V_6; ADJ_1; N_1	.88
PREP_35	PREP_24; N_7; DET_2; V_1; PPR_1	.68
others_61	same_56	.91
<b>664</b>	<b>612</b>	<b>.92</b>
En POS	PI POS	c
N_374	N_364; ADJ_9; PPA_1	.97
PREP_64	PREP_53; N_7; DET_4	.83
V_51	V_35; PPA_10; N_6	.69
ADJ_46	ADJ_42; N_2; V_1; DET_1	.91
DET_36	DET_33; N_2	.91
others_73	same_70	.95
<b>644</b>	<b>597</b>	<b>.93</b>

Table 4: French/Polish and English/Polish POS tag correspondences

Table 5 summarizes direct and indirect correspondences among syntactic dependency relations. It can be seen that direct correspondence rates for dependencies are lower than direct correspondences for POS tags: 78% when the source language is French source and 82% when it is

English. Moreover, the difference according to the source language—5% in favour of English—is more important than for POS tags—1% in favour of English. It is mainly due to the PMOD and PCOMP relations: the first connects a preposition to its governor and the second connects the dependent to a preposition. Since Polish is an inflected language, the connections between words are indicated through cases. In particular, it results in a noun not being necessarily linked to another noun by a preposition. This is also the case for English, as far as compounds are concerned, while in French a preposition is almost always required to form noun phrases. This is one of the reasons why the direct correspondence rate between English and Polish is higher than between French and Polish. The following example shows a direct MOD/MOD correspondence for the English/Polish pair and an indirect PMOD\_PCOMP/MOD correspondence for the French/Polish one.

*purity*  $\xrightarrow{\text{MOD}}$  *criteria*\_N *substances*\_N listed  
*les critères*\_N  $\xrightarrow{\text{PMOD\_de\_PCOMP}}$  *pureté* des  
*substances énumérés*  
*kryteria*\_N  $\xrightarrow{\text{MOD}}$  *czystości*\_N dla *substancji*  
*wymienionych*

Fr DEP	Pl DEP	c
PMOD_111	PMOD_56; MOD_51; OBJ_4	.50
MOD_106	MOD_106	1
PCOMP_35	PCOMP_25; MOD_7; OBJ_2; PMOD_1;	.71
OBJ_23	OBJ_16; MOD_5; PMOD_2	.69
SUJ_19	SUJ_18; OBJ_1	.94
others_38	same_38	1
<b>332</b>	<b>259</b>	<b>.78</b>
En DEP	Pl DEP	c
MOD_95	MOD_90; PMOD_5	.94
PMOD_93	PMOD_59; MOD_26; PCOMP_4; OBJ_3; SUBJ_1	.63
PCOMP_64	PCOMP_49; MOD_8; PMOD_7	.76
DET_29	DET_29	1
OBJ_23	OBJ_22; PMOD_1	.95
others_18	same_18	1
<b>322</b>	<b>267</b>	<b>.83</b>

Table 5: French/Polish and English/Polish syntactic correspondences

## 7 Discussion

The results of the projection of POS tags and dependencies concur with those reported in the related works presented in section 2. First, concerning the number and gender subcategories, Borin (2002) has found that the former is applicable across languages whereas the latter is less relevant, at least for the German/Swedish language pair. As seen in section 3, the projection of the number subcategory offers the highest score and the projection of the gender the lowest—0.59. It was to be expected that gender would perform the worst considering its arbitrary nature at least in French and Polish. Indeed, there are three genders in Polish, masculine, feminine and neutral, as well as in English, and two in French. Thus, not only the number of genders across French and Polish is different but they are not distributed in the same way in both languages. The information on gender was not available for English, gender being assigned according to the human/non-human feature.

Considering POS tags, the level of direct correspondence is the highest one when compared to the number and gender subcategories as well as to dependencies. The precision performed is however lower with respect to the figures obtained by Borin (2002) on the one hand, and Yarowski et al.’s (2001) on the other hand. In Borin’s study, precision was assessed provided the word alignments used to project POS tags were correct. In this study, precision has been evaluated regardless of possible errors prior to projection. When these errors are discarded, the precision rates are similar. In Yarowski et al.’s work (2001), the evaluation did not concern annotation projection but an induced tagger trained on 500K occurrences of automatically derived POS tag projections. Indeed, the authors claim that direct annotation projection is quite noisy. This study shows that such a simple approach can perform fairly well as far as precision is concerned. The results are likely to be improved by implementing basic POS tag conversion rules as suggested in (Borin, 2002).

For the projection of dependencies, defining such conversion rules seems necessary as suggested by the significant difference in precision when the projection of unlabeled and labeled dependencies are compared. Polish does not proceed in the same way to encode syntactic functions as compared to English or French. Nevertheless, some of the syntactic divergences observed seem regu-

lar enough to be used to derive indirect correspondences. Hwa et al. (2002) have noticed that applying elementary linguistic transformations considerably increases precision and recall when projecting syntactic relations, at least for the English/Chinese language pair. The present study suggests that this kind of approach is promising for the English/Polish and French/Polish pairs as well.

The exceptional status of the corpus certainly influences the quality of the results. Legislative texts of the EU in their different language versions are legally binding. Thus, they have to be as close as possible semantically and this constraint may favour the direct correspondences observed.

## 8 Conclusion

We have presented a simple yet promising method based on aligned corpora to induce linguistic annotations in Polish texts. POS tags and dependencies are directly projected to the Polish part of the corpus from the automatically annotated English or French part. As far as precision is concerned, the direct projection is fairly efficient for POS tags but appears to be too restrictive for dependencies. Nevertheless, the results are encouraging since they are likely to be improved by applying indirect correspondence rules. They validate the idea of the existence of direct or indirect yet regular correspondences on the English/Polish and French/Polish language pairs which has already been tested with some syntax-based alignment techniques (Ozdowska, 2004; Ozdowska and Claveau, 2005). The next step will consist in exploiting the indirect correspondences and the multiple sources of information provided by two different source languages. Moreover, using IBM-4 word alignments in one direction instead of the intersection will be considered.

This work mainly focusses on precision thus lacking information on recall. Larger scale evaluations would be necessary to validate the approach, particularly evaluations that could measure recall, since the amount of evaluation data used is this study could be considered too limited.

## References

Lars Borin. 2002. Alignment and tagging. In Lars Borin, editor, *Parallel corpora, parallel worlds: selected papers from a symposium on parallel and*

*comparable corpora at Uppsala University*, pages 207–217. Rodopi, Amsterdam/New York.

Didier Bourigault, Cécile Fabre, Cécile Frérot, Marie-Paule Jacques, and Sylwia Ozdowska. forthcoming. Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique. *T.A.L (Traitement Automatique des Langues)*.

Peter F. Brown, Stephen. A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Tomaž Erjavec, Camelia Ignat, Bruno Pouliquen, and Ralf Steinberger. 2005. Massive multilingual corpus compilation: Acquis communautaire and TOTALE. In *2nd Language and Technology Conference*.

Cécile Fabre and Didier Bourigault. 2001. Linguistic clues for corpus-based acquisition of lexical dependencies. In *Corpus Linguistic Conference*.

Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluating translational correspondence using annotation projection. In *40th Annual Conference of the Association for Computational Linguistics*.

Geoffrey Leech. 1997. Introducing corpus annotation. In Roger Garside, Geoffrey Leech, and Anthony McEnery, editors, *Corpus Annotation. Linguistic Information from Computer Text corpora*, pages 1–18. Longman, London/New York.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 1(29):19–51.

Sylwia Ozdowska and Vincent Claveau. 2005. Aligement de mots par apprentissage de règles de propagation syntaxique en corpus de taille restreinte. In *Conférence sur le Traitement Automatique des Langues Naturelles*, pages 243–252.

Sylwia Ozdowska. 2004. Identifying correspondences between words: an approach based on a bilingual syntactic analysis of French/English parallel corpora. In *Multilingual Linguistic Resources Workshop of COLING'04*.

Bruno Pouliquen and Ralf Steinberger. 2005. The acquis communautaire corpus. In *JRC Enlargement and Integration Workshop*.

Helmut Schmidt. 1994. Probabilistic part-of-speech tagging using decision trees. In *1st International Conference on New Methods in Natural Language Processing*.

David Yarowski, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *1st Human Language Technology Conference*.