TSNLP – Des jeux de phrases-test pour l'évaluation d'applications dans le domaine du TALN

Sabine Lehmann
ISSCO
Université de Genève
54, rte des Acacias
CH 1227 Genève
+41 - 22 - 705 79 33
Sabine.Lehmann@issco.unige.ch

Dominique Estival
Department of Linguistics
University of Melbourne
Parkville, Victoria 3052
Australia
+61 - 3 - 344 42 27
D.Estival@linguistics.unimelb.edu.au

Stephan Oepen
DFKI GmbH
CL Department
Stuhlsatzenhausweg 3
D 66123 Saarbrücken
+49 - 681 - 302 52 82
oe@dfki.uni-sb.de

1 Résumé

Le nombre d'applications dans le domaine du TALN n'a cessé d'augmenter lors de ces dernières années. Ce développement va de pair avec une demande croissante d'outils pour évaluer ces applications. Le projet TSNLP répond à cette demande en proposant une méthodologie et des outils pour l'évaluation à l'aide de jeux de phrases-test.

Mis à part une méthodologie élaborée pour la construction de matériel de test, TSNLP a créé la plus grand base de données de jeux de phrases-test actuellement disponible pour le français, l'anglais et l'allemand. En outre, ce projet a développé des outils qui facilitent la construction, le stockage et l'accès aux données.

Les résultats de TSNLP seront publiques. Le projet propose ainsi des ressources linguistiques qui pourraient devenir une proposition de standard pour un modèle d'évaluation pour tout utilisateur d'applications dans le domaine du TALN.

2 Introduction

Les jeux de phrases-test et les corpus de texte représentent les deux méthodes les plus utilisées pour l'évaluation des applications en TALN. Elles peuvent être considérées comme complémentaires: l'intérêt des corpus est que leurs données sont "naturelles" tandis que l'avantage des jeux de phrases-test réside dans leur caractère systématique et exhaustif. De plus, les phrases-test incluent aussi des données agrammaticales qui s'avèrent être indispensables pour l'évaluation diagnostique.

L'objectif principal du projet TSNLP[1] est de créer une méthodologie et des outils pour l'évaluation des applications en TALN avec des jeux de phrases-test.

[1]Le projet TSNLP (Test Suites for Natural Language Processing) est un projet relevant du programme LRE (Language Research and Engineering) de la Communauté européenne. Il a commencé en décembre 1993 et s'est terminé en mars 1996. Toutes les informations concernant le projet sont accessibles sur WWW (http://tsnlp.dfki.uni-sb.de/tsnlp/).

Cet article se base sur les résultats de recherche obtenus par l'ensemble du consortium. Nous remercions donc tous nos partenaires: Lorna Balkan, Frederik Fouvry et Doug Arnold de l'Université d'Essex (Colchester, Royaume-Uni); Klaus Netter, Judith Klein et Judith Baur du DFKI (Saarbrücken, Allemagne); Kirsten Falkedal du GMS (Berlin, Allemagne) – lors du projet, à l'ISSCO (Genève, Suisse); Sylvie Regnier-Prost, Eva Dauphin et Veronika

Les buts du projet peuvent être décrits de manière plus spécifique:

- l'élaboration d'une méthodologie pour la construction systématique et progressive des phrases-test;
- le développement d'un schéma d'annotations détaillé, et neutre par rapport à des théories linguistiques spécifiques et par rapport à des types d'applications particuliers.
- la construction d'un nombre important de phrases-test qui peuvent être utilisées pour différents types d'applications, à savoir des analyseurs, des correcteurs de grammaire et des correcteurs de langue contrôlée.
- la création d'outils qui facilitent la construction du matériel de test ainsi que l'élaboration et l'implémentation d'une base de données extensible pour le stockage, l'accès et la manipulation des données.

3 Les motivations

Quoique les phrases-test soient employées pour l'évaluation d'applications, il n'existe pas de méthodologie générale ni de directives pour construire des phrases-test. En outre, il n'y a pas de matériel de test adéquat disponible: une étude concernant les phrases-test existantes – réalisée au début du projet ([Estival et al., 1994a]) – a en effet révélé que la réutilisation des phrases-test est rendue difficile faute d'annotations ou d'informations supplémentaires suffisantes. La plupart de ces phrases-test ont été conçues pour des applications spécifiques et ne tiennent compte que des besoins particuliers de l'utilisateur ou du développeur de ces applications. Ces considérations représentent les motivations principales qui sont à la base du projet.

Idéalement, une phrase-test devrait donc être réutilisable tout en restant spécifique. C'est pourquoi TSNLP a abandonné le concept traditionnel selon lequel un jeu de phrases-test représente une liste monolithique de phrases-test afin de le remplacer par le concept d'une base de données contenant des phrases-test associées à un grand nombre d'annotations. Ainsi, il est possible d'extraire des sous-ensembles de jeux de phrases-test à partir d'annotations spécifiques pour une situation de test particulière. La base de données peut être considérée comme un "jeu de phrases-test virtuel" qui permet à un utilisateur de construire ses propres jeux de phrases-test.

4 Le schéma d'annotations

L'élaboration détaillée du schéma d'annotations s'est basée sur l'étude préliminaire portant sur les jeux de phrases-test existants ([Estival et al., 1994a]). Les annotations ne se réfèrent ni à une théorie linguistique spécifique, ni à un type d'application particulier ou encore à une situation d'évaluation spéciale ([Estival et al., 1994b]).

Les annotations et les données des phrases-test sont réparties sur quatre niveaux de représentation distincts.

- Les données centrales

  Ce niveau comprend la phrase-test ainsi que toute les informations indépendantes du phénomène linguistique ou de l'application à tester. Mis à part la phrase-test même, les

Lux de l'Aerospatiale (Suresnes, France).

annotations incluent (i) le nom de l'auteur, la date, le numéro d'identification, (ii) l'origine, le registre, la catégorie syntaxique (syntagme, phrase ou même paragraphe), ainsi que la grammaticalité de la phrase-test, (iii) les catégories (morpho-)syntaxiques des constituants principaux, leur position dans la phrase et (iv) une représentation (sous-spécifiée) de la structure fonctionnelle.

- Les données concernant le phénomène linguistique

  Ce niveau comprend le nom du phénomène linguistique, son super-type – dans la hiérarchie des phénomènes linguistiques (voir section 5) – et son numéro d'identification. L'interaction avec d'autres phénomènes est également indiquée. En outre, les annotations décrivent les paramètres (morpho-)syntaxiques qui sont caractéristiques pour un phénomène particulier (p.ex. le nombre et le type de compléments pour la complémentation) et qui permettent de dériver les données agrammaticales des données grammaticales. Une phrase-test particulière peut être attribuée à un ou plusieurs phénomènes.[2]

- Les séries de test

  Les données peuvent être regroupées dans des séries de test. Une série de test comprend généralement une donnée grammaticale (ou positive) et une ou plusieurs données agrammaticales (ou négatives). La relation entre phrases-test positives et négatives est l'un des problèmes les plus difficiles pour la construction des données. La construction des données négatives se fait à partir de la variation systématique des paramètres, définis pour chaque phénomène (voir section 5).

- Les paramètres concernant l'utilisateur et l'application

  Les annotations incluent des informations sur l'emploi de la phrase-test en ce qui concerne l'utilisateur et l'application qui est évaluée. Elles indiquent par exemple si la donnée en question est pertinente pour une application spécifique. Ce niveau est particulièrement important pour le processus d'adaptation *(customisation)*. Les utilisateurs de phrases-test sont invités à utiliser cet espace afin de le personnaliser en y intégrant les informations qui se rapportent à leurs applications et à leurs besoins spécifiques.

Les quatre niveaux de représentation comprennent également de l'espace pour ajouter des commentaires supplémentaires.

5 La construction des phrases-test

La couverture linguistique des jeux de phrases-test dans la base de données actuelle est l'intersection des phénomènes considérés comme cruciaux pour les applications de TALN dans nos trois langues.

Voici la liste des phénomènes sélectionnés:

- la complémentation;
- l'accord;
- la modification;

[2] Ainsi "*Il vient.*" peut servir de phrase-test pour la complémentation (`C_Complementation` dans la notation de TSNLP) ou pour l'accord (`C_Agreement`). Les données agrammaticales dérivées de cette phrase-test diffèrent cependant: "**Il le vient.*" correspond par exemple à une phrase-test négative pour la complémentation, tandis qu'une des donnes ngatives possibles pour l'accord, est "**Il viens.*".

- la diathèse;
- le mode, le temps et l'aspect;
- les types des phrases (interrogative, exclamative);
- la coordination;
- la négation;
- les phénomènes extra-grammaticaux (e.g. ponctuation, abréviations, etc.).

Le tableau suivant indique le nombre de phrases-test qui ont été construites à ce jour dans le cadre du projet:

Phenomenon	**English**	**French**	**German**	**Total**
C_Complementation	148 | 863	188 | 567	218 | 246	554 | 1676
C_Agreement	68 | 55	104 | 183	224 | 175	396 | 413
C_Modification		329 | 63		329 | 63
NP_Complementation	10 | 27	12 | 28		22 | 55
NP_Agreement	201 | 995	272 | 1082	299 | 1732	772 | 3809
NP_Modification	301 | 484		53 | 60	354 | 544
Diathesis	157 | 124	176 | 119	147 | 148	480 | 391
Tense Aspect Modality	157 | 39	77 | 275	186 | 134	420 | 448
Sentence Types	80 | 100	389 | 387	105 | 14	574 | 501
Coordination	147 | 186	379 | 319	105 | 429	631 | 934
Negation	289 | 129	68 | 100	82 | 210	439 | 439
Word Order		7 | 7	60 | 160	67 | 167
Extragrammatical	24 | 34		253 | 0	277 | 34
Total	1582 | 3036	2001 | 3130	1732 | 3308	5315 | 9474

Les phénomènes sont classés selon leur domaine, c'est-à-dire selon la catégorie syntaxique dans laquelle ils se produisent.

Il s'agit plus précisément de la phrase (S), la proposition (C), le syntagme nominal (NP), le syntagme adjectival, le syntagme prépositionnel (PP) et le syntagme adverbial (AdvP). Le nom du phénomène est précédé par son domaine (p.ex. C_Complementation, NP_Agreement). La plupart des phénomènes présentent aussi des sous-classes (voir l'exemple ci-dessous).

Pour renforcer la cohérence entre les annotations utilisées pour les trois langues, une liste canonique des catégories et des fonctions a été établie. Cette liste se veut très générale; elle essaie donc d'éviter toute référence à des particularités présentées par telle ou telle théories linguistique.

Le tableau suivant présente un exemple d'une phrase-test pour la C_Complementation avec les annotations concernant le niveau de la phrase-test même, l'analyse, le phénomène ainsi que les paramètres. L'identification des paramètres – qui caractérisent les propriétés d'un phénomène – s'est souvent révélée très difficile, ce qui explique qu'ils ne soient pas toujours définis de manière très détaillée.

Ils occupent cependant un rôle essentiel puisque la dérivation des phrases-test agrammaticales se base sur leur variation (pour une description plus détaillée, voir [Lehmann et al., 1996]).

Test Item				
item id: *24100101*		author: *issco*		date: *jan-95*
register: *formal*		format: *none*		origin: *invented*
difficulty: *1*		wellformedness: *1*		category: *S*
input: *L' entreprise suppose que l' ingénieur vient .*				length: *7*
comment:				
position	instance	category	function	domain
0:2	*L' entreprise*	*NP_sg-fem*	*subj*	*2:3*
2:3	*suppose*	*V*	*func*	*0:7*
3:7	*que l'ingénieur vient*	*C_cpl(que)*	*s-comp*	*2:3*

Phenomenon			
phenomenon id: *2410*	author: *issco*	date: *jan-95*	
name: *C_Complementation_subj(NP)_V_s-comp(C-cpl)*			
supertypes: *C_Complementation*			
presupposition: *C_Agreement, C_Wordorder, NP_Agreement, NP_Wordorder*			
restrictions: *neutral*	interactions: *none*	purpose: *test*	
comment: *Transitive verbs with subordinate clause (valency:2)*			
position	attribute	value	instance
	subcat	*2(subj, s-comp(C_cpl(que)))*	
0:2	*comp*	*subj*	*L'entreprise*
3:7	*comp*	*s-comp(C_cpl(que))*	*que l'ingénieur vient*

6 La technologie des phrases-test

La construction de matériel de test ainsi que son adaptation à des applications en TALN est un travail très laborieux qui, de surplus, est particulièrement sujet à erreur. C'est pourquoi le développement d'outils facilitant la construction et l'utilisation des phrases-test a joué un rôle prépondérant lors du projet. Cette section n'en résume que les aspects les plus importants (pour une vue d'ensemble se référer à [Oepen et al., 1996]).

6.1 La construction de phrases-test

Afin de faciliter la construction de phrases-test, un outil à interface graphique ("test suite construction tool" – `tsct`) a été implémenté. Il se base sur le schéma d'annotations présenté dans la section 4 et procède à une vérification (limitée) des valeurs pour les attributs. `Tsct` permet également de réutiliser des données déjà annotées, ce qui constitue une aide considérable.

En outre, un outil de génération de phrases-test a été développé en cours de projet. Basé sur une grammaire à clauses définies [Arnold et al., 1994]), il génère des phrases-test grammaticales et agrammaticales de manière systématique, et peut être utilisé pour la construction de phrases-test spécifiques.

6.2 La maintenance et l'interrogation de phrases-test

Le consortium du projet a décidé d'implémenter deux bases de données (i) afin de satisfaire aux différents besoins des développeurs et des utilisateurs dans le domaine du TALN, et (ii) pour fournir des interfaces qui d'une part conviennent à l'utilisateur "humain", et d'autre part, soient accessibles aux programmes d'applications externes.

La première base de donnée (`tsdb1`: test suite database 1) a été implémentée en `ANSI C` ce qui a permis de peaufiner le langage d'interrogation et l'interface. La deuxième (`tsdb2`) se base par contre sur une base de données commerciale, à savoir Microsoft FoxPro, disponible pour Apple Macintosh et les PC sous MS Windows.[3] Elle correspond aux standards courants de l'industrie et permet donc à tout utilisateur (industriel ou universitaire) d'avoir accès aux jeux de phrases-test de TSNLP.

6.3 Un exemple d'une interrogation et de l'extraction

L'interpréteur de commandes de `tsdb1` se base sur `ASCII` et utilise un langage d'interrogation comparable au style `SQL` simplifié. Voici un exemple d'interrogation précédé par la paraphrase française correspondante:[4]

"Donnez-moi toutes les phrases-test qui sont associées au phénomène C_Complementation et qui contiennent un objet pronominal":

```
select i-input
  where i-wf = 1 & p-name = "C_Complementation" &
        a-function = "obj" & a-category ~ "^PRON"
```

7 Evaluation des phrases-test

La validité du schéma d'annotations et des phrases-test a été évaluée pour trois types d'application, à savoir un correcteur d'orthographe pour le français (*Le Correcteur 101*), un correcteur de langue controllée pour l'anglais (*SECC*) et un analyseur pour l'allemand (le système HPSG développé au DFKI). Il s'agissait donc d'évaluations dans des situations en "boîte noire" et en "boîte transparente" [King & Falkedal, 1990].

Sans pouvoir entrer dans les détails, l'on peut dire que cette évaluation a fourni des résultats précieux qui ont permis d'améliorer la qualité des phrases-test (voir [Dauphin et al., 1995]). Si la situation d'évaluation en "boîte noire" a contribué fortement à identifier les défauts des applications, les tests réalisés sur l'analyseur allemand ont démontré que le matériel de test développé dans le cadre de TSNLP représentait un outil tout à fait adéquat pour l'évaluation diagnostique en situation de "boîte transparente".

8 Conclusion

Le projet TSNLP a développé des ressources linguistiques importantes pour l'évaluation diagnostique des applications dans le domaine du TALN. TSNLP dispose en effet de la plus grande base de données de jeux de phrases-test pour l'anglais, le français et l'allemand. Les phrases-test ont été construites et annotées selon une méthodologie systématique. Elles ont été intégrées dans deux bases de données différentes après avoir été évaluées à l'aide de trois applications.

[3] Une base de données supplémentaire qui se base sur MS Access est en train d'être développée: elle est comparable à `tsdb2` et sera également disponible pour les ordinateurs qui permettent d'exécuter MS Windows.

[4] D'autres exemples d'interrogation et de l'information supplémentaire concernant `tsdb1` sont accessibles sur la page d'accueil du WWW (http://tsnlp.dfki.uni-sb.de/tsnlp/).

Même si les résultats de ce projet constituent une innovation importante dans le domaine de l'évaluation des applications en TALN, il est évident que ce travail doit être poursuivi afin d'assurer son utilisation par la communauté du TALN. Les développements prévus comprennent l'extension du matériel de test, la création d'outils pour adapter les données à des applications et des domaines spécifiques ou encore l'élaboration de méthodes qui permettent une mise en relation entre phrases-test et corpus afin de déterminer la fréquence et la pertinence des données.

References

[Adriaens, 1994] Geerd Adriaens. 1994. SECC: Simplified English Checker and Style Correction in an MT Framework. In *Proceedings of the Language Engineering Convention*, Paris.

[Arnold et al., 1994] Doug Arnold, Martin Rondell, and Frederik Fouvry. 1994. Design and Implementation of Test Suite Tools. Report to LRE 62-089 D-WP5.1, University of Essex, UK.

[Balkan et al., 1996] Balkan, Lorna, Frederik Fouvry, and Sylvie Regnier-Prost (editors). 1996. TSNLP User Manual. Volume 1: Background, Methodology, Customization, and Testing. Technical report. University of Essex, UK.

[Dauphin et al., 1995] Eva Dauphin, Veronika Lux, Sylvie Regnier-Prost (principal authors), Doug Arnold, Lorna Balkan, Frederik Fouvry, Judith Klein, Klaus Netter, Stephan Oepen, Dominique Estival, Kirsten Falkedal, and Sabine Lehmann. 1995. Testing and Customisation of Test Items. Report to LRE 62-089 D-WP4, University of Essex, UK.

[Estival et al., 1994a] Dominique Estival, Kirsten Falkedal, Lorna Balkan, Siety Meijer, Sylvie Regnier-Prost, Klaus Netter, and Stephan Oepen. 1994. Survey of Existing Test Suites. Report to LRE 62-089 D-WP1, University of Essex, UK.

[Estival et al., 1994b] Dominique Estival, Kirsten Falkedal, Sabine Lehmann, Lorna Balkan, Siety Meijer, Sylvie Regnier-Prost, Klaus Netter, and Stephan Oepen. 1994. Test Suite Design — Annotation Scheme. Report to LRE 62-089 D-WP2.2, University of Essex, UK.

[Flickinger et al., 1987] Daniel Flickinger, John Nerbonne, Ivan A. Sag, and Thomas Wassow. 1987. Toward Evaluation of NLP Systems. Technical report, Hewlett-Packard Laboratories. Distributed at the 24^{th} Annual Meeting of the Association for Computational Linguistics (ACL).

[King & Falkedal, 1990] Margaret King and Kirsten Falkedal. 1990. Using Test Suites in the Evaluation of Machine Translation Systems. *Proceedings of the 13th International Conference on Computational Linguistics (COLING)*, Helsinki.

[Lehmann et al., 1996] Lehmann, Sabine, Dominique Estival, Kirsten Falkedal, Hervé Compagnion, Lorna Balkan, Frederik Fouvry, Judith Baur and Judith Klein. 1996. TSNLP User Manual. Volume 3: Test Data Documentation. Technical report. Istituto Dalle Molle per gli Studii Semantici e Cognitivi (ISSCO) Geneva, Switzerland.

[Nerbonne et al., 1993] John Nerbonne, Klaus Netter, Kader Diagne, Ludwig Dickmann, and Judith Klein. 1993. A Diagnostic Tool for German Syntax. *Machine Translation*, 8:85–107.

[Oepen et al., 1996] Oepen, Stephan, Frederik Fouvry, Klaus Netter, Tom Fettig, and Fred Oberhauser. 1996. TSNLP User Manual. Volume 2: Core Test Suite Technology. Technical report. Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) Saarbrücken, Germany.