Convertir des analyses syntaxiques en dépendances vers les relations fonctionnelles PASSAGE

Patrick Paroubek Munshi Asadullah Anne Vilnat
LIMSI-CNRS Bât. 508 Université Paris-Sud
91403 Orsay Cedex France
munshi@limsi.fr, pap@limsi.fr, anne@limsi.fr

RÉSUMÉ

Nous présentons ici les premiers travaux concernant l'établissement d'une passerelle bidirectionnelle entre d'une, part les schémas d'annotation syntaxique en dépendances qui ont été définis pour convertir les annotations du French Treebank en arbres de dépendances de surface pour l'analyseur syntaxique Bonsai, et d'autre part le formalisme d'annotation PASSAGE développé initialement pour servir de support à des campagnes d'évaluation ouvertes en mode objectif quantitatif boîte-noire pour l'analyse syntaxique du français.

ABSTRACT

Converting dependencies for syntactic analysis of French into PASSAGE functional relations

We present here a first attempt at building a bidrictionnal converter between, on the one hand the dependency based syntaxtic formalism which has been defined to map the French Treebank annotation onto surface dependency trees used by the Bonsai parser, on the other hand the PASSAGE formalism developped intially for French parsing quantitative black-box objective open evaluation campaigns.

MOTS-CLÉS : Analyse Syntaxique - Corpus arboré - Dependances -DepFTB - ConLL - PASSAGE.

KEYWORDS: Parsing - Treebank - Dependencies - DepFTB -ConLL - PASSAGE.

1 Introduction

Le travail que nous présentons ici repose sur deux constats : d'une part, si l'on s'intéresse aux analyseurs syntaxiques du français librement disponibles et prêts à l'emploi, on trouve essentiellement des analyseurs qui produisent une représentation utilisant le format ConLL(Buchholz et Marsi, 2006), et suivant les normes du French TreeBank (FTB) (Abeillé *et al.*, 2000), à l'image des parsers adaptés au français décrits dans (Candito *et al.*, 2010). D'autre part, quand on veut évaluer les performances de ces analyseurs peu de corpus existent avec des annotations du même format, à part Sequoia obtenu par conversion du FTB (Candito et Seddah, 2012) qui compte environ 3200 énoncés, alors qu'il existe le corpus des campagnes PASSAGE (Vilnat *et al.*, 2010; de la Clergerie *et al.*, 2008) qui a donné lieu à l'annotation manuelle d'au moins 14000 énoncés dont 8200 libres de droits [1], dont une partie comprenant divers genres issue de la

1. Ce corpus est accessible sur les serveur d'évaluation http ://passageval.limsi.fr et en cours de déploiement sur http ://www.elda.org, où l'on peut aussi obtenir la partie sous-droits du corpus.

campagne EASY (Paroubek *et al.*, 2006). Cependant, le corpus PASSAGE utilise un formalisme d'annotation comprenant un niveau de groupes syntaxiques et des fonctions grammaticales propres au projet. On voudrait donc pouvoir passer d'un schéma d'annotation à l'autre pour permettre l'usage de ces analyseurs et évaluer leur performances par rapport au corpus PASSAGE, ou bien pour certaines tâches d'analyse du langage, préférer utiliser le formalisme PASSAGE qui relache certaines contraintes des représentations syntaxiques classiques, ou encore disposer d'un convertisseur bidirectionnel pour effectuer des comparaisons entre les formalismes d'annotation. Les analyseurs syntaxiques des dernières années ont très souvent suivi des variantes du modèle de dépendances de l'analyseur de Stanford : SD, décrit dans (M.-C. de Marneffe et C D Manning, 2008), où les auteurs ont établis des comparaisons de formalismes entre SD, GR (*Grammatical Relation*) (Carroll *et al.*, 1999), développé pour faire des évaluations et PARC (Tracy Holloway King *et al.*, 2003), utilisé pour évaluer des analyseurs LFG (Kaplan *et al.*, 2004). Dans (Sagae *et al.*, 2008), les auteurs montrent comment convertir des représentation syntaxiques pour évaluer des analyseurs syntaxiques (fondés sur des formalismes différents) avec le Penn Tree Bank (PTB) et des textes de la littérature biomédicale académique, en utilisant des convertisseurs entre SD, GR, un analyseur syntaxique profond (une implémentation de HPSG (Miyao *et al.*, 2004)) et l'analyse de surface du PTB. Tous ces travaux ont été faits sur l'anglais, en faisant un large usage du PTB. On retrouve dans la majorité des travaux les mêmes familles de formalismes utilisés dans les différentes comparaisons que ceux pris ici comme exemple.

L'article présenté ici va donc d'abord présenter les formats de Bonsai, suivant ConLL, et PASSAGE. Pour vérifier l'hypothèse de compatibilité de ces deux formats, nous présenterons le corpus issu du FTB que nous avons constitué, et son annotation manuelle au format Passage. Les principes que nous avons suivis pour pemettre les conversions dans les deux sens (Passage vers ConLL, ou ConLL vers Passage) feront l'objet de la suite de cette présentation, et nous évaluerons enfin les résultats obtenus par ces deux convertisseurs, avant de conclure sur les suites de ce travail, et en particulier sur son application dans le projet PROJESTIMATE [2].

2 Dep-FTB pour le français

Nous nous intéressons aux schémas d'annotation décrits dans (Candito *et al.*, 2011) [3] qui décrivent les annotations produites par les analyseurs syntaxiques Bonsai [4]. Ces analyseur syntaxiques produisent des annotations dans le format de données ConLL, initialement développé pour des campagnes d'évaluation d'analyse syntaxique en dépendances (Buchholz et Marsi, 2006). Cette représentation utilise une représentation matricielle dont la première colonne contient les formes de l'énoncé, puis les autre colonnes suivantes leurs étiquettes morpho-syntaxiques, et ensuite les dépendances syntaxiques. C'est un format extensible, auquel on peut ajouter de nouvelles couches d'analyse par simple ajout de colonnes à la matrice de représentation. Les dépendances sont représentées aux moyen de deux colonnes, l'une pour le type de la dépendance, l'autre pour l'adresse de sa cible, qui référence une ligne de la matrice, la source de la dépendance étant la forme courante. La Table 1 illustre cette représentation.

Notons que nous aurons essentiellement deux types d'information, des étiquettes (morpho-

2. PROJESTIMATE est un projet FUI - CAP DIGITAL (2012-2015) sur l'analyse automatique de spécifications logicielles pour l'estimation de coûts, qui finance ces travaux

3. Guide d'annotation ConLL FTB : http ://alpage.inria.fr/statgram/frdep/Publications/FTB-GuideDepSurface.pdf

4. `http://alpage.inria.fr/statgram/frdep/fr_stat_dep_bky.html`

1	Je	cln	CL	CLS	s=suj	2	suj	2	suj
2	remercie	remercier	V	V	m=ind|n=s|p=3	0	root	0	root
3	le	le	D	DET	g=m|n=s|s=def	4	det	4	det
4	président	président	N	NC	g=m|n=s|s=c	2	obj	2	obj
5	en	en	P	P	p=3	4	dep	4	dep
6	exercice	exercice	N	NC	g=m|n=s|s=c	5	obj	5	obj
7	pour	pour	P	P	_	2	mod	2	mod
8	sa	son	D	DET	g=f|n=s|s=poss	9	det	9	det
9	réponse	réponse	N	NC	g=f|n=s|s=c	7	obj	7	obj
10	.	.	PONCT	PONCT	s=s	2	ponct	2	ponct

TABLE 1 – Extrait d'annotation ConLL issu du corpus Sequoia v4.0

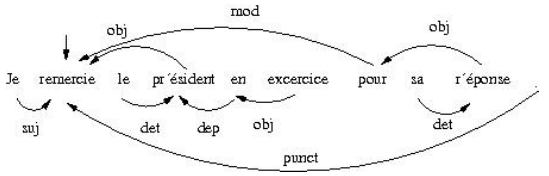


FIGURE 1 – Représentation graphique des dépendances de l'extrait d'annotation ConLL issu du corpus Sequoia v4.0

syntaxique, sémantique, se rapportant à un système de classification particulier, etc.) associées à une forme particulière, et des dépendances étiquetées qui vont relier des formes ensembles, avec la condition qu'une forme ne peut engendrer qu'une dépendance, mais par contre plusieurs dependances peuvent aboutir sur une même forme. Le guide d'annotation Bonsai recense 12 dépendances pour les gouverneurs verbaux : *suj (Sujet), obj (objet), de_obj (SP argumental en de, non locatif), a_obj (SP argumental en à, non locatif), p_obj (autre SP argumental), ats (Attribut du sujet), ato (Attribut de l'objet), mod (Modifieur), aux_tps (auxiliaires de temps), aux_pass (auxiliaires du passif), aux_caus (verbe causatif, en cas de complexe causatif + inf), aff (clitiques figés)* et 8 dépendances pour gouverneurs non verbaux : *mod (Modifieurs repérés structuralement, comme par exemple les adjectifs épithètes, autres que les relatives), mod_rel (Relatives adnominales), coord (Relation portée par un coordonnant, avec comme gouverneur le coordonné immédiatement précédent), arg (utilisé dans le cas de prépositions « liées », ex. « Charybde en Scylla », dep_coord (Relation portée par un coordonné différent du premier, avec comme gouverneur le coordonnant immédiatement précédent), det (Relation portée par les déterminants), ponct (Relation portée par tout dépendant typographique, sauf pour les virgules jouant le rôle de coordonnant), dep (Relation sous-spécifiée, pour les dépendants prépositionnels (pas de gestion de la distinction argument / ajout pour les gouverneurs non verbaux)*. Il s'agit là des représentations utilisées pour l'annotations automatique. L'annotation manuelle proposée dans (Candito *et al.*, 2011) ajoute les 8 dépendances suivantes : *mod_loc (Modifieurs sémantiquement locatifs, au propre ou au figuré), mod_cleft (Pour la subordonnée dans le cas d'une clivée), p_obj_agt (Pour les compléments d'agent, passif ou causatif), p_obj_loc (Dépendants argumentaux locatifs, source, destination, ou localisation), suj_impers (Pour le sujet explétif il), aff_moyen (Pour le clitique se en cas de moyen), arg_comp (Utilisé pour relier une comparative et son gouverneur), et finalement arg_cons (Utilisé pour relier une consécutive et son gouverneur adverbial.* La Figure 1 illustre certaines de ces dépendances.

3 Le formalisme d'annotation PASSAGE

PASSAGE(Vilnat *et al.*, 2010; de la Clergerie *et al.*, 2008) annote à la fois des groupes et des relations de dépendance [5]. Les groupes sont des constituants minimaux non récursifs. Il y en a 6 : le groupe nominal (GN), prépositionnel (GP), adjectival (GA), adverbial (GR), le noyau verbal (NV) et verbal prépositionnel (PV). Les 14 relations sont établies entre ces groupes ou entre les formes au sein de ces groupes. Il s'agit de lier le sujet au verbe (SUJ-V), le verbe à son auxiliaire (AUX-V), l'objet direct (COD-V), ou le complément [6] (CPL-V) , ou tous les autres modifieurs optionnels (MOD-V) au verbe. On annote aussi tous les autres types de modifieurs : du nom (MOD-N), de l'adjectif (MOD-A), de l'adverbe (MOD-R) ou de la préposition (MOD-P). On identifie aussi l'attribut du sujet ou de l'objet (ATT-SO), ainsi que le lien entre l'introducteur d'une subordonnée et son noyau verbal (COMP). Les trois dernières relations sont la coordiantion (COORD), la juxtaposition (JUXT) et l'apposition (APP). La Figure 2 illustre ce schéma d'annotation.

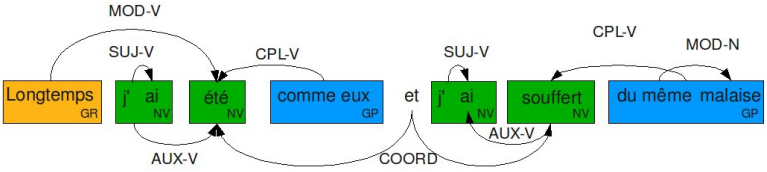


FIGURE 2 – Exemple d'annotation au format PASSAGE

Une comparaison entre les annotations PASSAGE transposées sur l'anglais et celles de SD et de PARC est présentée dans (Paroubek *et al.*, 2009).

4 Dep-FTB vers PASSAGE

Par rapport au schéma d'annotation Dep-FTB, les annotations PASSAGE sont moins spécifiques, la conversion vers les annotations PASSAGE ne pose pas de problème majeur et peut-être effectuée avec un système à base de règles hiérarchisées, déclenchées par des patrons mélangeant annotations et formes. L'ordre de déclenchement des règles va du plus spécifique au moins spécifique. Actuellement 45 règles ont été nécessaires pour traiter l'intégralité des phénomènes linguistiques présents. La Figure 3 donne un exemple de règles de projection des annotations.

Les éventuels problèmes que l'on va rencontrer lors de la conversion, vont concerner d'une part les différences de segmentation en mots entre les deux formalismes et d'autre part le positionnement précis des frontières des groupes syntaxiques PASSAGE cibles (pour les règles qui en définissent). Nous avons contourné les différences de segmentation au moyen d'un algorithme de programmation dynamique (Makhoul *et al.*, 1999) pour le réalignement de formes et la détermination précises des frontières de groupes syntaxiques cibles est un problème secondaire,

5. Guide d'annotation en français : http ://www.limsi.fr/Individu/pap/PEAS_reference_annotations_v2.2.html
6. qu'il s'agisse d'un adjoint ou d'un indirect (qu'on ne cherchera pas à distinguer)

```
AUX_TPS(?var1, ?var2) → AUX_V(?var2, ?var1)
```

```
Les enfants ont vu le concert
AUX_TPS(vu-3, ont-2) → AUX_V(ont-2, vu-3)
```

```
COORD(?var1, ?var2) + DEP_COORD(?var2, ?var3) → COORD(?var2, ?var1, ?var2)
```

```
Jean aime Marie et Paul aime Virginie!
COORD(aime-1, et), DEP_COORD(et, aime-3) → COORD(et, aime-1, aime-3)
```

```
AUX_CAUS(?var1, ?var2) + SUJ(?var1, ?var3) →
                          COD_V(?var2, ?var1) + SUJ_V(?var3, ?var2)
```

```
Paul fait entrer Marie
AUX_CAUS(entrer, fait), SUJ(entrer, Paul) →
                          COD_V(entrer, fait) + SUJ_V(Paul, entrer)
```

FIGURE 3 – Trois exemples de règles de projection de Dep-FTB vers PASSAGE : auxiliaire de temps (AUX-TPS), coordination (COORD) et factitif (AUX-CAUS).

car ils sont présents dans peu de règles et les annotations PASSAGE offrent la possibilité de n'avoir qu'une annotation en relations grammaticales. Néanmoins, nous avons considéré, quand cela était possible, la génération des groupes syntaxiques dans nos règles de conversion, avec l'espoir de pouvoir utiliser la couche d'annotations morpho-syntaxiques pour aider au placement correct de leurs frontières. Nous avons évalué les résultats obtenues par l'analyseur Berkeley Parser v1.0 adapté au français [7] et le convertisseur Dep-FTB sur un extrait du corpus PASSAGE (texte du parlement Européeen EP & JRC) de 1584 énoncés [8]. La Figure 4 donne une vue synoptique des différents composants de ce convertisseur, avec en clair le convertisseur Dep-FTB vers PASSAGE. Le texte est extrait de la référence de PASSAGE, analysé par Sequoia, puis converti au format PASSAGE, et aligné pour être ensuite évalué par les outils développés dans PASSAGE.

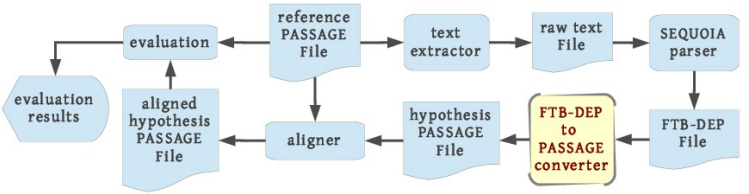
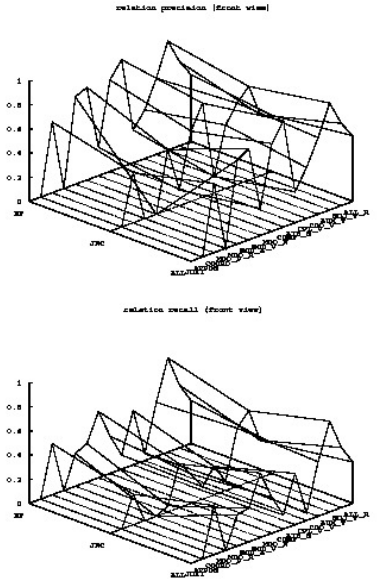


FIGURE 4 – Vue synoptique des convertisseurs DepFTB-PASSAGE.

La Figure 5 donne les mesures de performance pour les relations obtenues avec cet analyseur sur un extrait du corpus PASSAGE (texte du parlement Européeen EP & JRC) de 1584 énoncés.

7. http://alpage.inria.fr/statgram/frdep/fr_stat_dep_bky.html
8. Les corpus et les outils présentés dans cet article sont librement accessibles à l'URL http://www.limsi.fr/Individu/pap/Dep-FTB_PASSAGE.html



Relation	p	r	f
ALL	0.544	0.343	0.421
SUJ-V	0.681	0.498	0.575
AUX-V	0.885	0.752	0.813
COD-V	0.633	0.450	0.526
CPL-V	0.421	0.060	0.106
MOD-V	0.225	0.534	0.317
COMP	0.882	0.116	0.205
ATB-SO	0.256	0.426	0.320
MOD-N	0.725	0.298	0.423
MOD-A	0.756	0.277	0.406
MOD-R	0.694	0.259	0.377
MOD-P	0	0	0
COORD	0.572	0.406	0.475
APPOS	0	0	0
JUXT	0	0	0

FIGURE 5 – Mesures de précision et rappel pour les relations (comparaison en égalité stricte) obtenues avec l'analyseur Berkeley Parser v1.0 adapté au français et le convertisseur Dep-FTB sur un extrait du corpus PASSAGE (texte du parlement Européeen EP & JRC) de 1584 énoncés.

5 De PASSAGE à Dep-FTB

La conversion de PASSAGE vers Dep-FTB est plus problématique, car elle va nécessiter de résoudre des ambiguïtés pour lesquelles il faut disposer d'informations sémantiques afin de passer à des annotations plus spécifiques. L'ancrage des dépendances va aussi être problématique, car sous-spécifié dans PASSAGE au niveau des groupes syntaxiques, il est actuellement réalisé en choisissant la dernière forme du groupe dont les annotations morpho-syntaxiques sont compatibles avec la relation pour laquelle on cherche un ancrage.

Le convertisseur de PASSAGE vers Dep-FTB reprend les règles décrites précédemment ainsi que le contenu de la boîte à outils PASSAGE pour une grande part (C++). Une règle de correspondance est représentée par un ensemble d'énoncés PASSAGE dont les mots sont des formes explicites, des listes de patrons morpho-syntaxiques ou des variables. Si nous prenons comme exemple la traduction de la relation complément du verbe de PASSAGE CPL-V en Dep-FTB, elle peut produire des dépendances A_OBJ, DE_OBJ, POBJ_LOC, MOD et AFF (Candito *et al.*, 2011). Si la traduction peut utiliser les types des groupes relations et des groupes syntaxiques impliqués lorsque la relation relie plusieurs groupes syntaxiques (MOD versus COMP, ou bien la présence d'un groupe PV), dans le cas des relations CPL-V interne à un groupe syntaxique (cas des pronoms clitiques), il faudra distinger les différents cas de traduction en fonction de formes spécifiques ou de la classe sémantique des verbes (encodée dans le système actuel sous forme de listes de patron morpho-syntaxiques [9].) pour distinguer la production d'une dépendance : A_OBJ (*Paul*

9. Dans le cas où les annoations PASSAGE ne contiennent pas d'annotation morpho-syntaxiques, le convertisseur fait

y pense.), P_OBJ_LOC (*Paul y va.*), MOD (*Le chomage y est grand*) ou bien encore AFF (*Il y a 3 ans.*). À ce jour le convertisseur de PASSAGE vers Dep-FTB étant encore en phase de test, ses performances ne sont pas évaluables.

6 Conclusion

Nous avons présenté la première version d'un convertisseur bidirectionnel pour les formalismes d'annotation Dep-FTB et PASSAGE. Ce convertisseur permet de projeter la part des annotations syntaxiques qui ont un correspondant dans le formalisme cible.

La conversion de PASSAGE vers Dep-FTB est réalisée et les premières évaluations sont données. La conversion dans l'autre sens n'est pas complètement achevée, mais une première implémentation est en cours.

Le code du convertisseur ainsi que les différents corpus ayant servi aux expériences présentées dans cet article sont librement diponibles à l'url `http://www.limsi.fr/Individu/pap/Dep-FTB_PASSAGE.html`

Références

Abeillé, A., Clément, L. et Kinyon, A. (2000). Building a treebank for french. *In Proceedings of the 2nd International Conference on Language Ressources and Evaluation (LREC)*, pages 1251–1254, Athènes, Grèce.

Benarmara, F., Hatout, N., Muller, P. et Ozdowska, S., éditeurs (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

Buchholz, S. et Marsi, E. (2006). Conll-x shared task on multilingual dependency parsing. *In Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City. Association for Computational Linguistics. http ://acl.ldc.upenn.edu/W/W06/W06-2920.pdf.

Candito, M., Crabbé, B. et Falco, M. (2011). *Dépendances syntaxiques de surface pour le français - Schéma d'annotation pour un corpus en dépendances obtenu par conversion du FrenchTreebank*. http ://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html.

Candito, M., Nivre, J., Denis, P. et Anguiano, E. H. (2010). Benchmarking of statistical dependency parsers for french. *In Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, COLING '10, pages 108–116, Stroudsburg, PA, USA. Association for Computational Linguistics.

Candito, M. et Seddah, D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. *In Actes de TALN'2012*, Grenoble, France.

Carroll, J., Minnen, G. et Briscoe, T. (1999). Corpus annotation for parser evaluation. *In In Proceedings of the EACL workshop on Linguistically Interpreted Corpora (LINC)*.

Kaplan, R., Riezler, S., Tracy Holloway King, John T Maxwell, Vasserman, A. et Crouch, R. (2004). Speed and accuracy in shallow and deep stochastic parsing. *In* Susan Dumais, D. M. et

appel au TreeTagger et au lemmatiseur Flemm

ROUKOS, S., éditeurs : *HLT-NAACL 2004 : Main Proceedings*, pages 97–104, Boston, Massachusetts, USA. Association for Computational Linguistics.

MAKHOUL, J., KUBALA, F., SCHWARTZ, R. et WEISCHEDEL, R. (1999). Performance measures for information extraction. *In Proceedings of DARPA Broadcast News Workshop*, Herndnon VA.

DE LA CLERGERIE, E., HAMON, O., MOSTEFA, D., AYACHE, C., PAROUBEK, P. et VILNAT, A. (2008). PASSAGE : from French Parser Evaluation to Large Sized Treebank. *In* ELRA, éditeur : *In proceedings of the sixth international conference on Language Resources and Evaluation (LREC)*, Marrakech, Morroco.

M.-C. de MARNEFFE et C D MANNING (2008). The stanford typed dependencies representation. *In Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation at COLING 2008*, pages 1–8, Manchester. Association for Computational Linguistics.

Tracy Holloway KING, CROUCH, R., RIEZLER, S., DALRYMPLE, M. et Ronald M KAPLAN (2003). The parc 700 dependency bank. *In Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 2003*.

MIYAO, Y., NINOMIYA, T., et TSUJII, J. (2004). Corpus-oriented grammar development for acquiring a head-driven phrase structure grammar from the penn treebank. *In Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04)*, numéro 3248 de Lecture Notes in Computer Science, Hainan Island, China. Asia Federation of Natural Language Processing, Springer.

PAROUBEK, P., DE LA CLERGERIE, E., LOISEAU, S., VILNAT, A. et FRANCOPOULO, G. (2009). The PASSAGE Syntactic Representation. *In Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories*, pages 91–102, Gröningen. Netherlands Graduate Schools of Linguistics (LOT).

PAROUBEK, P., ROBBA, I., VILNAT, A. et AYACHE, C. (2006). Data, Annotations and Measures in EASY - the Evaluation Campaign for Parsers of French. *In proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*, pages 315–320, Genoa, Italy. ELRA.

SAGAE, K., MIYAO, Y., MATSUZAKI, T., et TSUJII, J. (2008). Challenges in mapping of syntactic representations for framework-independent parser evaluation. *In Proceedings of the Workshop on Automated Syntatic Annotations for Interoperable Language Resources at the First International Conference on Global Interoperability for Language Resources (ICGL'08)*, Hong-Kong.

VILNAT, A., PAROUBEK, P., de la CLERGERIE, E. V., FRANCOPOULO, G. et GUÉNOT, M.-L. (2010). PASSAGE Syntactic Representation : a Minimal Common Ground for Evaluation. *In* CHAIR), N. C. C., CHOUKRI, K., MAEGAARD, B., MARIANI, J., ODIJK, J., PIPERIDIS, S., ROSNER, M. et TAPIAS, D., éditeurs : *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).