An Analysis of the Performances of the CasEN Named Entities Recognition System in the Ester2 Evaluation Campaign

Damien Nouvel, Jean-Yves Antoine, Nathalie Friburger, Denis Maurel
Université François Rabelais Tours, LI
Antenne Universitaire de Blois, 3 place Jean Jaurès, F-41000 Blois, France
E-mail: {Damien.Nouvel, Jean-Yves.Antoine, Nathalie.Friburger, Denis.Maurel}@univ-tours.fr

Abstract

In this paper, we present a detailed and critical analysis of the behaviour of the CasEN named entity recognition system during the French Ester2 evaluation campaign. In this project, CasEN has been confronted with the task of detecting and categorizing named entities in manual and automatic transcriptions of radio broadcastings. At first, we give a general presentation of the Ester2 campaign. Then, we describe our system, based on transducers. Next, we depict how systems were evaluated during this campaign and we report the main official results. Afterwards, we investigate in details the influence of some annotation biases which have significantly affected the estimation of the performances of systems. At last, we conduct an in-depth analysis of the effective errors of the CasEN system, providing us with some useful indications about phenomena that gave rise to errors (e.g. metonymy, encapsulation, detection of right boundaries) and are as many challenges for named entity recognition systems.

1. Introduction

The CasEN named entity recognition system, described in this paper, participated to the French Ester2 evaluation campaign. Jointly organized by the French-speaking Speech Communication Association (AFCP) and the French Defense expertise and test center for speech and language processing (DGA/CEP), this campaign has concerned a large variety of speech and spoken language processing tasks that can be classified among segmentation, transcription and information extraction (Galliano et al. 2009). This campaign focused on French speaking radio broadcastings and targeted a wide variety of speaking styles and accents. In particular, the test corpora didn't restrict to broadcast news, but also contained entertainment shows and debates. The evaluation also considered French speaking African radio channels exhibiting strong accents. On the whole, the training, development and test corpora contained French speaking broadcastings from a large variety of sources: France Inter, Radio France International, France Culture, Radio Classique, Africa One, Radio Congo and TVME (Morocco).

The Named Entity (NE) detection task was the only information extraction task. Two subtasks were defined, which only differ in the processed material: reference (manual) transcriptions or transcriptions produced by Automatic Speech Recognition (ASR) systems. Three ASR transcripts (generated by three different systems) have been considered, in order to measure the impact of speech recognition errors on NE recognition. Every system had to detect and categorize the NEs that were present in the corpora. The reference consisted of a tag set of seven main categories: persons (*pers*), locations (*loc*), organizations (*org*), (human) products (*prod*), amounts (*amount*), times (*time*) and positions (*fonc*). This tag set has been divided among 38 sub-categories, but this fine-grained categorization has not been evaluated. The official evaluation measure used was the Slot Error Rate (SER) (Makhoul et al. 1999) but precision, recall and f-score were also computed for further analysis.

Seven systems, implementing a large variety of approaches participated to these tasks, among which our system, CasEN. Five systems were entirely rule-based (LIMSI, LINA, LI, Synapse, Xerox). Two of them carry out only a local analysis, whereas three involved a deep syntactic analysis. Finally, the last two systems (LIA, LSIS) used a machine learning approach based on Conditional Random Fields (CRF).

2. CasEN: named entity recognition using transducers

The NE recognition system CasEN relies on the CasSys system (Friburger, 2002). This platform processes texts using cascades of transducers. CasSys applies transducers in a predefined order: every transducer deletes or modifies text strings that match a specific pattern. The advantage of using transducers within a cascade (rather than one transducer) is that we first look for "islands of certainty" (Abney 1996), thus reducing the search space for further transducers.

CasSys uses the Unitex[1] toolkit to design, compile and apply transducers, and also provides additional behaviors to those packaged with the toolkit. Transducers describe linguistic constructs containing morphological, lexical and syntactic patterns to be searched in texts, and define actions (insertion or replacements) to be taken on the resulting strings. Such a system can be used for any task that needs to write rules, like chunking (Antoine et al. 2008), syntactic analysis or NE recognition for example.

CasEN is a cascade of transducers dedicated to NE recognition that runs on the CasSys platform. The first

[1] http://www-igm.univ-mlv.fr/~unitex/

version of CasEN was conceived for NE recognition on written texts. It includes about 150 transducers, which are each dedicated to the recognition of sequences of words that shall contain a NE (Friburger & Maurel 2004). Our experiments on a test corpus (from Le Monde newspaper) have exhibited a recall of 93% and a precision of 94% on proper names (Friburger, 2006). CasEN was involved in the *VariLing* project (Maurel et al. 2009), where it was greatly improved for the recognition of ENs in texts. The version of the system involved in the Ester2 campaign is an adaptation of the latter to speech transcripts and to spoken language.

Figure 1 shows a transducer, as it is designed using Unitex. This one is aimed at recognizing political organization. Each part of the string to be recognized is visualized as a box, that contains alternatives of words or syntactical categories to match. The whole expression to be detected is simply a path through this graph.

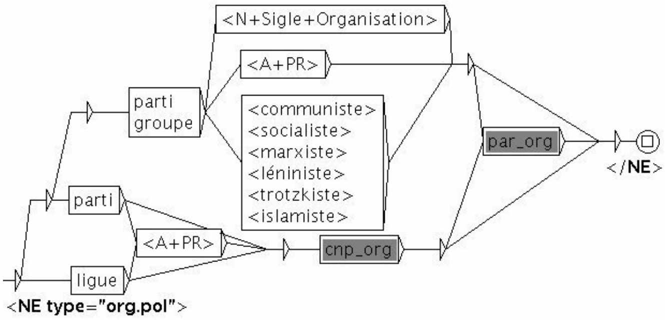


Figure 1: A transducer for political organizations

3. Results analysis

Tables 1 and 2 summarize the results of the EN Ester2 campaign (Galliano et al. 2009) on manual and the available ASR transcripts, which accuracy was evaluated by their Word Error Rate (WER). Among rule-based systems, those performing a deep syntactic analysis (Synapse, Xerox) get best results for manual transcript. However, this advantage is lost on ASR transcripts, where a machine learning approach (LIA) came first, closely followed by rule-based systems (LIMSI, LINA).

	Manual transcript		
	SER	P	R
LIA	23,9	86,4	71,8
LIMSI	30,9	81,1	70,9
LINA	37,1	80,7	55,4
LI Tours	33,7	79,3	65,8
LSIS	35,0	82,6	73,0
Synapse	9,9	93,0	89,3
Xerox	9,8	93,6	91,5

Table 1: Ester2 evaluation campaign scores (SER, Precision, Recall), manual transcript

	ASR1 (capitalized, WER: 12,11)		ASR2 (non capitalized, WER: 17,83)		ASR3 (non capitalized, WER: 26,09)	
	SER	ΔSER	SER	ΔSER	SER	ΔSER
LIA	43,4	-19,5	51,6	-27,7	56,8	-5,2
LIMSI	45,3	-14,4	55,5	-24,6	61,2	-5,7
LINA	54,0	-16,9	60,4	-23,3	65,2	-4,8
LI Tours	50,7	-17,0	80,8	-47,1	82,9	-2,1
LSIS	55,3	-20,3	86,5	-51,5	88,6	-2,1
Synapse	44,9	-35,0	60,7	-50,8	66,2	-5,5
Xerox	44,6	-34,8	-	-	-	-

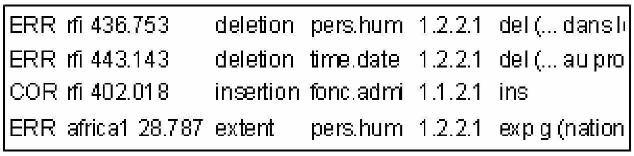
Table 2: Ester2 evaluation campaign scores (SER, SER variation), ASR transcripts

This shows us that symbolic and statistical approaches have potentially comparable performances. Considering the SER, one can see that our CasEN (LI Tours) system is ranked in 5th or 6th position, depending on the corpus. If our precision is reasonably good, the recall is much lower and is a weakness of our system.

Our system had difficulties to process manual transcriptions, probably because it was initially designed to process written text using evidences (McDonald 1996) to describe regular forms of NEs. But its results on ASR1 transcript are quite satisfactory considering the difficulty of the task, maybe because it doesn't rely on a deep syntactic analysis.

Regarding ASR2 and ASR3, one shall mention that ASR1 did provide capitalized proper nouns, whereas others did not. Our system didn't implement a dedicated module to detect missing capitalizations, what partly explain the great difference of results between ASR1 and and the two other ASR transcripts. On the opposite, it seems that our system is reasonably affected by the increase of automatic transcription errors (WER), as shown by the slight differences of SER between ASR2 and ASR3. On the whole, we consider encouraging that our system was not overwhelmed by those specifically designed for spontaneous speech. But there is room for improvement and we will here focus on this question.

This paper analyses the results of this evaluation campaign to determine on what difficulties our system has been the most challenged. Since the annotation reference and the scoring software are available, we can evaluate ourselves and assess improvements. Every error logged by the scoring software has been annotated with: its location, the error type (deletion, insertion, erroneous tag, extent error…), the rule from Ester2's convention that applies in that specific situation and some indications about the context within which the error appeared (Figure 2). We examined half of the reference corpus (41 Kwords, 5890 NE, 1180 errors) so as to determine what directions should be investigated to improve our results.



```
ERR rfi 436.753     deletion  pers.hum   1.2.2.1  del (... dans l
ERR rfi 443.143     deletion  time.date  1.2.2.1  del (... au pro
COR rfi 402.018     insertion fonc.admi  1.1.2.1  ins
ERR africa1 28.787  extent    pers.hum   1.2.2.1  exp g (nation
```

Figure 2: The error characterization file

Furthermore, we also decided to correct to some extent the reference annotation. As noticed by other participants, this reference contained errors or inconsistencies with the annotation guide, what obviously prevaricates evaluation. The idea behind this correction of the reference is to have as much confidence as possible in the score that is computed over these files: every error identified by the scoring software was classified as a (real) error or, on the opposite, as an annotation error.

4. Annotation bias for evaluation

4.1 Detected errors within the reference

We tried to estimate the influence of the annotation errors or inconsistencies on the overall results of the Ester2 campaign. Within those corrections, some are NEs that have not been found by annotators (43 over 99 corrections of the reference) but that were correctly detected by our system. Consider for instance the following sentence: “*Ensuite c'est le [président] de l'association [...]*” (transl. “*Afterwards, it is the [president] of the association [...]*”). This EN is not present in the reference, while CasEN correctly detected it: the annotation guide recommends considering “*president*” as a NE:

(2.3.1.3) Annotate a position even if the person holding it is not named

As expected, these errors significantly penalized the system. Table 3 presents the variation of the different error metrics after correcting the reference annotation. One should observe for instance a reduction of almost 10% of the SER (31.0 vs 33.7 in the official results).

SER	P	R	F
31,0 (-2,7)	82,5 (+3,2)	66,9 (+1,1)	0,74 (+0,02)

Table 3: Variation of scores (SER, Precision, Recall, F-score) after reference correction

Moreover, we lately realized that rule *(1.1.6.1)* of the annotation guide, which restricts imbrication of NEs to *pers* with a *fonc* was several times violated. We found 44 exceptions to this rule, that should therefore not have been tagged. The impact of those annotation errors has not been assessed, but it reveals the great difficulty to have an evaluation we can rely on.

4.2 Influence of named entity categorization

With 7 NE main types, the Ester2 campaign has introduced a NE categorisation which is more precise than those considered by previous evaluations (see the MUC conferences, for instance). This classification has a limited but indisputable drawback: the differences between some categories (and/or sub-categories) are sometimes slight; this also explains that annotators met difficulties to classify an NE in the dedicated categories.

In order to measure the influence of this classification, we have conducted experimentations on potentially conflictual couples of sub-categories that belong to different main NE tags. For instance, the distinction between *loc.admi* (an administrative location) and *org.gsp* (a geo-political organization) or *org.div* (entertainment or sport organization) is not trivial: France may be considered either as a geographical entity, as a political organization or as a sport team, depending on circumstances. The category assignment may be controversial, even for a human. A great part of those conflicts are caused by metonymy (Markert & Hahn 2002), when using a proper name in a sense that is somehow related to its literal value. Consider for instance the following annotations for the NE *“Maroc” (transl. “Morocco”)* in reference corpus:

<u>(1) Administrative localisation:</u>
“le stade le plus grand du [Maroc] sera construit [...]”
(transl. “biggest stadium in [Morocco] will be built [...]”)
<u>(2) Political organization:</u>
“l'unité territoriale du [Maroc] [...]”
(transl. “the territorial unity of [Morocco] [...]”)
<u>(3) Sport team:</u>
“La guinée a battu le [Maroc] trois à deux [...]”
(transl. “Guinea defeated [Morocco] three to two [...]”)

The differences between these annotations are sometimes very slight. In particular, the example (2) has been annotated as a political organization. One may however wonder whether it shouldn't be considered as an administrative location, as shown by the introducing context “*territorial unity*”. CasEN recognized the latter, what was considered as an error. Likewise, the distinction between the Ester2 *time.date* (a date or a period located on a calendar) and *amount.phy.dur* (a duration) categories should be questioned.

These categories misclassifications lead to type errors. To quantify their impact on scores, we evaluated our system after the merge of some of the conflictual subcategories. More precisely, we decided to merge *loc.admi* and *org.gsp* types, considering *org.gsp* was inherently too ambiguous. On the contrary, we still counted the *time.date* and *amount.phy.rel*, types separately, judging this distinction makes sense. Table 4 shows that both merges lead to a non-negligible improvement of the performances that should, to a certain extent, concern all participants.

	#NE	SER	P	R	F
loc.admi + org.gsp	477 + 156	28,3 (-2,7)	85,4 (+2,9)	69,3 (+2,4)	0,77 (+,03)
time.date + amount.phy.dur	631+ 53	27,7 (-0,6)	87,6 (+2,2)	70,4 (+1,1)	0,79 (+,02)

Table 4: Variation of scores (SER, Precision, Recall, F-score) after categories merges

4.3 Encapsulated NEs and boundary errors

For any evaluation campaign, artefactual errors are often found, which are due to differences between the system and the reference representation schemes. During the Ester2 campaign, CasEN has faced two kinds of such errors that could have easily been avoided, since our system was designed for another project.

The assessed version of our system didn't detect encapsulated NEs. But, as explained above, imbrication of NEs had to be detected in the Ester2 campaign: a name (*pers*) with a contiguous position (*fonc*) should be encapsulated within a *pers* NE. For instance, the string *"le président Museveni"* (transl. *"the president Musuveni"*), should be tagged as an encompassing *pers "[[president] [Museveni]]"*, containing *"[president]"* as a *fonc* and *"[Museveni]"* as an included *pers*. We didn't focus on implementing this feature, hence, 33 of those NEs were undetectable for our system.

Furthermore, the annotation guide explained whether the determiner should be included within NE tag, depending on its type. Most of the time, the guide required to include it solely for *time* and *amount* NEs. Since these annotations rules were not consistent from one NE type to another, we decided to pass over these constraints, what obviously caused unavoidable extent errors.

We achieved a few simple adaptations after the end of the Ester2 campaign, so as to have an idea of how much our system's performances decreased due to these lacks. Table 5 presents the corresponding score variations.

SER	P	R	F
25,5 (-2,8)	86,0 (+0,6)	71,7 (+2,4)	0,79 (+0,02)

Table 5: Variation of scores (SER, Precision, Recall, F-score) after system adaptations

4.4 Annotation bias : conclusion

The errors pointed out in the annotating procedure clearly advocate for a much more reliable and transparent process before evaluating systems. In that campaign, the annotation error rate is of 3% (100 errors overs 3000 NEs) and 9% of our evaluation errors (100 annotation errors overs 1100 errors issued from evaluation). These issues and related questions are an emerging topic for further investigation and research (Fort et al. 2009).

The scores improvements obtained by merging categories, emphasizes the great importance of the taxonomy, for systems to have confidence in their NE recognition. Ambiguities among NE categories, may be quite significant and therefore have great impact on scores. On Ester1 campaign, ambiguity rate (the proportion of sequence of words belonging to at least two subcategories, as Morocco to *org.gsp*, *loc.admi* and *org.div*) has been measured from 40% (development corpus) to 32% (test corpus) (Favre et al. 2005).

But it is also obvious that our system had some deficiencies regarding the annotation requirements. Indeed, this is inherent to every evaluation campaign: results are partly determined by the amount of time teams devote for improving and adapting system to the evaluation process. From a general point of view, we do not consider those annotation-specific difficulties as relevant to assess the quality of our system.

We will now detail results obtained by analysing the insights of CasEN, to determine the most promising directions so as to enhance our system.

5. Qualitative analysis of our results

5.1 NE types and error characterization

Figure 3 presents the results of CasEN according to NE categories. The precision is quite satisfactory, especially for *amount*, *pers* and *time*. Scores are very low on *prod* category: those NEs often involve metonymic uses, they are less frequent (less attention is devoted to them) and other participants met difficulties on this category too. Recall varies significantly from one category to another, and is quite low on categories *org* and *fonc*.

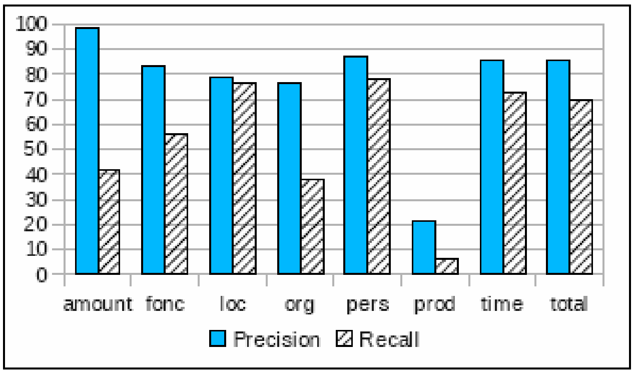


Figure 3: CasEN results by category

African broadcastings are mainly responsible for this situation. The CasEN vocabulary clearly lacks coverage for African actual proper names. For instance *"Hissene Habré"*, (former Tchad president) appears five times in one single transcription among 26. It is never recognized by our system, since it corresponds to Out Of Vocabulary (OOV) words for our system. Facing this problem, one may consider using encyclopaedias and large coverage lexicons, while others would rather look for morphological or contextual information so as to detect those NEs. The campaign has shown that having larger vocabularies gave a crucial advantage to participants, what does not prevent the investigation of additional approaches to reduce the influence of OOV.

Table 6 presents the distribution of errors according to the five main EN types. Erroneous categorisations (Category Conflicts) mainly corresponds to the misclassifications studied on § 4.2. Reference Errors are related to the errors in manual annotation of the reference that we have detected. As explained before, Encapsulated NEs have been missed because the assessed version of CasEN didn't implement their detection. Not found NE are errors described by the scoring software as “deletion”, NEs we didn't find (those include the inner part of encapsulated NEs). Finally, Wrong Extent corresponds to errors of delimitations (boundaries of NEs). While one should consider that the three first error types do not identify real errors or are corresponding to problems which are currently solved (encapsulated NE), the last two error types clearly challenged our system. They represent around 60% of the official Ester2 errors, what enables to situate more precisely the real performances of our system. We will discuss below what causes this two kinds of errors.

Annotation bias			CasEN effective errors	
Category Conflicts	Reference Errors	Encapsulated NE	Not found NE	Wrong Extent
10,6 %	8,2 %	4,4 %	30,99 %	34,9 %

Table 6: CasEN errors distribution

5.2 Transducers evaluation

A careful investigation of the behaviour of every transducer can provide useful information on the main sources of errors of CasEN. We conducted an experiment with the most recent version of our system to see how much confidence we could have in every transducer individually. For this purpose, we logged errors occurring for each transducer during recognition, what allowed us to evaluate its precision. Regarding recall, it is not so straightforward to know, for each transducer, what NEs are missed, thus the computed metrics do not include this kind of errors.

Transducers *“loc_tpays”*, *“loc_tville”* and *“loc_tgeo”* search for locations (countries, cities a.s.o.), *“person102”*, *“balais_pers”* and *“tpresident”* for person names (presidents for the latter), *“org1”* recognizes organizations and *“dettps”*, time expressions. Figure 4 shows error impact (proportion of errors generated by a transducer over all errors) for transducers that generated the most errors.

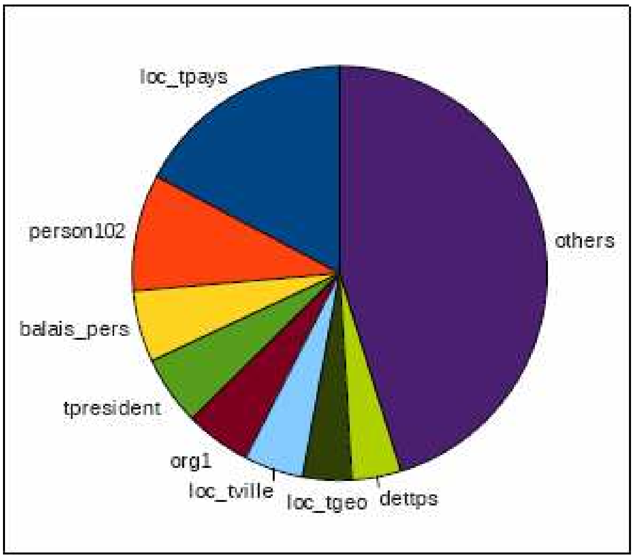


Figure 4: Transducers error impact

We notice that eight transducers generate 54% of the errors. However, since those transducers recognize 60% of the NEs, additional information is necessary about their respective accuracy. For each, figure 5 depicts its SER and Average error rate (number of errors a transducer generated over how many NE it recognized).

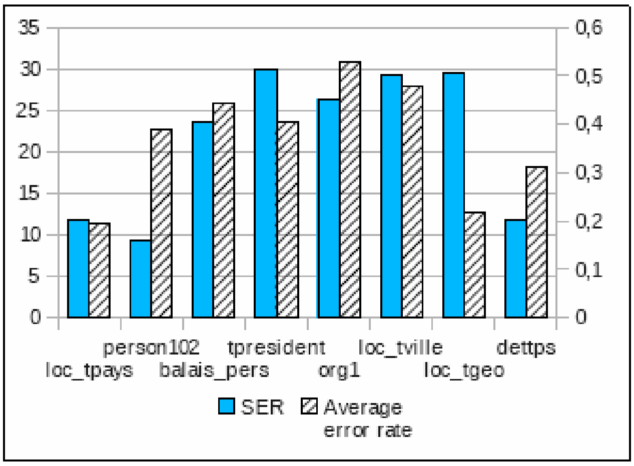


Figure 5: Transducers SER and Average error rate

The transducers looking for locations are mainly erroneous due to metonymic uses, frequently related to sport (e.g. *“Egypt”* as *org.div*, a football team) or to political organization (e.g. *“Paris”* denoting the France government). For these issues, we are currently testing text mining algorithms, to find what relevant context surrounding these NEs would allow to disambiguate.

Transducers detecting persons are still challenged by detection of encapsulated NEs, especially because of their nested position (*fonc*). Among them, presidents or ministers have frequently very long titles, whose ending boundary is hard to find. These complex situations may also occur for organizations and time expressions. Consider this example: *“la [chambre régionale des* *comptes d'Ile-de-France] [...]"* (transl. *"the [regional accountability office of Ile-de-France] [...]"*). The right boundary of such a large spanned NE is hard to detect. To overcome this issue, we intend to implement a chunker (Antoine et al. 2008) to provide linguistically consistent groups of words for delimiting NEs.

6. Conclusions and future work

In this paper, we have discussed in details the performances of the CasEN recognition system in the Ester2 evaluation campaign. We have pointed out some weaknesses of our system. Some of them may be easily explained by the fact that our system was dedicated to process written input and not adapted for handling ASR degradations: from this point of view, our results on speech transcripts are encouraging. But other errors are due to the complexity of the considered NEs themselves (metonymic uses, boundaries of NEs) and have certainly challenged other participants.

One should reasonably suppose that limitations of systems differ from one participant to another. It would indeed have been very interesting to evaluate the behaviour of a global system that merged the outputs of participants through a voting procedure, as done by (Brun et al. 2009). This idea was expressed during the closing workshop of the Ester2 campaign and should be taken up for future campaigns. Besides, systems implementing a deep syntactic analysis seem to obtain better results, at least for reference (manual) transcript: knowing more precisely how much such deep processing (Brun & Hagège 2008) contributes to the overall process would help determining the most promising approaches dedicated to NE recognition.

Our conclusions may also be related to a general trend within NLP: corpus-based approaches and machine learning techniques (symbolic, pattern mining, statistical) may address some issues for robustly processing large amounts of data, by inferring lexicons and descriptions of a wide variety of forms within a language, so as to reach a high recall for Information Extraction tasks. We are currently working on sequence mining approaches, more specifically frequent episode mining (Mannila et al. 1997). Some previous experiments give us hope that pattern mining and association rules (Budi & Bressan 2003) may help regarding coverage. We consider using encyclopaedias (Charton & Torres-Moreno 2009), while keeping in mind inherent limitations due to the dependency of NEs over time (Favre et al. 2005). For metonymy uses, patterns as LSR (Plantevit et al. 2009) could be well-suited to find relevant context as a half-constrained sequence of words.

7. Acknowledgements

This work was supported by the two projects Variling (ANR-06-CORP-023), Epac (ANR-00-MDCA-006-03) founded by the French research agency (ANR), and also by the Feder (European Regional Development Fund) project *Entités nommées et nommables*.

8. References

Abney, S. (1996). Partial Parsing via Finite-State Cascades. *Workshop on Robust parsing, 8th European Summer School in Logic, Language and Information (ESSLLI)*, pp. 8-15. Prague, Czech Republic.

Antoine, J.Y., Mokrane, A., Friburger , N. (2008). Automatic rich annotation of large corpus of conversational transcribed speech : the chunking task of the EPAC project. *Proceedings of the 6th European Conference on Language Resources and Evaluation (LREC).* Marrakech, Maroc.

Brun, C., Dessaigne, N., Ehrmann, M., Gaillard, B., Guillemin-Lanne, S., Jacquet, G., Kaplan, A., Kucharski, M., Martineau, C., Migeotte, A., Nakamura, T., Voyatzi, S. (2009). Une expérience de fusion pour l'annotation d'entités nommées. *Actes TALN'2009*. Senlis, France.

Brun, C., Hagège, C., (2008). Vérification sémantique pour l'annotation d'entités nommées. *Actes TALN'2008*. Avignon, France.

Budi, I., Bressan, S., (2003). Association Rules Mining for Name Entity Recognition. *Proceedings of the 4th International Conference on Web Information Systems Engineering (WISE)*. Roma, Italy.

Charton, E., Torres-Moreno, J.M. (2009). Classification d'un contenu encyclopédique en vue d'un étiquetage par entités nommées. *Actes TALN'2009*. Senlis, France.(Eshkol et al. 2010)

Eshkol, I., Maurel, D., Friburger, N. (2010). Eslo : from transcription to speakers' personal information annotation. *Proceedings of the 7th European Conference on Language Resources and Evaluation (LREC).* Valetta, Malte.

Favre, B., Béchet, F., Nocéra, P. (2005). Robust Named Entity Extraction from Spoken Archives. *Proceedings of the joint conference Human Language Technology Empirical Methods for Natural Language Processing (HTL-EMNLP)*. Vancouver, Canada.

Fort, K., Ehrmann, M., Nazarenko, A. (2009). Vers une méthodologie d'annotation des entités nommées en corpus ? *Actes TALN'2009*. Senlis, France.

Friburger, N. (2002). Reconnaissance automatique des noms propres : application à la classification automatique de textes journalistiques. PhD thesis, Université François-Rabelais Tours (directed by Maurel, D.).

Friburger, N. (2006). Linguistique et reconnaissance automatique des noms propres. In *Meta: Translators' Journal*. Les Presses de l'Université de Montréal, vol. 51-4, pp. 637-650.

Friburger, F., Maurel, D. (2004). Finite-state transducer cascades to extract named entities in texts. In *Theoretical Computer Science*. Essex, UK: Elsevier Science Publishers, vol. 313, pp. 94-104.

Galliano, S., Gravier, G., Chaubard, L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. *Proceedings of the 10th Conference of the International Speech Communication Association (Interspeech),* pp.

2583-2586. Brighton, UK.

Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R. (1999). Performance measures for information extraction. *Proceedings of DARPA Broadcast News Workshop*. Herndon, VA: Morgan Kaufmann, pp. 249-252.

Mannila, H., Toivonen, H., Verkamo, A.I. (1997). Discovery of Frequent Episodes in Event Sequences. In *Data Mining and Knowledge Discovery*. Hingham, MA: Kluwer Academic Publishers, pp. 259-289.

Markert, K., Hahn, U. (2002). Understanding metonymies in discourse. In *Artificial Intelligence*. Essex, UK: Elsevier Science Publishers, vol. 135, pp. 145-198.

McDonald, D.D. (1996). Internal and External Evidence in the Identification and Semantic Categorisation of Proper Names. In *Corpus processing for lexical acquisition*. Cambridge, MA: MIT Press, pp. 21-39.

Plantevit, M., Charnois, T., Klema, J., Rigotti, C., Cremilleux, B (2009). Combining Sequence and Itemset Mining to Discover Named Entities in Biomedical Texts: A New Type of Patter. In *International Journal of Data Mining, Modelling and Management*. Geneva, Switzerland: Inderscience Publishers, vol. 1-2, pp. 119-148.