Automatic Named Entity Pre-Annotation for Out-of-Domain Human Annotation

Sophie Rosset$^\alpha$, **Cyril Grouin**$^\alpha$, **Thomas Lavergne**$^{\alpha,\beta}$, **Mohamed Ben Jannet**$^{\alpha,\beta,\gamma,\delta}$
Jérémy Leixa$^\epsilon$, **Olivier Galibert**$^\gamma$, **Pierre Zweigenbaum**$^\alpha$.
$^\alpha$LIMSI–CNRS $^\beta$Université Paris-Sud $^\gamma$LNE
$^\delta$LPP, Université Sorbonne Nouvelle $^\epsilon$ELDA
{rosset,grouin,lavergne,ben-jannet,pz}@limsi.fr
leixa@elda.org, olivier.galibert@lne.fr

Abstract

Automatic pre-annotation is often used to improve human annotation speed and accuracy. We address here out-of-domain named entity annotation, and examine whether automatic pre-annotation is still beneficial in this setting. Our study design includes two different corpora, three pre-annotation schemes linked to two annotation levels, both expert and novice annotators, a questionnaire-based subjective assessment and a corpus-based quantitative assessment. We observe that pre-annotation helps in all cases, both for speed and for accuracy, and that the subjective assessment of the annotators does not always match the actual benefits measured in the annotation outcome.

1 Introduction

Human corpus annotation is a difficult, time-consuming, and hence costly process. This motivates research into methods which reduce this cost (Leech, 1997). One such method consists of automatically pre-annotating the corpus (Marcus et al., 1993; Dandapat et al., 2009) using an existing system, e.g., a POS tagger, syntactic parser, named entity recognizer, according to the task for which the annotations aim to provide a gold standard. The pre-annotations are then corrected by the human annotators. The underlying hypothesis is that this should reduce annotation time while possibly at the same time increasing annotation completeness and consistency.

We study here corpus pre-annotation in a specific setting, *out-of-domain named entity annotation*, in which we examine specific questions that we present below. We produced corpora and annotation guidelines for named entities which are both hierarchical and compositional (Grouin et al., 2011),[1] and which we used in contrastive studies of news texts in French (Rosset et al., 2012). We want to rely on the same named entity definitions for studies on two types of data we did not cover: parliament debates (*Europarl* corpus) and regional, contemporary written news (*L'Est Républicain*), both in French. To help the annotation process we could reuse our system (Dinarelli and Rosset, 2011), but needed first to examine whether a system trained on one type of text (our first Broadcast News data) could be used to produce a useful pre-annotation for different types of text (our two corpora).

We therefore set up the present study in which we aim to answer the following questions linked to this point and to related annotation issues:

- can a system trained on data from one specific domain be useful on data from another domain in a pre-annotation task?
- does this pre-annotation help human annotators or bias them?
- what importance can we give to the annotators' subjective assessment of the usefulness of the pre-annotation?
- can we observe differences in the use of pre-annotation depending on the level of expertise of human annotators?

Moreover, as the aforementioned annotation scheme is based on two annotation levels (*entities* and *components*), we want to answer these questions taking into account these two levels.

We first examine related work on pre-annotation (Section 2), then present our corpora and annotation task (Section 3). We describe and discuss experiments in Section 4, and make subjective and

[1] Corpora, guidelines and tools are available through ELRA under references ELRA-S0349 and ELRA-W0073.

Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pages 168–177,
Sofia, Bulgaria, August 8-9, 2013. ©2013 Association for Computational Linguistics

quantitative observations in Sections 5 and 6. Finally, we conclude and present some perspectives in Section 7.

2 Related Work

Facilitating human annotations has been the topic of a large amount of research. Two different approaches can be distinguished: active learning (Ringger et al., 2007; Settles et al., 2008) and pre-annotation (Marcus et al., 1993; Dandapat et al., 2009). Our work falls into the latter type.

Pre-annotation can be used in several ways. The first is to provide annotations to be corrected by human annotators (Fort and Sagot, 2010). A variant consists of merging multiple automatic annotations before having them corrected by human curators to produce a gold-standard (Rebholz-Schuhmann et al., 2011). The second type consists of providing clues to help human annotators perform the annotation task (Mihaila et al., 2013).

This work addresses the first type, a single-system pre-annotation with human correction. An objective is to examine whether a system trained on one type of text can be useful to pre-annotate texts of a different type. Most previous studies have been performed on well-behaved tasks such as part-of-speech tagging on in-domain data, i.e., the model used for pre-annotating the target data had been trained on similar data. For instance, Fort and Sagot (2010) provide a precise evaluation of the usefulness of pre-annotation and compare the impact of different quality levels in POS taggers on the Penn TreeBank corpus. They first trained different models on the training part of the corpus and applied them to the test corpus. The pre-annotated test corpus was then corrected by humans. They reported gains in accuracy and inter-annotator agreement. The study focused on the minimal quality (accuracy threshold) of automatic annotation that would prove useful for human annotation. They reported a gain for human annotation when accuracy ranged from 66.5% to 81.6%. On the contrary, for a semantic-frame annotation task, Rehbein et al. (2009) observed no significant gain in quality and speed of annotation even when using a state-of-the-art system.

Generally speaking, annotators find the pre-annotation stage useful (Rehbein et al., 2009; South et al., 2011; Huang et al., 2011). Annotation managers consider that a bias may occur depending on how much human annotators trust the pre-annotation (Rehbein et al., 2009; Fort and Sagot, 2010; South et al., 2011). In their frame-semantic argument structure annotation, Rehbein et al. (2009) addressed a specific question considering a two-level annotation scheme: is the pre-annotation of frame assignment (low-level annotation) useful for annotating semantic roles (high-level annotation)? Although for the low-level annotation task they observed a significant difference in quality of final annotation, for the high-level task they found no difference.

Most of these studies used a pre-annotation system trained on the same kind of data as those which were to be annotated manually. Nevertheless some system-oriented studies have focused on the results obtained by systems trained on one type of corpus and applied to another type of corpus, e.g., for a Latin POS tagger (Poudat and Longrée, 2009; Skjærholt, 2011) or for a CoNLL named entity tagger for German (Faruqui and Padó, 2010) for which the authors noticed noticed a reduction of the F-measure when going from in-domain (newswire data, F=0.782 for their best system) to out-of-domain (Europarl data, F=0.656).

One of our objectives is then to examine whether a system trained on one type of text can be useful to pre-annotate texts of a different type. We set up experiments to study precisely the possible induced bias and whether the level of experience of the annotators would make a difference in such a context. In this study, we used two different kinds of corpora, which were both different from the corpus used to train the pre-annotation system.

3 Task and corpus description

3.1 Task

In this work, we used the structured named entity definition we proposed in a previous study (Grouin et al., 2011): entities are both hierarchical (types have subtypes) and compositional (types and components are included in entities) as in Figure 1.

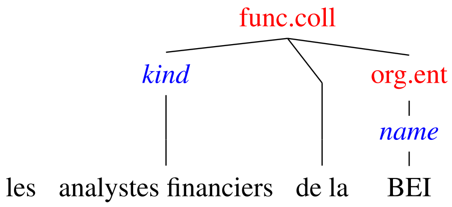

Figure 1: Multi-level annotation of entity subtypes (red tags) and components (*blue* tags): *the financial analysts of the EBI*

This taxonomy of entity types is composed of 7 types (*person, location, organization, amount, time, production* and *function*) and 32 sub-types (individual person *pers.ind* vs. group of persons *pers.coll*; administrative organization *org.adm* vs. services *org.ent*; etc.). Types and subtypes constitute the first level of annotation.

Within these categories, components are second-level elements (*kind, name, first.name*, etc.), and can never be used outside the scope of a type or subtype element.

3.2 Corpora

Two French corpora were sampled from larger ones:

Europarl: Prepared speech (*Parliament Debates—Europarl*): 15,306 word extract;

Press: Local, contemporary written news (*L'Est Républicain*): 11,146 word extract.

These corpora were automatically annotated using the system described in (Dinarelli and Rosset, 2011). This system relies on a Conditional Random Field (CRF) model for the detection of components and on a probabilistic context-free grammar (PCFG) model for types and sub-types. These models have been trained on Broadcast News data. This system achieved a Slot Error Rate (Makhoul et al., 1999) of 37.0% on Broadcast conversation and 29.7% on Broadcast news, and ranked first in the Quaero evaluation campaign (Galibert et al., 2011).

4 Experiments

In this section we present the protocol we designed to study the usefulness of pre-annotation under different conditions, and its overall results.

4.1 Protocol

We defined the following protocol, similar to the one used in Rehbein et al. (2009).

Corpora. Four versions of our two corpora were prepared: (i) raw text, (ii) pre-annotation of types, (iii) pre-annotation of components, and (iv) full pre-annotation of both types and components. Each of these four versions was split into four quarters.

Annotators. Eight human annotators were involved in this task. Among them, four are considered as expert annotators (they annotated corpora in the previous years) while the four remaining ones are novice annotators (this was the first time they annotated such corpora; they were given training sessions before starting actual annotation). We defined four pairs of annotators, where each pair was composed of an expert and a novice annotator.

Quarter allocation. We allocated each corpus quarter in such a way that each pair of annotators processed, in each corpus, material from each one of the four pre-annotated versions (see Table 3). The same allocation was made in both corpora.

4.2 Results

For each corpus part, a reference was built based on a majority vote by confronting all annotations. The resulting reference corpus is presented in Table 1.

Corpus		# comp.	# types	# entities	# words
Press	Q1	481	310	791	3047
	Q2	367	246	673	2628
	Q3	495	327	822	2971
	Q4	413	282	695	2600
Europarl	Q1	362	259	621	3926
	Q2	309	221	530	3809
	Q3	378	247	625	3604
	Q4	413	299	712	3967

Table 1: General description of the reference annotations: number of components, types, entities (the sum of components and types), and words

Table 2 presents the performance of the automatic pre-annotation system against the reference corpus. We used the well known F-measure and in addition the Slot Error Rate as it allows to weight different error classes (deletions, insertions, type or frontier errors). Fort and Sagot (2010) reported a gain in human annotation when pre-annotation accuracy ranged from 66.5% to 81.6%. Given their results we can hope for a gain in both accuracy and annotation time when using pre-annotation.

Table 3 presents all results obtained by each annotators given each pre-annotation condition (raw, components, types and full) in terms of precision, recall and F-measure.

Corpus	#	Raw text			Components			Types			Full		
		R	P	F	R	P	F	R	P	F	R	P	F
Press	Q1	0.874	0.777	0.823	0.876	0.741	0.803	0.824	0.870	**0.846**	0.852	0.800	0.825
		0.810	0.766	0.788	0.815	0.777	0.796	0.645	0.724	0.683	0.844	0.785	**0.813**
	Q2	0.765	0.796	0.780	0.870	0.773	0.819	0.822	0.801	0.812	0.917	0.773	**0.839**
		0.558	0.654	0.602	0.826	0.775	**0.800**	0.815	0.777	0.795	0.816	0.752	0.783
	Q3	0.835	0.715	0.771	0.888	0.809	0.847	0.884	0.796	0.837	0.887	0.859	**0.873**
		0.792	0.689	0.736	0.904	0.780	**0.837**	0.876	0.771	0.820	0.780	0.827	0.803
	Q4	0.802	0.757	0.779	0.845	0.876	0.860	0.900	0.702	0.789	0.914	0.840	**0.876**
		0.794	0.727	0.759	0.696	0.715	0.705	0.812	0.701	0.752	0.802	0.757	**0.779**
Europarl	Q1	0.809	0.728	0.766	0.800	0.568	0.665	0.776	0.862	**0.817**	0.754	0.720	0.736
		0.754	0.720	**0.736**	0.720	0.609	0.660	0.687	0.607	0.644	0.736	0.638	0.683
	Q2	0.776	0.792	**0.784**	0.782	0.617	0.690	0.797	0.645	0.713	0.821	0.526	0.641
		0.563	0.498	0.529	0.802	0.619	**0.699**	0.698	0.553	0.617	0.769	0.566	0.652
	Q3	0.747	0.459	0.569	0.749	0.624	0.681	0.805	0.800	**0.803**	0.735	0.744	0.739
		0.732	0.598	0.658	0.736	0.717	0.726	0.822	0.738	**0.777**	0.808	0.734	0.769
	Q4	0.742	0.624	0.678	0.874	0.760	**0.813**	0.732	0.480	0.580	0.743	0.608	0.669
		0.721	0.566	0.634	0.695	0.652	**0.672**	0.707	0.600	0.649	0.738	0.603	0.664

Table 3: Overall recall, precision and F-measure for each pair of annotators *(blue: pair #1, ocre: pair #2, green: pair #3, white: pair #4)* on each corpus quarter *(Q1, Q2, Q3, Q4)*, depending on the kind of pre-annotation *(raw text, only components, only types, full pre-annotation)*. Expert annotator is on the upper line of each quarter, novice annotator is on the lower line. Boldface indicates the best F-measure for each novice and expert annotator among all pre-annotation tasks in a given corpus quarter

Corpus		Components		Types		Full	
		F	SER	F	SER	F	SER
Press	Q1	72.4	37.9	63.5	46.3	68.9	41.0
	Q2	77.2	32.2	66.8	43.5	73.1	36.6
	Q3	76.1	34.1	68.3	41.7	73.1	36.9
	Q4	76.1	33.3	63.3	45.7	71.0	38.2
Europarl	Q1	61.9	49.9	57.5	55.4	60.1	52.2
	Q2	61.2	51.3	54.6	54.3	58.5	52.5
	Q3	61.6	50.1	53.3	55.7	58.2	52.2
	Q4	57.1	57.0	48.1	59.7	53.3	58.1
Broad.		88.3	29.1	73.1	39.1	73.2	33.1

Table 2: F-measure and Slot Error Rate achieved by the automatic system on each kind of annotation and on in-domain broadcast data

We also computed inter-annotator agreement (IAA) for each corpus considering two groups of annotators, *experts* and *novices*. We consider that the inter-annotator agreement is somewhere between the F-measure and the standard IAA considering as *markables* all the units annotated by at least one of the annotators (Grouin et al., 2011). We computed Scott's Pi (Scott, 1955), and Cohen's Kappa (Cohen, 1960). The former considers one model for all annotators while the latter considers one model per annotator. In our case, these two values are almost the same, which means that the proportions and kinds of annotations are very similar across experts and novices. Figure 2 shows the IAA (Cohen's Kappa and F-measure) obtained on the two corpora given the four pre-annotation conditions (no pre-annotation, components, types, and full pre-annotation). As we can see, IAA is systematically higher for the *Press* corpus than for the *Europarl* corpus, which can be linked to the higher performance of the automatic pre-annotation system on this corpus. We also can see that pre-annotation always improves agreement and that full pre-annotation yields the best result. We observe that, as expected, pre-annotation leads human annotators to obtain higher consistency.

5 Subjective assessment

An important piece of information in any annotation campaign is the feelings of the annotators about the task. This can give interesting clues about the expected quality of their work and on the usefulness of the pre-annotation step. We asked the annotators a few questions concerning several features of this project, such as the annotation

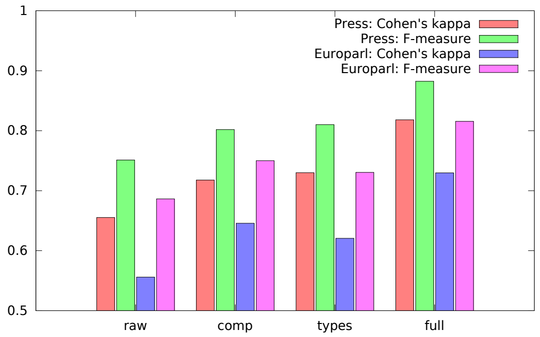

Figure 2: Cohen's Kappa (red and blue) and F-measure (green and pink) measuring agreement of experts and novices on Press and Europarl corpora in four pre-annotation conditions. Each measure compares the concatenated annotations of the four experts with the four novices.

manual, or how they assessed the benefits of pre-annotation in the different corpora (Section 5.1). Another important point is the experience of the annotators, which we also examine in the light of theirs answers to the questionnaire (Section 5.2).

5.1 Questionnaire

The questionnaire submitted to the annotators contained 4 questions, dealing with their feedback on the annotation process:

1. According to you, which level of pre-annotation has been the most helpful during the annotation process? Types, components, or both?

2. To what extent would you say that pre-annotation helped you in terms of precision and speed? Did it produce many errors you had to correct?

3. If you had to choose between the Europarl corpus and the Press corpus, could you say that one has been easier to annotate than the other?

4. Concerning the annotation manual, are there topics that you would like to change, or correct? In the same way, which named entities caused you the most difficulties to deal with?

All 8 annotators answered these questions. We summarize below what we found in their answers.

5.1.1 Level of pre-annotation

Most of the annotators preferred the corpora that were pre-annotated with types only. The reason, for the most part, is that a pre-annotation of types allows the annotator to work faster on their files, because guessing the components from the types is easier than guessing types from components.[2] Indeed, the different types of entities defined in the manual always imply the same components, be they specific (to one entity type) or transverse (common to several entity types). On the contrary, a transverse component, such as <kind>, can be part of any type of named entity. The other reason for this choice of pre-annotation concerns the readability brought to the corpora. An annotation with types only is easier to read than an annotation with components, and less exhausting after many hours of work on the texts.

5.1.2 Gain in precision and speed

What motivated the answers to the second question mainly concerns the accuracy of the different pre-annotation methods. While all of them presented errors that needed to be corrected, the pre-annotation of types was the one that they felt presented the smaller number of errors. Thus, annotators spent less time reviewing the corpora in search of errors, compared to the other pre-annotated corpora (with components, and with both types and components), where more errors had to be spotted and corrected. This search for incorrect pre-annotations impacted the time spent on each corpus. Indeed, most annotators declared that pre-annotation with types was quicker to deal with than other pre-annotation schemes.

5.1.3 Corpus differences

About one half of the annotators agreed that the Europarl corpus had been more difficult to annotate. Despite obvious differences in register, sentence structure and vocabulary between the two corpora, Europarl seemed more redundant and complex than the other corpus. For instance, one of the annotators declared:

> *The Europarl corpus is more difficult to annotate in the sense that the existing types and components do not always match the realities found in the corpus, either because their definitions*

[2] This feeling is supported by results about ambiguity presented in Fort et al. (2012).

> *cannot apply exactly, or because the required types and components are missing (mainly for frequencies: "five times per year").*

The other half of the annotators did not feel any specific difficulties in annotating one corpus or the other. According to them, both corpora are the same in terms of register and sentence structure.

5.1.4 Improvements in guidelines

All of the annotators were unanimous in thinking that two points need to be modified in the manual. First of all, the distinction between the <org.adm> and <org.ent> subtypes is too difficult to apprehend, above all in the Europarl corpus where these entities are too ambiguous to be annotated correctly. Secondly, the distinction between the <pers> and <func> types has also been difficult to deal with. The other remarks about potential changes mainly concerned the introduction of explicit rules for frequencies, which are recurrent in the Europarl corpus.

5.2 Experience

As mentioned earlier in Section 4.1, we will now see if the differences in experience between annotators impacted their difficulty in annotating the corpora. First of all, when we look at the answers given to question 3, we notice that both novice and expert annotators consider the Europarl corpus the most difficult to annotate. Most of their answers deal with the redundancy and the formal register of the data. Moreover, as everyone answered in question 4, both <func> and <org> entities have to be modified to be easier to understand and to use. This unanimous opinion about what needs to be reviewed in the manual allows us to think that the annotators' level of experience has a low impact on their apprehension of the corpora, both *Europarl* and *Press*. To confirm this, we can look at the answers given to questions 1 and 2, as indicated in the previous paragraph. As has been explained, every annotator correctly pointed at the many errors found in pre-annotation, regardless of their experience. Besides, the assessment of the benefits of pre-annotation is the same for almost everyone, regardless of their experience too: both novice and expert annotators agree that pre-annotation with type adds efficiency and speed to annotation.

To conclude, according to our observations based on the questionnaire, we cannot assert that there has been a difference between novice and expert annotators. Both groups agreed on the same difficulties, pointed at the same errors, and criticized the same entities, saying that their definitions needed to be clarified.

6 Quantitative observations

In this section we provide results of quantitative observations in order to support, or not, the annotators' subjective assessment.

6.1 Corpus statistics

The annotators reported different feelings depending on the corpora. Some of them reported that the Europarl corpus was more difficult to annotate, with more complex sentence structures, or usage of fewer proper nouns.

To explore these differences, we computed some statistics over the two original, un-annotated corpora (which are much larger than the samples annotated in this experiment) as well as over the original broadcast news corpus used to train the pre-annotation system. Each of these corpora contains several million words.

Table 4 reports simple statistics about sentences in the three corpora. Based on these statistics, while the *Europarl* (Euro) corpus is very similar to the original *Broadcast News* (BN), the *Press* corpus shows differences: sentences are 20% shorter, with fewer but larger chunks, confirming the impression of simpler, less convoluted sentences.

	BN	Press	Euro
Mean sentence length	30.2	23.9	29.7
Mean chunk count	10.9	6.7	10.4
Mean chunk length	2.7	3.6	2.8

Table 4: Sentence summary of the three corpora

Looking more closely at the contents of these sentences, Figure 3 summarizes the proportions of grammatical word classes. The sentiment of extensive naming of entities in the *Press* corpus is confirmed by the four times higher rate of proper nouns. On the other hand, entities are more often referred to using nouns with an optional adjective in the *Europarl* corpus, leading to a more frequent usage of the latter.

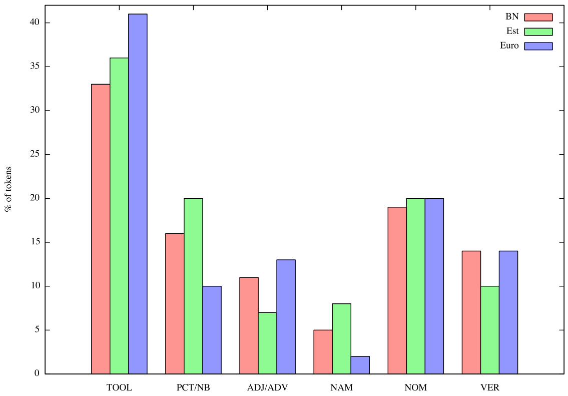

Figure 3: Frequency of word classes in the three corpora (BN = Broadcast News, Est = Press, Euro = Europarl). TOOL = grammatical words, PCT/NB = punctuation and numbers, ADJ/ADV = adjectives and adverbs, NAM = proper name, NOM = noun, VER = verb.

6.2 Influence of pre-annotation on the behaviour of annotators

As already mentioned, it is often reported that a bias may occur depending on human confidence in the pre-annotation (Fort and Sagot, 2010; Rehbein et al., 2009; South et al., 2011). An important unknown is always the influence of pre-annotation on the behaviour of annotators, and at which point pre-annotation induces more errors than it helps. This may obviously depend on pre-annotation quality. Table 5 summarizes the error rates of the automatic annotator in the studied data (*Press* + *Europarl*) and in comparison to in-domain data. *Insertions* (Ins) are extra annotations, *deletions* (Del) missing annotations, and *substitutions* (Subs) are annotations that are incorrect in type, boundaries, or both. We can see that going out-of-domain increased deletions, probably through a lack of knowledge of domain vocabulary. But it did not influence the other error rates significantly. It is also noticeable that deletion is the type of error most produced by the system, with every third entity missed. Automatic, full pre-annotation of *Press* + *Europarl* obtains a precision of 0.79 and a recall of 0.56.

Domain	Pre-annotation	Ins	Del	Subs
Out	Components	4.4%	33.6%	7.8%
	Types	7.0%	36.2%	12.7%
	Full	5.5%	34.6%	9.7%
In	Full	3.7%	23.4%	10.6%

Table 5: Pre-annotation errors and comparison with in-domain (Broadcast News) data

Human annotator performance can then be measured over the same three error types (Table 6). We can see that annotation quality was systematically improved by pre-annotation, with the best global result obtained by full pre-annotation. In addition there was no increase in deletions (had the human stopped looking at the unannotated text) or insertions (had the human always trusted the system) as might have been feared. This may be a side effect of the high deletion rate, making it obvious to the human that the system was missing things. In any case, the annotation was clearly beneficial in our experiment with no ill effects seen in error rates compared to the gold standard.

Pre-annotation	Ins	Del	Subs
Raw	8.9%	18.9%	12.8%
Components	5.9%	16.7%	11.3%
Types	7.1%	16.5%	12.0%
Full	7.1%	16.5%	10.1%

Table 6: Mean human annotation error levels for each pre-annotation scheme

6.3 Is pre-annotation useful and to whom?

All annotators asserted that pre-annotation is useful, specifically with types. In this section, we provide observations concerning variations in annotation both in terms of accuracy (F-measure is used) and duration.

	Raw	Comp.	Types	Full
Experts	0.748	0.786	0.778	0.791
Novices	0.682	0.737	0.721	0.742

Table 7: Mean F-measure of experts and novices, for each pre-annotation scheme

	Raw	Comp.	Types	Full
Experts	109.0	52.5	64.0	39.13
Novices	151.7	135.5	117.9	103.88

Table 8: Mean duration (in minutes) of annotation for experts and novices, for each pre-annotation scheme (two corpus quarters)

Tables 7 and 8 confirm the hypothesis that automatic pre-annotation helps annotators to annotate faster and to be more efficient. All pre-annotation levels (components, types and both) seem to be helpful for both experts and novices. Experts reached a higher accuracy (F=0.791) and they were more than twice faster with components or full pre-annotation. Similarly, novices performed better when working on a full pre-annotation (F=0.742) and reached a faster working time (48mn less than with no pre-annotation). This last observation contradicts the annotators' reported experience: the annotators felt more comfortable and faster with a types-only pre-annotation than with full pre-annotation (see Section 5.1.2). The results show that full pre-annotation was the best choice for both quality and speed.

These results confirm that pre-annotation is useful, even with a moderate level of performance of the system. Does it help to annotate components and types equally? To answer this question, we computed the F-measure of novices and experts for both components and types separately (see Figure 4).

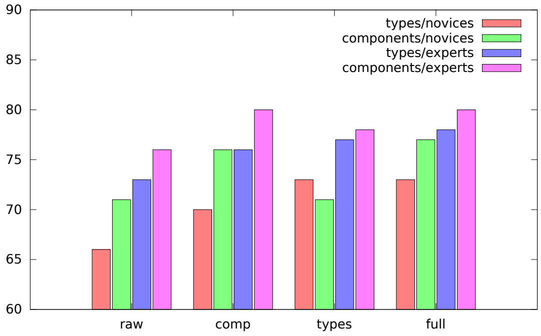

Figure 4: Mean F-measure on each pre-annotation level for expert and novice annotators

For experts we can see that all pre-annotation levels allow them to improve their performance on both types and components. However for novices, pre-annotation with types does not improve their performance in labeling components. We also notice that pre-annotation in both types and components allows experts and novices to reach their best performance for both types and components.

7 Conclusion and Perspectives

Conclusion. In this paper, we studied the interest of a pre-annotation process for a complex annotation task with only an out-of-domain annotation system available. We also designed our experiments to check whether the level of experience of the annotators made a difference in such a context. The experiment produced in the end a high-quality gold standard (8-way merge including 2 versions without pre-annotation) which enabled us to measure quantitatively the performance of every pre-annotation scheme.

We noticed that the pre-annotation system proved relatively precise for such a complex task, with 79% correct pre-annotations, but with a poor recall at 56%. This may be a good operating point for a pre-annotation system to reduce bias though.

In our quantitative experiments we found that the fullest pre-annotation helped most, both in terms of quality and annotation speed, even though the quality of the pre-annotation system varied depending on the annotation layer. This contradicted the feelings of the annotators who thought that a type-only pre-annotation was the most efficient. This shows that in such a setting self-evaluation cannot be trusted. On the other hand their remarks about the problems in the annotation guide itself seemed rather pertinent.

When it comes to experts vs. novices, we noted that their behaviour and remarks were essentially identical. Experts were both better and faster at annotating, but had similar reactions to pre-annotation and essentially the same feelings.

In conclusion, even with an out-of-domain system, a pre-annotation step proves extremely useful in both annotation speed and annotation quality, and at least in our setting, with a reasonably precise system (at the expense of recall) no bias was detectable. In addition, no matter what the annotators feel, as long as precision is good enough, the more pre-annotations the better. Pre-filtering either of our two levels did not help.

Perspectives. Based upon this conclusion, we plan to use automatic pre-annotation in further annotation work, beginning with the present corpora. As a first use, we plan to propose a few changes to the annotation principles in the guidelines we used. To annotate existing corpora with these changes, automatic pre-annotation will be useful.

As a second piece of future work, we plan to annotate new corpora with the existing annotation framework. We also plan to add new types of named entities (e.g., events) to extend the annotation of existing annotated corpora, using the pre-annotation process to reduce the overall workload.

Acknowledgments

This work has been partially funded by OSEO under the Quaero program and by the French ANR VERA project.

References

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Sandipan Dandapat, Priyanka Biswas, Monojit Choudhury, and Kalika Bali. 2009. Complex linguistic annotation – no easy way out! A case from Bangla and Hindi POS labeling tasks. In *Proc of 3rd Linguistic Annotation Workshop (LAW-III)*, pages 10–18, Suntec, Singapore, August. ACL.

Marco Dinarelli and Sophie Rosset. 2011. Models cascade for tree-structured named entity detection. In *Proc of IJCNLP*, pages 1269–1278, Chiang Mai, Thailand.

Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a German named entity recognizer with semantic generalization. In *Proc of Konvens*, Saarbrücken, Germany.

Karën Fort and Benoît Sagot. 2010. Influence of pre-annotation on POS-tagged corpus development. In *Proc of 4th Linguistic Annotation Workshop (LAW-IV)*, pages 56–63, Uppsala, Sweden. ACL.

Karën Fort, Adeline Nazarenko, and Sophie Rosset. 2012. Modeling the complexity of manual annotation tasks: a grid of analysis. In *Proceedings of COLING 2012*, pages 895–910, Mumbai, India, December. The COLING 2012 Organizing Committee.

Olivier Galibert, Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum, and Ludovic Quintard. 2011. Structured and extended named entity evaluation in automatic speech transcriptions. In *Proc of IJCNLP*, Chiang Mai, Thailand.

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard. 2011. Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proc of 5th Linguistic Annotation Workshop (LAW-V)*, pages 92–100, Portland, OR. ACL.

Minlie Huang, Aurélie Névéol, and Zhiyong Lu. 2011. Recommending MeSH terms for annotating biomedical articles. *Journal of the American Medical Informatics Association*, 18(5):660–7.

Geoffrey Leech. 1997. Introducing corpus annotation. In Roger Garside, Geoffrey Leech, and Tony McEnery, editors, *Corpus annotation: Linguistic information from computer text corpora*, pages 1–18. Longman, London.

John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel. 1999. Performance measures for information extraction. In *Proc. of DARPA Broadcast News Workshop*, pages 249–252.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn TreeBank. *Computational Linguistics*, 19(2):313–330.

Claudiu Mihaila, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. 2013. Biocause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*, 14:2.

Céline Poudat and Doninique Longrée. 2009. Variations langagières et annotation morphosyntaxique du latin classique. *Traitement Automatique des Langues*, 50(2):129–148.

Dietrich Rebholz-Schuhmann, Antonio Jimeno, Chen Li, Senay Kafkas, Ian Lewin, Ning Kang, Peter Corbett, David Milward, Ekaterina Buyko, Elena Beisswanger, Kerstin Hornbostel, Alexandre Kouznetsov, René Witte, Jonas B Laurila, Christopher JO Baker, Cheng-Ju Kuo, Simone Clematide, Fabio Rinaldi, Richárd Farkas, György Móra, Kazuo Hara, Laura I Furlong, Michael Rautschka, Mariana Lara Neves, Alberto Pascual-Montano, Qi Wei, Nigel Collier, Md Faisal Mahbub Chowdhury, Alberto Lavelli, Rafael Berlanga, Roser Morante, Vincent Van Asch, Walter Daelemans, José L Marina, Erik van Mulligen, Jan Kors, and Udo Hahn. 2011. Assessment of NER solutions against the first and second CALBC silver standard corpus. *J Biomed Semantics*, 2.

Ines Rehbein, Josef Ruppenhofer, and Caroline Sporleder. 2009. Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation. In *Proc of 3rd Linguistic Annotation Workshop (LAW-III)*, pages 19–26, Suntec, Singapore. ACL.

Eric Ringger, Peter McClanahan, Robbie Haertel, George Busby, Marc Carmen, James Carroll, Kevin Seppi, and Deryle Lonsdale. 2007. Active learning for part-of-speech tagging: Accelerating corpus annotation. In *Proc of Linguistic Annotation Workshop (LAW)*, pages 101–108. ACL.

Sophie Rosset, Cyril Grouin, Karën Fort, Olivier Galibert, Juliette Kahn, and Pierre Zweigenbaum. 2012. Structured named entities in two distinct press corpora: Contemporary broadcast news and old newspapers. In *Proc of 6th Linguistic Annotation Workshop (LAW-VI)*, pages 40–48, Jeju, South Korea. ACL.

William A Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quaterly*, 19(3):321–325.

Burr Settles, Mark Craven, and Lewis Friedland. 2008. Active learning with real annotation costs. In *Proc of the NIPS Workshop on Cost-Sensitive Learning*.

Arne Skjærholt. 2011. More, faster: Accelerated corpus annotation with statistical taggers. *Journal for Language Technology and Computational Linguistics*, 26(2):151–163.

Brett R South, Shuying Shen, Robyn Barrus, Scott L DuVall, Özlem Uzuner, and Charlene Weir. 2011. Qualitative analysis of workflow modifications used to generate the reference standard for the 2010 i2b2/VA challenge. In *Proc of AMIA*.