Report on NLP-based CALL Workshop

Sylvianne Granger, Georges Antoniadis, Cédrick Fairon, Julia Medori, Virginie Zampa

D39.3.1 (Final)

Report on NLP-based CALL Workshop

Main author : Sylviane Granger (UCLOUVAIN)

Nature of the deliverable : Report

Dissemination level : Public

Planned delivery date : June 2006

Prepared for the European Commission, DG INFSO, under contract N°. IST 507838
as a deliverable from WP39
Submitted on 28-06-2006

Summary

Report on NLP-based Computer-Assisted Language Learning workshop

History

Filename	Status	Release	Changes	Uploaded
D39-03-01-F.pdf	Final	1		28/06/2006

Contributor(s)

Name, First name	Contractor
Antoniadis Georges	(USTENDHAL)
Fairon Cédrick	(UCLOUVAIN)
Granger Sylviane	(UCLOUVAIN)
Medori Julia	(UCLOUVAIN)
Zampa Virginie	(USTENDHAL)

Report on the Workshop on NLP-based CALL

Table of Contents:

1. Contribution of the JEIRP to the Kaleidoscope network

The main objective of the JEIRP (WP39: Digital Language Learning: An Integrated Perspective) is to integrate existing research on the use of digital technologies in language learning and teaching and to develop new tools and methodologies that bridge the gap between two fields that should have a lot to contribute to each other but unfortunately fail to communicate: Natural Language Processing and Computer-assisted Language Learning. Natural Language Processing (NLP) is a branch of artificial intelligence that deals with analyzing, understanding and generating language that humans use naturally. It embraces a variety of computational techniques that perform tasks on natural language machine-readable free text, such as translation, summarization, information extraction, etc. Computer-assisted Language Learning (CALL) may be defined as "the search for and study of applications of the computer in language teaching and learning" (Levy 1997). It embraces all the computerized activities that aid to the presentation, reinforcement and assessment of learning material and learners' production. Although the integration of NLP into CALL systems would seem like a natural step in the evolution of CALL, NLP-based applications are not very widespread. According to Kraif et al (2004) this is due to three main reasons: "NLP techniques often lack reliability, NLP products and resources are quite expensive and difficult to implement, and the end-users (teachers, learners, conceptors, editors...) are not aware of NLP possibilities".

The lack of communication between NLP and CALL is primarily due to a divergence between computer scientists and didactics experts regarding the notion of "language". For Information Technology specialists, it is considered as a sequence of codes, while for didactics experts, it is a system of forms and concepts. Currently, most language learning software uses simple pattern-matching algorithms with no linguistic analysis. This leads to errors such as the example shown in Antoniadis & Ponton 2004, where a student answers 'la casa' with two spaces but his/her answer is not accepted for the expected answer is 'la casa' with only one space. More importantly, the reliance on crude pattern-matching means that very simple but crucial processes, such as the disambiguation of 'open' as an adjective and 'open' as a verb, or the grouping of all inflected forms of a verb (go, goes, going, gone, went), are out of bounds for CALL.

The only way to deal with these issues is to integrate NLP into CALLware. As pointed out by Chanier 1998, Brun et al 2002, Antoniadis & Ponton 2002, Antoniadis & Ponton 2004 and Antoniadis & Chanier 2005, only the use of NLP methods makes it possible to consider and process language as a system of forms and concepts.

The JEIRP aims to contribute to the integration of NLP into CALL by focusing more particularly on:
* the input to the language learning process: NLP tools and techniques are used to enrich the texts used as input for the language learning process by means of reliable NLP techniques (mainly lemmatization and grammatical tagging);
* the output of the language learning process: NLP techniques are used to provide intelligent and adapted feedback on learners' productions.

It brings together an interdisciplinary team of researchers that have acquired considerable expertise in the NLP, CALL and/or language didactics.

The main two partners are:

* the Université catholique de Louvain (Belgium) with the combined expertise of two research labs: the CENTAL (Centre for Natural Language Processing) and the CECL (Centre for English Corpus Linguistics). The UCL is the coordinator of the JEIRP.
* the Université Stendhal – Grenoble 3 (France) which provides its expertise in CALL with the LIDILEM and in language teaching with the Centre Universitaire d'Etudes Françaises (Centre for French studies).

working in collaboration with two research teams which are not funded by the Kaleidoscope network:

* the UWiLL (Ubiquitous Web-based Intelligent Language Learning) team at Tamkang University (Taiwan).
* the Department of French as a Foreign Language at York University (Canada).

2. Object of this deliverable

This deliverable is a report on the one-day workshop on NLP-based CALL that was organized on 13 April 2006 by the JEIRP partners as a contribution to an increased visibility of the project and the Kaleidoscope network. The workshop's bilingual title was:

"Integrating Natural Language Processing and Computer-Assisted Language Learning: current status and future prospects"
"Traitement automatique des langues et Apprentissage des langues assisté par ordinateur : quelles perspectives d'intégration?"

The workshop was organized within the framework of the 13th edition of the *Traitement Automatique du Langage Naturel (TALN)* conference (http://www.taln.be/), a yearly conference organized under the aegis of the French association for NLP called *ATALA (Association pour le Traitement Automatique des Langues).* The *TALN 2006* conference was held in Leuven (Belgium) from 10-13 April 2006. It gathered more than 150 researchers from 14 different countries.

The Louvain and Grenoble teams who organized the workshop took care to ensure international dissemination of the event by posting the call for papers on a large number of websites and international discussion lists. To ensure high scientific quality, they put together a scientific committee composed of a large number of international experts:

Dominique Abry, CUEF, Université Stendhal-Grenoble 3, France
Georges Antoniadis, LIDILEM, Université Stendhal-Grenoble 3, France
Aimé Avolonto, York University, Canada
Lars Borin, Göteborg Univesity, Sweden
Jill Burstein, Educational Testing Service, NJ, USA
Carol Chapelle, Iowa State University, USA

Cristelle Cavalla, CUEF, Université Stendhal-Grenoble 3, France
Cédrick Fairon, CENTAL, UCLouvain, Belgium
Sylviane Granger, CECL, UCLouvain, Belgium
Marie-Josée Hamel, Dalhousie University, Canada
Trude Heift, Simon Fraser University, Canada
Olivier Kraif, LIDILEM, Université Stendhal-Grenoble 3, France
Claudia Leacock, Pearson Knowledge Technologies, CO, USA
Thomas Lebarbé, LIDILEM, Université Stendhal-Grenoble 3, France
Fanny Meunier, CECL, UCLouvain, Belgium
Claude Ponton, LIDILEM, Université Stendhal-Grenoble 3, France
Frédérique Segond, Xerox XRCE, France
Cornelia Tschichold, University of Wales Swansea, United Kingdom
Serge Verlinde, ILT, KULeuven, Belgium
David Wible, Tamkang University, Taiwan

Each of the contributions received was reviewed by two members of the scientific committee. In the end, seven contributions were selected for presentation at the workshop and inclusion in the conference proceedings, which also included an introduction to the field by the organizers (G. Antoniadis, C. Fairon, S. Granger, J. Medori & V. Zampa. Title: *Quelles machines pour enseigner la langue?*).

Upon reception of the revised versions of the contributions, the Louvain and Grenoble teams edited the papers for inclusion in the conference proceedings, in which the workshop proceedings were allotted a specific section (cf. **Appendix 1**).

3. General outline of the workshop

The workshop started with an introduction to the field by the organizers (G. Antoniadis, C. Fairon and S. Granger, J. Medori, V. Zampa). The first part of the introduction was devoted to a presentation of the Kaleidoscope Network and the JEIRP. A copy of the Powerpoint presentation is enclosed in **Appendix 2**. In the introduction we also introduced the JEIRP's website, which was created in order to increase the visibility of the field of Integrated Digital Language Learning (IDILL)

(URL: www.idill.org) and act as a portal on the use of NLP in Computer-Assisted Language Learning. It provides a select bibliography, a list of events and links to other websites dealing with the same subject.

After this introduction came the workshop presentations. The titles of the presentations and their authors are listed below:

* *Intelligent CALL*: Cornelia Tschichold (University of Wales Swansea, Great-Britain).

* *The use of NLP tools for Basque in a multiple user CALL environment and its feedback*: Itziar Aldabe, Bertol Arrieta, Arantza Díaz de Ilarraza, Montse Maritxalar, Ianire Niebla, Maite Oronoz, Larraitz Uria (University of the Basque Country, Department of Computer Languages and Systems, Spain).

* *Le TALN au service de la didactique du français langue étrangère écrit* : Isabelle Audras, Jean-Gabriel Ganascia (Université Pierre et Marie Curie – LIP 6, France)

* *Automated Analysis of Students' Free-text Answers for Computer-Assisted Assessment*: Matthieu Hermet (University of Ottawa, School of Information Technology and Engineering), Stan Szpakowicz (Polish Academy of Sciences, Institute of Computer Science, Polan), Lise Duquette (University of Ottawa, Second Language Institute, Canada).

* *Capitalisation d'une ressource en or : le dictionnaire* : Michael Zock (Laboratoire d'Informatique Fondamentale (LIF) – CNRS, France)

* TAEMA : Traitement Automatique de l'Ecriture de Mots Affectifs : Pierre-André Buvet, Fabrice Issac (Université Paris 13 – Laboratoire de Linguistique Informatique, CNRS, France)

* *A Computational Approach to the Discovery and Representation of Lexical Chunks* : David Wible (Department of English, Tamkang University, Taipei, Taiwan), Chin-Hwa Kuo (Computer and Network Lab, Tamkang University, Taipei, Taiwan)

Meng-Chang Chen, Nai-Lung Tsao, Tsung-Fu Hung (Institute of Information Science, Academia Sinica, Taipei, Taiwan)

The workshop also included a demo of *ALFALEX*, a learning environment that relies on a large electronic dictionary of French for the automatic generation of lexical exercises (S. Verlinde, Katholieke Universiteit Leuven, Belgium).

The presentations demonstrated that NLP clearly has a legitimate place in CALL but that much remained to be done, in particular to ensure teacher- and learner-friendly didactic implementation. Actual didactic integration and validation are often the 'parents pauvres' of NLP-based CALL projects.

While the workshop only tackled some of the main issues involved in the integration of NLP into CALL, one major restriction being that all the papers focused on the treatment of written data, it made it possible to bring out some key areas which clearly need to figure prominently on the future agenda of NLP-based CALL research:

1. **Need for close collaboration between commercial companies, NLP specialists and language teachers**

One of the main challenges facing the use of NLP in CALL is a lack of communication between the people involved in developing CALL software, NLP specialists and teachers. This is a pity as one partner's weakness proves to be the other's strength and much would be gained from closer collaboration. Commercial systems know how to make eye-catching products with an attractive interface and cutting-edge multimedia content, but often lack sound pedagogical principles. NLP specialists have the necessary expertise to add NLP modules to CALL systems but research often leads to prototypes rather than fully-grown software systems that teachers can use. Language teachers, on the other hand, have the pedagogical knowledge based on their teaching experience and the pedagogical contents they develop are therefore targeted and well-suited for their students. However, the systems they implement are rarely up-to-date with the latest technologic and linguistic developments. Although they have complementary expertise, these partners seldom communicate. The future therefore clearly lies in interdisciplinary projects.

2. Integration of electronic dictionaries

A second major thread running through the workshop was the integration of electronic dictionaries into learning environments. This development can be related to the growing role of the lexicon in foreign language learning (cf. Lewis's lexical approach 1993 & 1997). Innovative applications take learner needs into account, notably the fact that learners often have an imperfect knowledge of the word they are looking for. NLP techniques relying on phonetic, orthographic or semantic approximation can help him/her identify the correct entry in the dictionary. Another interesting development automatically generates exercises from the electronic dictionary. Also of interest are applications which aim at aiding both comprehension and production. Electronic dictionaries which contain syntactic and semantic information (hyperonyms, hyponyms, etc.) and specify the constraints on the arguments (e.g. subject: human) prove to be highly effective writing tools.

3. Learner errors: detection, correction and feedback

A third major issue addressed at the workshop was the detection and correction of learner errors, a very popular domain at the moment as demonstrated by a recent issue of the CALICO journal which is entirely dedicated to this subject (Heift et Schulze, 2003). The main challenge for NLP-based CALL is to evaluate and provide feedback to learners' answers to open questions, i.e. questions that allow learners to write free text. Intelligent diagnostic tools would allow CALL to move beyond the fill-in-the-blanks exercises or multiple choice questions that are its staple diet today.

References

ANTONIADIS G., PONTON C. (2002). Le TAL: une nouvelle voie pour l'apprentissage des langues. UNTELE'2002, Compiègne, France.

ANTONIADIS G., PONTON C. (2004). Mirto : un système au service de l'enseignement des langues, UNTELE'2004, Compiègne, France.

ANTONIADIS G., ECHINARD S., KRAIF O., LEBARBÉ Th., LOISEAU M., PONTON C. (2004). NLP-based scripting for CALL activities. eLearning for Computational Linguistics and Computational Linguistics for eLearning. International Workshop in Association with COLING 2004. 28 août 2004, Genève (Suisse)

ANTONIADIS G., CHANIER, T. (eds.) (2005). *ALSIC* 8 (2), numéro thématique « TAL et apprentissage des langues ».

ANTONIADIS G., ECHINARD S., KRAIF O., LEBARBÉ Th., PONTON C. (2005). « Modélisation de l'intégration de ressources TAL pour l'Apprentissage des Langues : la plateforme MIRTO ». In *ALSIC* 8 : 65-79

BORIN L. (2002). « What have you done for me lately? The fickle alignment of NLP and CALL ». Reports from Uppsala Learning Lab.

BRUN C., PARMENTIER T., SANDOR A., SEGOND F. (2002). « Les outils de TAL au service de la e-formation en langues », *Multinguisme et traitement de l'information.* F. Segond, ed., pages 223-250, Hermès Science Publications, Paris, France.

CHANIER T. (1998). Relations entre le TAL et l'ALAO ou l'ALAO un « simple » domaine d'application du TAL ? International conference on natural language processing and industrial application (NLP+IA'98), Moncton, Canada.

HEIFT T., SCHULZE M. (eds) (2003). Special issue of *CALICO* on Error Analysis and Error Correction in Computer-Assisted Language Learning 20 (3).

KRAIF O., ANTONIADIS G., ECHINARD S., LOISEAU M., LEBARBE T., PONTON C. (2004) « NLP Tools for CALL : the Simpler, the Better ». In Proceedings of InSTIL/ICALL2004 – NLP and Speech Technologies in Advanced Language Learning Systems – Venice 17-19 June, 2004.

LEVY M. (1997) *CALL: Context and Conceptualisation*, Oxford: Oxford University Press.

LEWIS M. (1993). *The Lexical Approach: The State of ELT and a Way Forward*. Language Teaching Publications, Hove.

LEWIS M. (1997). *Implementing the Lexical Approach: Putting Theory into Practice.* Language Teaching Publications, Hove.

Appendix 1: Workshop proceedings

Integrating Natural Language Processing and Computer-Assisted Language Learning : current status and future prospects

Workshop organised on April, 13th 2006 within the Framework of TALN 2006

Antoniadis, G., Fairon, C. & Granger, S. (eds)

Centre de traitement automatique du langage (UCLouvain)

Centre for English Corpus Linguistics (UCLouvain)

Laboratoire de Linguistique et Didactique des Langues Etrangères et Maternelles (Université Stendhal Grenoble)

The papers hereafter were published in :

Mertens Piet, Fairon Cédrick, Dister Anne, Watrin Patrick (eds). *Verbum ex machina.* In Proceedings of the 13th Conference *Traitement automatique des langues naturelles.* Louvain-la-Neuve: Presses universitaires de Louvain : 793-875 (Cahiers du Cental, 2).

ATELIER « Traitement automatique des langues et apprentissage des langues assisté par ordinateur : quelles perspectives d'intégration ? »

Quelles machines pour enseigner la langue ?

Georges Antoniadis[1], Cédrick Fairon[2], Sylviane Granger[3], Julia Medori[2-3], Virginie Zampa[1]

[1] Université Stendhal de Grenoble, Laboratoire LIDILEM
{Georges.Antoniadis ; Virginie.Zampa}@u-grenoble3.fr

[2] Université catholique de Louvain, Centre de Traitement Automatique du Langage
{fairon,medori}@tedm.ucl.ac.be

[3] Université catholique de Louvain, Centre for English Corpus Linguistics
granger@lige.ucl.ac.be

Résumé

Cet article présente dans un premier temps l'historique de l'enseignement assisté par ordinateur (EAO) en situant ses origines aux années 1920 avec les premières machines à enseigner mécaniques. L'arrivée de l'ordinateur a par la suite permis de proposer à l'apprenant de langues différents types d'activités : tâches de compréhension, simulations, etc. Cependant, celles-ci ont des limites qui ne peuvent être surmontées sans l'apport du traitement automatique des langues (TAL). Nous présentons ici la problématique de l'intégration du TAL aux systèmes d'ALAO en dressant un bilan des défis que cette intégration doit aujourd'hui relever et nous faisons une synthèse des présentations de l'atelier. Celles-ci proposent des problématiques diverses allant de la détection et la correction d'erreurs à l'enrichissement de dictionnaires électroniques en passant par la mise en œuvre d'outils complets d'aide à l'apprentissage des langues. Nous verrons que la clé de l'intégration du TAL dans l'ALAO réside dans le travail pluridisciplinaire entre informaticiens, didacticiens des langues et spécialistes du TAL.

Mots-clés : TAL, ALAO, systèmes d'apprentissage des langues, intégration du TAL à l'ALAO, correction d'erreurs, dictionnaire électronique, phraséologie.

Abstract

This paper first presents a history of Computer-Assisted Learning (CAL), setting its origins in the 1920s with the invention of mechanical learning machines. The use of the computer then allowed the development of different types of language learning activities: comprehension tasks, simulations, etc. However, without the contribution of natural language processing (NLP), these activities are of limited use. We address the problem of the integration of NLP in CALL systems while summing up the challenges this integration has to overcome today and synthesize the workshop presentations. These presentations deal with a range of issues from error detection and correction to the extension of electronic dictionaries through the implementation of comprehensive language learning tools. We will see that the key to the integration of NLP in CALL is in the pluridisciplinary work between didacticians, IT and NLP specialists.

Keywords: NLP, CALL, language learning systems, integration of NLP in CALL, error correction, electronic dictionary, phraseology.

1. Introduction

L'atelier « Traitement automatique des langues et Apprentissage des langues assisté par ordinateur : quelles perspectives d'intégration ? » organisé conjointement avec TALN 2006 à Leuven (Belgique) est une tentative de réponse à la question du titre : quelles perspectives

peut-on espérer de l'intégration du TAL dans les systèmes d'ALAO ? Cette question sous-entend une autre : l'ALAO peut-il se passer des procédures du TAL ? Si la réponse à cette dernière question est négative, une troisième question est alors à poser : quelle est la plus-value didactique d'une telle intégration ?

La question de la nécessité d'intégration du TAL dans les systèmes d'ALAO est abordée depuis plusieurs années dans plusieurs colloques internationaux, notamment dans des ateliers organisés par le « Special Interest Group in Language Processing » d'EUROCALL et par l'Integrating Speech Technology in (Language) Learning SIG. Un atelier organisé à Grenoble en octobre 2004 dans le cadre d'une journée ATALA[1] a donné lieu à un numéro thématique de la revue ALSIC (Antoniadis et Chanier, 2005). Même si cet atelier ne concernait, en grande partie, que des travaux francophones, la conclusion fut unanime, seul le TAL permet de considérer la langue, objet de l'apprentissage, en tant que système de formes et d'associations forme-sens.

Une telle affirmation n'entraîne pas nécessairement l'utilisation *a priori* du TAL pour les systèmes d'ALAO ; à notre avis, une telle intégration ne peut se justifier que par la plus-value didactique qu'elle induit. En ce sens, l'objet de cet atelier est d'essayer de définir cette plus-value, de la matérialiser, de la quantifier, de la formaliser, tant que faire se peut. Vaste programme pluridisciplinaire !

2. Des machines à enseigner à l'ALAO

Dès 1912, Thorndike rêvait d'un livre manuel mécanisé : « si, par le miracle et l'ingéniosité mécanique, un livre pouvait être agencé de telle façon que seulement pour celui qui aurait fait ce qui est demandé à la première page, la page deux devienne visible, et ainsi de suite, beaucoup de ce qui requiert actuellement de l'instruction personnelle pourrait être assuré par le livre » (Thorndike, 1912, cité par Bruillard, 1997, p. 33-34). L'idée de l'enseignement programmé (EP) mécanisé est ainsi née dès le début du XX^{e} siècle en réponse à des critiques émises à l'encontre de l'enseignement « classique » et afin essentiellement de permettre un rythme d'apprentissage adapté à l'élève et une activité continue pour ce dernier.

2.1. Les premières machines à enseigner et l'enseignement programmé

2.1.1. La machine de Pressey

La première machine à enseigner (ME) est celle élaborée par Sidney Pressey (1927), dans les années 1920. Il s'agit d'une machine automatisée pour corriger les Questions à Choix Multiples (QCM) avec quatre boutons correspondant aux réponses possibles à la question présentée. L'apprenant ne passe à la question suivante que lorsque sa réponse est juste et la machine garde une trace des actions de l'apprenant. Certains, dont Skinner (1968), lui reprochent d'avoir fondé sa machine sur des connaissances insuffisantes du phénomène d'apprentissage. Ils vont ainsi se focaliser sur les phénomènes d'apprentissage et créer l'enseignement programmé.

2.1.2. L'enseignement programmé

De Montmollin (1971) définit l'enseignement programmé comme « une méthode pédagogique qui permet de transmettre des connaissances sans l'intermédiaire direct d'un professeur ou d'un moniteur, ceci tout en respectant les caractéristiques de chaque élève pris

[1] Journée ATALA : « TAL et apprentissage des langues », 22 octobre 2004, Grenoble (France), http://w3.u-grenoble3.fr/lidilem/talal/.

individuellement ». Cet enseignement repose sur quatre principes : le principe de structuration de la matière à enseigner (il s'agit de découper et de présenter la matière de manière à faciliter la compréhension et la mémorisation), le principe d'adaptation (l'enseignement doit être adapté à l'élève), le principe de stimulation et le principe de contrôle. Le cheminement est soit linéaire (exemple : la machine de Skinner) soit ramifié (exemple : la machine de Crowder (1963)).

Skinner s'appuie sur les résultats de ses travaux en psychologie du comportement et en partant des résultats du conditionnement opérant en tant que théorie du contrôle des mécanismes d'apprentissage, il envisage la création d'une technologie scientifique de l'enseignement qui utilise l'enseignement programmé susceptible d'être dispensé par une machine à enseigner. Pour Skinner, l'efficacité de l'apprentissage repose sur cinq principes :

- le principe de la participation active : le sujet doit construire sa propre réponse et non la choisir (le QCM entraîne des erreurs que l'élève n'aurait jamais commises sans cette suggestion) ;
- le principe des petites étapes : il faut fragmenter la difficulté pour que même les plus faibles puissent répondre ;
- le principe de progression graduée ;
- le principe de l'allure personnelle : chacun doit pouvoir avancer à son rythme ;
- le principe des réponses correctes : trop d'échecs découragent les élèves, il faut les guider.

Ainsi, dans sa machine, les exercices se trouvent sur un rouleau que l'apprenant fait défiler grâce à une molette. Les questions apparaissent dans une fenêtre, l'élève inscrit sa réponse sur un espace blanc réservé à cet effet, puis il compare sa réponse à la correction et actionne le levier pour passer à la question suivante.

Mais l'EP de Skinner est rapidement critiqué et, dès 1959 Crowder propose un système alternatif. Sa machine diffère de celle de Skinner par différents points, notamment le cheminement et le type d'exercices. En effet, la machine présente des informations qui sont suivies par un QCM. Contrairement à Skinner qui cherche à limiter l'erreur, Crowder lui attribue une fonction importante. De plus, il considère qu'apprendre, c'est souvent apprendre à distinguer, à discriminer. Une fois la réponse corrigée, si elle est bonne, l'apprenant passe à l'information suivante, si elle est mauvaise, l'apprenant est dirigé vers des exercices de rattrapage pour ensuite revenir à l'exercice auquel il a échoué. Lorsqu'il répond correctement à plusieurs questions, il passe par des raccourcis. Ce genre de ME permet ainsi une meilleure adaptation à l'apprenant, et prend ainsi une place prépondérante dans l'EP.

L'EP a ouvert de nouvelles pistes de recherche sur les méthodes et théories d'enseignement et d'apprentissage et a marqué le début de l'enseignement assisté par ordinateur.

2.2. L'utilisation de l'ordinateur pour l'enseignement

Au départ, l'utilisation de l'ordinateur se limitait à automatiser ce qui était fait mécaniquement par les ME. L'enseignement assisté par ordinateur (EAO) n'est réellement né qu'au début des années 60 et ce n'est que dans les années 70, avec les travaux sur les systèmes experts qu'apparaissent les premières tentatives de rendre « intelligent » l'EAO. Ces recherches avaient pour finalités de combler les limites existantes, c'est-à-dire :

- de dialoguer avec l'apprenant en langage naturel ;
- de sélectionner la suite de ce qui doit être enseigné ;

- d'anticiper, de diagnostiquer et de comprendre les erreurs de l'apprenant ;
- d'améliorer les stratégies d'enseignement et de le modifier en fonction de l'apprenant.

Puis les années 70 sont marquées par les premiers micro-mondes, les années 80 par les tuteurs intelligents et les années 90 par les systèmes coopératifs et les environnements interactifs d'apprentissage avec ordinateur.

Les exercices proposés en ALAO se divisent, pour Mangenot (1997), en six catégories (cf. ci-dessous). En fonction du support (DVD / Web), du mode d'utilisation (formation avec un enseignant, autoformation, etc.), de l'âge des personnes à qui il est adressé, etc., ces différentes catégories occupent une place plus ou moins importante.

- **Les tâches de compréhension.** L'évolution des ordinateurs a permis rapidement d'introduire des séquences audio, puis des séquences vidéo dans les tâches de compréhension. Mais le plus souvent, les exercices qui s'y rapportent, sont sous la forme de QCM ou d'objets (séquences audio/vidéo, mots, etc.) à cliquer ou glisser et la correction est en termes de vrai/faux avec parfois quelques commentaires oraux ou écrits tels que « bravo », « c'est bien », etc.
- **Les exercices ayant pour but l'acquisition de connaissances discursives.** Il s'agit le plus souvent de puzzle, de repérage de séquence, d'appariement. La correction est en termes de vrai/faux.
- **Les enregistrements d'énoncés et les exercices oraux de transformation d'énoncés.** Ce type d'exercices correspond à une tâche de répétition ou de transformation d'énoncés. L'apprenant a la possibilité de s'enregistrer, de s'écouter. Dans certains logiciels sa voix est utilisée pour doubler un personnage, dans d'autres l'apprenant peut regarder son sonagramme et le comparer à celui de la personne qu'il répète. Dans tous les cas, il est en situation d'autocorrection.
- **Les simulations.** Mangenot distingue trois formes de simulations : celles qui permettent de laisser le choix à l'utilisateur (ce qui correspond au fonctionnement des « livres dont vous êtes le héros ») par exemple, entre aller au restaurant et aller visiter un monument ; celles qui consistent à prendre connaissance de documents des personnages et à faire la même chose qu'eux, par exemple, un CV ; le troisième type de simulation consiste à associer des éléments graphiques pour obtenir une réaction du système. Ce type d'exercice n'entraîne pas de correction mais plutôt des commentaires.
- **Les productions écrites.** La production écrite n'est pas très présente dans les logiciels. Ceci est dû au fait que sa correction est trop difficile. Il existe déjà des problèmes pour les questions à réponses obligatoirement courtes (exemple : les textes à trous), mais le problème est bien plus grand avec les véritables productions pour lesquelles il faut prendre en compte la syntaxe et la sémantique. De ce fait, en ALAO, les corrections apportées lors d'exercices de production écrites sont soit formulées sous forme de correction type que l'apprenant doit comparer à sa production, soit d'indications

Que ce soit sur CD-Rom ou sur Internet, les logiciels d'ALAO ne comportent que très peu d'exercices de lecture et d'écriture. Les exercices les plus présents sont ceux nécessitant le moins de corrections, c'est-à-dire ceux pour lesquels la réponse est donnée grâce à un clic de souris.

3. Le TAL au secours de l'ALAO

Les premières tentatives d'utilisation du TAL pour l'apprentissage des langues datent des années 80. Maladroites au départ (voire scientifiquement contestables lorsque, par exemple, elles essaient d'assimiler le TAL à l'utilisation des programmes tels que ELIZA ou SHRDLU), elles se développent au début des années 90. Le nombre de symposiums et de travaux européens (Swartz et Yazdani, 1992), nord américains (Holland *et al.*, 1995) ou français (Chanier *et al.*, 1993) pendant cette période atteste de cette activité. Ce constat est partagé par Jung (2005), qui situe le pic des travaux dans la période 85-95. Cela correspond au plein développement des ordinateurs personnels et à la première tentative d'insertion des TIC (Technologies de l'Information et de la Communication) dans les dispositifs pédagogiques. Une trace de ces travaux est présente dans toutes les revues traitant de l'apprentissage des langues : *CALICO, Language Learning and Technology, ReCALL, CALL, ALSIC, System.*

Les deux formes de la langue, écrite et orale, sont concernées ; néanmoins, l'état d'avancement du traitement de la langue écrite comparativement à celui de la parole influence directement le nombre, et souvent la qualité, de solutions et de systèmes proposés. Ainsi, les travaux et systèmes concernant la forme écrite sont nettement majoritaires et touchent un nombre plus important de facettes et situations de l'apprentissage des langues.

Implicitement ou explicitement, une idée sous-tend tous ces travaux : l'apprentissage des langues assisté par ordinateur demande des travaux pluridisciplinaires en partenariat, sur un pied d'égalité entre les différentes disciplines. Seule l'association intime des problématiques de chaque discipline ou domaine concernés (Didactique des langues, Informatique, Linguistique, TAL) permet de proposer des solutions et des systèmes opérationnels, dignes d'intérêt pour les apprenants et capables de leur offrir une plus-value didactique par rapport aux méthodes et systèmes classiques. Dans ce travail en association, la tâche essentielle pour chaque discipline ou domaine est la définition de la partie du terme « apprentissage des langues assisté par ordinateur » qui la (le) concerne. Ainsi, c'est la Didactique qui est la plus apte à déterminer le terme « apprentissage », la Linguistique le terme « langue », l'Informatique celui « assisté par ordinateur ». Chacune de ces définitions doit être confrontée aux autres, adaptée aux contraintes des définitions partenaires, apporter sa part à l'élaboration de la solution globale.

L'utilisation des corpus pour l'apprentissage des langues est une « conséquence » majeure, à notre avis, de l'utilisation du TAL et, bien sûr, de l'informatique. Le TAL a permis aux enseignants des langues d'utiliser la richesse des corpus, la diversité des situations langagières qu'ils contiennent ; de disposer d'une source pratiquement intarissable d'exemples de la langue « réelle », celle qu'il faut apprendre, celle que les apprenants auront à utiliser lors des situations communicatives ; et d'automatiser l'analyse des productions des apprenants eux-mêmes. Cette approche d'apprentissage a donné lieu à un grand nombre de travaux, des ressources et des systèmes (Tribble et Barlow, 2001 ; Granger *et al.*, 2001 ; Granger 2002). Le développement de l'Internet a permis le partage et la diffusion de ces ressources et systèmes ce qui, sans doute, a fortement contribué à l'extension de l'utilisation des corpus pour l'apprentissage des langues. Plusieurs systèmes d'ALAO utilisent et exploitent des corpus, bruts ou annotés; le soin apporté à leur constitution et annotation comme la pertinence de leur exploitation détermine souvent la qualité du système qui les utilise.

Si, dans les années 80, la problématique de l'utilisation du TAL pour l'ALAO essaie de se construire en tâtonnant, elle se stabilise dans les années 90, et elle a peu varié depuis. Mis à part, la question « peut-on se passer du TAL pour les systèmes d'ALAO ? » qui a préoccupé

et préoccupe toujours des travaux du domaine (Chanier 1998 ; Brun *et al.*, 2002, Antoniadis 2004), quatre problèmes constituent la majeure partie de la problématique du domaine :

- **Il faut définir et évaluer l'apport du TAL pour l'ALAO.** L'apport potentiel du TAL découle de sa problématique et du but qu'il s'est fixé. Lui seul permet de considérer la forme langagière, non pas comme une suite de signes dépourvus d'interprétation, mais comme des éléments d'un système à deux niveaux (forme et sens). Dans le cadre de l'apprentissage des langues, le fonctionnement de chaque niveau de ce système doit être considéré, détaillé, manipulé, mis en pratique d'une manière ciblée ; les liens entre les deux niveaux doivent être mis en évidence concrètement et la polysémie, source de difficultés en apprentissage des langues, doit pouvoir trouver la place et le traitement appropriés. Seule l'utilisation du TAL permet actuellement d'espérer atteindre ce but, de créer des systèmes et des outils permettant aux enseignants des langues (et par extension aux apprenants) de manipuler la langue telle qu'elle est définie dans leur discipline et non telle que l'informatique est capable de la considérer.

  Comme les enseignants des langues, un système d'apprentissage des langues ne peut être valide, voire acceptable, que s'il est capable, d'une part d'engendrer uniquement des connaissances langagières correctes (non fausses) et d'autre part d'expertiser correctement les productions langagières des apprenants. L'utilisation des procédures du TAL ne peut être envisageable que si elle permet de satisfaire à cette double condition. Concrètement, l'utilisation du TAL n'est possible que lorsque les résultats de ses traitements sont sûrs ou quasi-sûrs ; l'ambiguïté, problème essentiel en TAL, n'est concevable en ALAO que si elle est maîtrisée par le système. Toute maîtrise partielle peut avoir comme conséquence l'apprentissage de concepts langagiers partiellement ou complètement erronés. C'est en ce sens que l'apport du TAL pour l'ALAO doit être défini et évalué. Les travaux actuels tentent, par petits pas, de cerner les résultats du TAL utilisables pour la construction des systèmes d'ALAO.

- **Il faut connaître l'apprenant pour mieux définir son apprentissage.** L'approche individualisée d'apprentissage n'est une idée ni nouvelle, ni propre à l'ALAO. S'il est généralement admis qu'individualiser l'apprentissage permet de l'optimiser, force est de constater que les systèmes maîtrisant une telle démarche sont à créer. Ce n'est pas tant l'évaluation des connaissances de l'apprenant qui constitue la difficulté majeure, mais l'exploitation de cette évaluation et, surtout, la création automatique d'activités en fonction du niveau mesuré de l'apprenant. Le TAL peut apporter des outils aussi bien pour l'évaluation que la création automatique d'activités. Certes ces outils ne sont pas la panacée ; ils constituent néanmoins actuellement la seule piste pour apporter des réponses, partielles le plus souvent, à ces questions.

- **Il faut être capable de détecter, d'expliquer et de corriger automatiquement les fautes de l'apprenant si l'on souhaite qu'il puisse travailler en autonomie.** La question de la correction automatique des productions d'apprenants comme celle de la génération automatique d'explications pour l'apprenant "hantent" les travaux en ALAO depuis ses origines. Néanmoins, aller au delà d'une correction du type "vrai/faux" à partir d'une liste de "vérités" préétablie, ou de l'affichage d'explications préenregistrées, demande la prise en compte des aspects langagiers des réponses fournies, l'évaluation dans le contexte de l'activité pédagogique et en fonction du profil de l'apprenant, la génération, éventuellement, d'explications en langue, en tenant compte des paramètres précédents. En ce sens, et par définition du domaine, le TAL peut apporter toute une panoplie de procédures et d'outils grâce auxquels les questions d'évaluation et d'explications peuvent se poser sur d'autres bases, dans une démarche

constructive et cumulative. Néanmoins, cet apport du TAL est à considérer en tenant compte de ses possibilités actuelles : modéliser et « mesurer » l'attendu, en rejetant les erreurs (l'inattendu). En ce sens l'analyse d'énoncés erronés[2] et la génération d'explications associées ne peut pas être une application directe des procédures TAL; des méthodes et des heuristiques spécifiques restent à inventer.

- **Il faut que les outils et systèmes proposés ne demandent pas de compétences techniques spécifiques.** En tant que domaines de connaissances, l'informatique comme le TAL, la linguistique comme la didactique, manipulent des concepts qui leur sont propres et demandent l'usage d'outils spécifiques, fondés, le plus souvent, sur ces concepts. Pour le profane, l'utilisation de tels outils peut entraîner de longs apprentissages et suppose l'acquisition d'un minimum de concepts du domaine à l'origine de chaque outil. Dans le cadre de l'apprentissage des langues, si les enseignants sont *a priori* des spécialistes de didactique, leurs compétences en informatique ou en TAL, voire en linguistique, sont le plus souvent bien plus limitées, voire minimes ou inexistantes. Cet état peut rendre problématique l'appropriation didactique ou pédagogique de tout outil en dehors du domaine des préoccupations scientifiques des enseignants (et *a fortiori* des apprenants). En ce sens, tout produit de l'ALAO, destiné *a priori* à des didacticiens des langues, doit satisfaire à deux impératifs : son utilisation ne doit demander qu'un minimum de compétences non didactiques et doit permettre le maniement de concepts didactiques en vue de la mise en œuvre de solutions didactiques ou pédagogiques.

Plus de vingt ans après le début des travaux en « TAL et ALAO », force est de constater que si un certain nombre de prototypes ou de systèmes expérimentaux existent ou ont existé (Chanier *et al.*, 1995 ; Brun *et al.*, 2002 ; Antoniadis *et al.*, 2005), les systèmes commercialisés sont extrêmement rares. L'avancée insuffisante des recherches du domaine n'est qu'une explication partielle. À notre avis, deux facteurs sont la cause principale de cet état : la méconnaissance du TAL de la part des didacticiens des langues[3], voire des informaticiens, et le coût des ressources et produits issus du traitement automatique de la langue. Non standardisés, ces derniers restent encore difficilement utilisables en l'état et demandent souvent d'importantes adaptations pour être déployés à profit dans le cadre de l'ALAO.

Ne soyons pas pessimistes ! Un véritable effort de standardisation est en train de s'accomplir dans la communauté TAL, et l'on voit poindre de plus en plus de ressources "génériques" dans la mouvance du logiciel libre (concordanciers, étiqueteurs, lemmatiseurs, etc.). Le plus souvent, ces briques de base ne nécessitent pas d'autre investissement que l'adaptation des formats d'entrée/sortie et la prise en main des paramètres de base. Dans l'état de l'art, la mise en œuvre des outils les plus simples est susceptible d'apporter une plus-value qui dépasse de loin ce modeste investissement. Quant au travail en collaboration entre didacticiens des langues et spécialistes du TAL, l'existence de cet atelier et les travaux qui y sont présentés tendent à prouver qu'un langage commun de travail est en train de se forger.

4. Synthèse de l'atelier

Dans la première présentation de l'atelier, Tschichold aborde le problème majeur qui se présente à quiconque aborde la place du TAL dans l'ALAO, à savoir le manque d'interaction et de collaboration active entre les partenaires principaux : les entreprises commerciales qui

[2] Il ne s'agit pas ici de détecter seulement l'existence d'une faute, mais de calculer en plus le différentiel entre la forme erronée et celle qui aurait dû être à sa place.

[3] L'inverse étant aussi vrai, le travail en commun n'est souvent que le résultat du hasard.

produisent des didacticiels, les enseignants de langues et les spécialistes du TAL. Elle met en évidence les qualités ainsi que les imperfections des différents acteurs de l'apprentissage des langues assisté par l'ordinateur. D'une part, les systèmes commerciaux savent rendre leurs produits attractifs par des interfaces travaillées et des outils multimédia de pointe mais souvent ne prévoient pas de scénarisation pédagogique. D'autre part, les professeurs de langue ont des connaissances didactiques basées sur leur expérience de l'enseignement et les contenus qu'ils développent sont dès lors ciblés et bien adaptés à leur public. Cependant, les systèmes ALAO qu'ils produisent sont souvent peu sophistiqués du point de vue technologique et linguistique. Finalement, les linguistes disposent des outils de TAL qui leur permettent d'ajouter des fonctionnalités aux logiciels d'ALAO mais leurs recherches aboutissent souvent à des prototypes plutôt qu'à des outils réellement utilisables par les enseignants. Ces différents acteurs bien que complémentaires ne communiquent que très rarement entre eux. Cette constatation a été abordée entre autres par Salaberry (1996) et Borin (2002). Borin évoque ce problème comme étant dû à des différences de culture et une mauvaise compréhension entre les différentes disciplines. Salaberry réfute l'idée que l'ordinateur pourrait se substituer à l'enseignant mais défend plutôt l'idée qu'il servirait comme outil de support à l'utilisateur. Nous verrons que la plupart des articles présentés lors de l'atelier évoquent effectivement des outils d'aide à l'apprenant.

L'environnement ALAO pour le basque développé par Aldabe *et al.* illustre bien le type d'approche interdisciplinaire prônée par Tschichold. Leur environnement, qui intègre une **large gamme d'outils de TAL**, est destiné à répondre aux besoins de trois types d'utilisateurs : les apprenants, les enseignants et les linguistes computationnels. Il comprend de nombreuses ressources (conjugueur, concordancier, etc.) destinées à aider les apprenants à résoudre des exercices ou produire des textes et permet, en outre, de collecter des corpus d'apprenants et de traiter leurs erreurs. L'interface a été évaluée par une population d'étudiants, qui l'a jugée très utile pour l'aide à la résolution d'exercices. L'environnement proposé comporte de nombreuses similitudes avec l'environnement d'aide à l'écrit développé par Wible *et al.* (2001) pour l'anglais. Parmi d'autres plateformes intégrant des outils de TAL, citons GLOSSER (Nerbonne *et al.*, 1998), un outil d'aide à la lecture en français par des étudiants de langue maternelle néerlandaise, et MIRTO (Antoniadis *et al.*, 2005), qui intègre des fonctions de TAL basiques que l'enseignant peut combiner de diverses manières pour créer des scénarios pédagogiques pour ses élèves.

L'environnement décrit par Aldabe *et al.* attache une importance particulière à la **détection et la correction des erreurs** des apprenants, un domaine qui est l'objet de nombreux travaux actuellement, comme l'atteste notamment un numéro récent de la revue *CALICO* qui est entièrement consacré à ce sujet (Heift et Schulze, 2003). Deux autres présentations à l'atelier sont centrées sur cette problématique. Audras et Ganascia présentent un outil de diagnostic de l'apprenant basé sur l'utilisation d'un analyseur morphosyntaxique et d'un analyseur stylistique, le *Littératron*. Ce logiciel extrait des motifs syntaxiques récurrents dans un corpus afin de repérer les tournures caractéristiques des étudiants de français de langues maternelles et de niveaux d'apprentissage différents. Hermet *et al.* quant à eux ont développé un environnement virtuel dont le but est d'améliorer la compréhension à la lecture d'étudiants de français langue seconde. Le système, qui comporte un parseur, un dictionnaire de synonymes et un dictionnaire de dérivations, permet de reconnaître et d'évaluer les réponses à des questions ouvertes. En matière de correction automatique, le TAL est plus généralement utilisé pour la résolution automatique d'exercices de type QCM, exercices à trous, questions fermées où les possibilités d'erreur de correction sont limitées. Beaucoup s'accordent pour dire qu'il n'est pas acceptable de donner des informations erronées à l'apprenant au risque de perdre sa confiance en l'outil qu'il utilise. Des systèmes de correction automatique de

questions ouvertes existent basés sur des similarités avec des réponses types ou un vocabulaire attendu dans la réponse pré-entrée dans la machine (cf. par exemple, Perez 2004).

Un autre domaine clé est celui des **dictionnaires électroniques** dont l'importance en ALAO est appelée à croître en raison notamment du rôle grandissant joué par le lexique dans l'apprentissage des langues (cf. la « lexical approach » de Lewis 1993 et 1997). Deux articles abordent cette question : Zock ainsi que Buvet et Issac suggèrent une extension des informations stockées dans un dictionnaire classique, à savoir les définitions et les traductions. La démarche de Buvet et Issac est clairement onomasiologique, celle de Zock est à la fois onomasiologique et sémasiologique. Zock propose un outil d'aide à la compréhension et à la production. Il part du principe qu'un apprenant connaît souvent le mot qu'il recherche mais n'en a qu'une idée approximative. La forme entrée par l'utilisateur ressemble soit par la forme soit par le sens au mot recherché. Pour les approximations de forme, le système utilise des algorithmes orthographiques pour retrouver la forme exacte. Pour ce qui est des approximations de sens, le dictionnaire peut être enrichi d'un réseau lexical et présenter à l'utilisateur la liste des mots proches par le sens. Ainsi, un apprenant qui cherche le mot 'infirmière' passera par le mot 'hôpital' pour le retrouver. À partir de ce dictionnaire, Zock envisage la création automatique d'exercices d'aide à la compréhension avec la présentation d'une liste de mots à mémoriser et dès que l'un d'eux est retenu, il est enlevé de la liste; mais aussi des exercices d'aide à la production avec un exercice de construction de phrases à partir d'un modèle et de mots à utiliser. Buvet et Issac eux, proposent un outil d'aide à la rédaction qui génère des phrases simples du domaine de l'affect dans le but de familiariser l'apprenant avec le lexique concerné et son emploi. Pour cela, chaque entrée du dictionnaire est formalisée suivant le modèle des classes d'objet contenant des informations syntaxiques, sémantiques (hyperonymes, synonymes, etc.) et les contraintes sur les arguments (*sujet : humain*, par exemple). La génération de phrases se produit par l'intermédiaire d'une interface qui, à partir d'un concept et des arguments choisis par l'apprenant, génère toutes les phrases possibles exprimant cet affect. Ces deux projets sont dans la même lignée que le projet ALEXIA (Chanier, 1995), un environnement proposant une aide à la compréhension avec la construction d'une base de données lexicale personnalisée, mais aussi un outil d'aide à la rédaction qui analyse le texte produit par l'apprenant sur un thème précis. Parmi d'autres dictionnaires d'aide à la compréhension ou à la production, citons le DAFLES (Selva, 2002), un dictionnaire utilisé pour la génération automatique d'exercices lexicaux, et le lexique bilingue de Pecman (2005), qui a pour but de fournir une aide à la rédaction scientifique en anglais et en français.

L'exploitation par le TAL des corpus de textes électroniques a revitalisé et profondément modifié le domaine de la **phraséologie**, qui touche à l'extraction, la catégorisation et l'intégration des unités polylexicales (cf. Granger et Meunier, à paraître). L'article de Wible *et al.* décrit un outil d'extraction et de représentation de séquences préfabriquées (« lexical chunks ») ainsi qu'un outil, le *Collocator*, servant à mettre en évidence les collocations dans un texte afin d'aider l'apprenant à reconnaître les expressions figées. Cet outil permet à l'apprenant de découvrir les collocations en temps réel à l'intérieur des pages Web qu'il parcourt. Il est basé sur un système de statistiques qui permet de trouver des cooccurrences de suites d'un ou de plusieurs mots même si celles-ci sont entrecoupées d'autres mots.

5. Conclusion

Le champ couvert par l'application du TAL à l'ALAO est vaste. Il recouvre de nombreux sujets dont certains sont totalement absents de cet atelier, en particulier ce qui touche au traitement de l'oral. En ce qui concerne le traitement de l'écrit, cependant, les thématiques

abordées donnent une bonne idée des nombreux défis auxquels les chercheurs sont confrontés à l'heure actuelle. Une chose apparaît de manière sûre : si les chercheurs qui travaillent à l'intersection du TAL et de l'ALAO veulent avoir une chance de voir leurs travaux mis en pratique dans l'enseignement des langues, il est indispensable qu'ils accordent une grande importance à la scénarisation pédagogique et il faut bien avouer que cet aspect est souvent marginal, voire inexistant. La clé réside dans le travail pluridisciplinaire et la mise en commun de pratiques et de techniques. Nous espérons que cet atelier contribuera à fédérer les travaux menés par des chercheurs d'horizons divers et à faciliter l'intégration des résultats de leurs recherches dans des outils d'ALAO conviviaux et performants.

Références

ANTONIADIS G. (2004). « Les logiciels d'apprentissage des langues peuvent-ils ignorer le TAL ? ». In *Les cahiers de l'APLIUT* XXIII (2), juin 2004 : 81-97.

ANTONIADIS G., CHANIER, T. (Éds.) (2005). *ALSIC* 8 (2), numéro thématique « TAL et apprentissage des langues ».

ANTONIADIS G., ECHINARD S., KRAIF O., LEBARBÉ Th., PONTON C. (2005). « Modélisation de l'intégration de ressources TAL pour l'Apprentissage des Langues : la plateforme MIRTO ». In *ALSIC* 8 : 65-79

BORIN L. (2002). « What have you done for me lately? The fickle alignment of NLP and CALL ». Reports from Uppsala Learning Lab.

BRUILLARD E. (1997). *Les machines à enseigner.* Hermès, Paris.

BRUN C., PARMENTIER T., SANDOR A., SEGOND Fr. (2002). « Les outils de TAL au service de la e-formation en langues ». In Fr. Segond (dir.), *Multilinguisme et traitement de l'information.* Hermès, Paris : 223-250.

CHANIER T. (1998). « Relations entre le TAL et l'ALAO ou l'ALAO un simple domaine d'application du TAL ? » In *International conference on natural language processing and industrial application (NLP+IA'98).* Moncton.http://lifc.univ-fcomte.fr/RECHERCHE/P7/pub/Moncton/index.htm

CHANIER T., RENIÉ D., FOUQUERÉ C. (dir.) (1993). *Actes du colloque SCIAL'93 (sciences cognitives, informatique et apprentissage des langues)*. Université Blaise Pascal, Clermont-Ferrand. Préface et sommaire en ligne à http://archive-edutice.ccsd.cnrs.fr/edutice-00001148

CHANIER T., FOUQUERÉ C., ISSAC F. (1995). « AlexiA : Un environnement d'aide à l'apprentissage lexical du français langue seconde ». In *Conférence Environnements Interactifs d'Apprentissage avec Ordinateur* (EIAO'95). Eyrolles, Paris : 79-90.

CROWDER N. (1963). « On the difference between linear and intrinsic programming ». In *Phi Delta Kappan* 44 : 250-254.

DE MONTMOLLIN M. (1971). *L'enseignement programmé.* PUF, Paris.

GRANGER S. (2002). « A bird's eye view of learner corpus research ». In S. Granger, J. Hung et S. Petch-Tyson (eds), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, *Language Learning and Language Teaching* 6. Benjamins, Amsterdam/Philadelphia : 3-33.

GRANGER S., MEUNIER F. (eds) (à paraître). *Phraseology : An Interdisciplinary Perspective*. Benjamins, Amsterdam/Philadelphia.

GRANGER S., VANDEVENTER A., HAMEL M.J. (2001). « Analyse des corpus d'apprenants pour l'ELAO basé sur le TAL ». In *TAL* 42 (2) : 609-621.

HEIFT T., SCHULZE M. (eds) (2003). S*pecial issue of CALICO on Error Analysis and Error Correction in Computer-Assisted Language Learning* 20 (3).

HOLLAND V.M., KAPLAN J.D., SAMS M.R. (dir) (1995). *Intelligent Language Tutors. Mahwah.* Lawrence Erlbaum Associates, NJ.

LEWIS M. (1993). *The Lexical Approach : The State of ELT and a Way Forward*. Language Teaching Publications, Hove.

LEWIS M. (1997). *Implementing the Lexical Approach : Putting Theory into Practice.* Language Teaching Publications, Hove.

MANGENOT F. (1997). « Synthèse de trois cours de FLE sur CD-ROM ». In *Les Cahiers de l'asdifle, multimédia et langue étrangère, Actes des 19e et 20e rencontres* : 79-88.

NERBONNE J., DOKTER D., SMIT P. (1998). « Morphological Processing and Computer-Assisted Language Learning ». In *Computer-Assisted Language Learning* 11 (5) : 543-559.

PECMAN M. (2005). « Compilation, formalisation and presentation of bilingual phraseology ». In C. Cosme, G. Gouverneur, F. Meunier, M. Paquot (éds), *Phraseology 2005. The Many Faces of Phraseology. An Interdisciplinary Conference.* Université catholique de Louvain, Louvain-la-Neuve : 335-338.

PEREZ D. (2004). *Automatic Evaluation of User's Short Essays by Using Statistical and Shallow Natural Language Processing Techniques.* Advanced Studies Diploma Work. Universidad Autonoma de Madrid.

PRESSEY S.L. (1927). « A machine for automatic teaching of drill material ». In *School and Society* 25 : 549-552.

SALABERRY R. (1996). « A Theoretical Foundation for the Development of Pedagogical Tasks in Computer Mediated Communication ». In *CALICO* 14 (1) : 5-34.

SELVA T. (2002). « Génération Automatique d'Exercices Contextuels de Vocabulaire ». In *Actes de TALN 2002* : 185-194.

SKINNER F.B. (1968). *La révolution scientifique de l'enseignement*. Charles Dessart, Bruxelles.

SWARTZ M., YAZDANI M. (dir.) (1992). *The Bridge to International Communication : Intelligent Tutoring Systems for Foreign Language Learning*. Springer-Verlag.

THORNDIKE E.L. (1913). *Educational Psychology 2. The Psychology of Learning*. Teachers College, New-York.

TRIBBLE C., BARLOW M. (éds) (2001). *Using Corpora in Language Teaching and Learning.* Special issue of *Language Learning and Technology* 5 (3).

WIBLE D., KUO C-H., CHIEN F-Y., LIU A., TSAO N-L. (2001). « A Web-based EFL writing environment : integrating information for learners, teachers, and researchers ». In *Computers and Education* 37 : 297-315.

Intelligent CALL: The magnitude of the task

Cornelia Tschichold

University of Wales Swansea – CALS
c.tschichold@swansea.ac.uk

Résumé

La qualité de la plupart des programmes ALAO n'est pas bien équilibrée en ce qui concerne l'utilisation de la technologie informatique et le contenu linguistique. Ce déséquilibre peut être expliqué par les contraintes très divergentes agissant sur le développement des didacticiels. Les produits ALAO commerciaux souffrent surtout d'un manque d'adaptation à l'apprenant et du manque de réponse intelligente à la production linguistique de l'apprenant. Les approches TAL n'ont pas encore atteint une qualité suffisante à cause de la distance énorme entre le langage des apprenants et les genres linguistique pour lesquels les outils TAL ont été développés. La solution proposée est une approche locale, centrée sur le lexique (bilingue), approche qui combine les ressources existantes pour arriver à des didacticiels plus créatifs et interactifs.

Mots-clés : méthodologie, évaluation et réponse, lexique, ALAO, TAL.

Abstract

The quality of most CALL programs is not well balanced with respect to the use of computer technology and of language content and processing. This imbalance can be explained by a number of constraints pulling CALL developers in diverging directions. For commercial CALLware the poor learner fit and lack of feedback is a serious impediment. So far ICALL approaches trying to overcome this have not been of a sufficiently high quality due to the vast distance between most learner language and the text genres NLP is helpful for. The way forward suggested here is for ICALL to take a localized, (bilingual) lexicon-centred approach that combines sophisticated resources with improved learner fit for more creative and interactive CALLware.

Keywords: methodology, feedback, lexicon, ICALL, CALL, NLP.

1. Introduction

The two fields of computer-assisted language learning (CALL) and the applied branch of computational linguistics, natural language processing (NLP) would seem to be natural allies when considering ways to improve the way computers are used for learning foreign languages, and one might therefore wonder why the development of Intelligent CALL (ICALL) programs is not more advanced at the present moment. Multimedia computers certainly seem to be capable of representing reality in all its possible (and also in many impossible) forms, so the hardware cannot be the problem. There are numerous well-functioning applications of NLP, from spell-checkers and electronic dictionaries to sophisticated machine translation programs. And we also have a huge and constantly growing body of research on second language acquisition (SLA) that should allow us to make the right decisions in the framework of CALL development. So why does it appear to be so difficult to combine all of this knowledge with the necessary hardware to come up with a really intelligent CALL program?

A first non-trivial problem is of course the fact that ultimately it is not so simple to combine several very different types of knowledge needed for such a goal, knowledge which is

typically held by different groups of people, so that it becomes very challenging to achieve a fruitful cooperation in a converging work culture. Computer hardware and software specialists, computational linguists, SLA researchers and foreign language teachers tend to work in very different paradigms, with widely diverging background assumptions and typical constraints on their work. Bringing people from such different traditions together and balancing their contributions to the common goal is the first step that is necessary for an ICALL project on a reasonable scale to be successful.

If we take a look at existing programs, we can see that this is indeed not such an easy step to make. CALL programs are clear evidence that the emphasis during the development stage of the software was either on delivering good linguistic content or rather on good, up-to-date use of technology. Many of the commercial programs make excellent use of the technology available and look very attractive at first sight, but are disappointing where it comes to the language content and the pedagogical principles seen at play there. These shortcomings can often be explained quite easily by the commercial pressures at work in such projects. Not only do the developers have to work to tight time schedules, they also need to come up – at frequent intervals – with a new product that appeals to a wide market such as the worldwide EFL market.

On the other hand, we find programs where much more time and thought has been spent on the language content and how much and what kind of language exposure learners should have and how the learners' linguistic production is to be processed. These programs tend to be developed either by enthusiastic teachers with extensive teaching experience and a sound knowledge of their particular student group, or by academics with a background in computational linguistics.

The programs developed by teachers are typically small and well adapted to their own student population, but often use technology less than ideally when compared to commercial programs and are limited in their software functions. And then there are programs developed by groups of academically based linguists trying to use innovative approaches to process language and generate feedback, but this group is also less concerned about the use of up-to-date multimedia features.

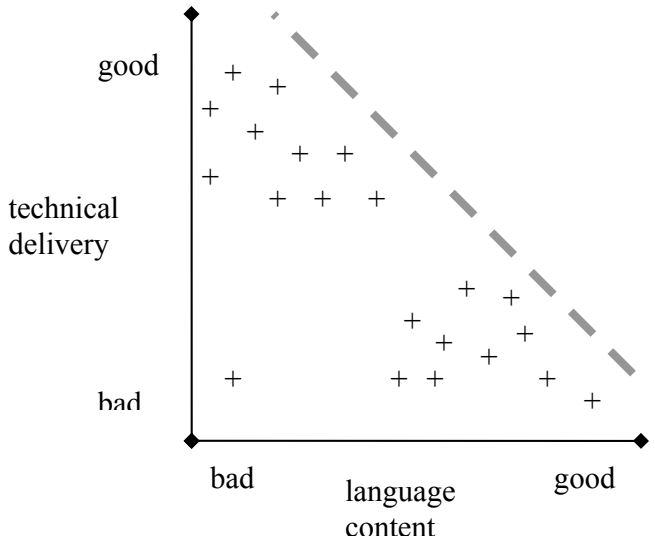


Figure 1. Delivery mode vs. content quality in CALL

The common element in all these types is that they are typically less than perfect in (at least) one of the two aspects of technical and linguistic components. If we plot the quality of a random number of them (each '+' standing for one hypothetical CALL program) on a graph, with technology and delivery on one axis, and language content and feedback on the other, we might get a picture as seen in figure 1. There seems to be a certain bar (- - - -), with an area above it that is difficult to reach for CALL programs. For a general evaluation, where all of these aspects have to be taken into account, technical delivery should meet a certain minimal threshold (albeit one that is rising all the time). The language content is less easy to evaluate, but ultimately it is the learning success that can be achieved with the help of the program which should determine the overall mark. This aspect can usefully be broken down into facets such as consistency, learner fit, etc.

The best existing programs are thus found near the bar; unfortunately, we also find a number of programs quite far away from that line. But no program to date can deliver the dream of a customized, all-purpose language course that makes learning a foreign language an entertaining and expeditious experience, gives us practice in small talk, corrects our written texts, clearly explains the errors we have made, and produces a couple of remedial exercises before proceeding to the next tailored lesson. Such expectations are well beyond today's technology. The reasons for this are to be found in a combination of forces or constraints pulling in different directions.

2. Three factors at play (at least)

These diverging constraints on the development of CALLware tend to lead to dissimilar end products by the different groups involved in authoring CALL. While commercial developers tend to yield to the pressures of time, money and the use of the latest computer technology, teachers will most likely focus on adapting the language content, and computational linguists will primarily be interested in improving feedback mechanisms by processing (more or less) free text. A number of other factors can easily be imagined to play a role in the development process, as illustrated in figure 2, but these three diverging constraints will be the focus here.

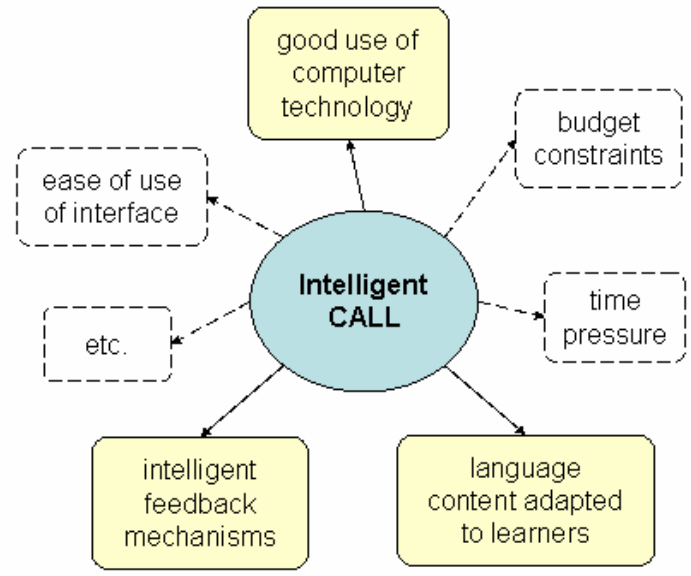


Figure 2. Opposing constraints in CALL development

2.1. Commercial CALL

While much of commercial CALLware uses the latest computer hardware and software for the delivery of the content in useful ways, the language content itself is frequently less than ideal. The attractive interface sometimes seems to be more important than the less easily visible language content (Hewer *et al.*, 1999). Commercially developed programs, but not only those, often seem to show a certain lack of awareness of (not to say disregard for) the complexity of the task. Clifford (1998) lists some of these in his seven misperceptions about CALL, and as his list has lost none of its relevance today, I quote it here.

1. language is a simple phenomenon;
2. language acquisition is a simple task;
3. presentation of information is all that is required for language learning;
4. access to information eliminates the need for presentations;
5. every learning activity is appropriate for all learners;
6. automating poor teaching practices will improve the instructional process;
7. poorly designed 'shovelware' is better than well designed, but costly, learning activities. (Clifford, 1998, p.1)

In such a (common, but unconscious) view, language is seen as a finite body of knowledge (1), which simply needs to be presented (2, 3) to the learner and then practised via a restricted set of exercises (2, 5). CALL programs developed with these misconceptions in their developers' minds typically use rudimentary presentation followed by multiple choice questions and other simple to mark tasks. The language meant to be taught is rarely presented in a way that would be conducive to learning it; and the possibilities of practising it are hampered by the lack of meaningful feedback beyond the "wrong – try again"-type.

Reasonably critical reviews of CALL programs such as those on the CALICO website often mention the lack of sound pedagogical principles behind much new CALLware. But descriptions of programs frequently focus on functionalities and other technical aspects[1] and say comparatively little on the language content and the didactic aspects of the program in question. In view of the aim of CALL, *i.e.* language learning, the largest proportion of text of a CALL review should be devoted to the content of the program. The recent review by Burston (2005) of a vocabulary training program is a case in point. While the words themselves (given as examples in the review) belong to a beginner's vocabulary, the example sentences shown are only suitable for advanced learners. Burston rightly questions some of the methodology and the feedback mechanism offered by the program, but the overall mark seems to evaluate the program more against other similar CALL programs than against other, non-computer-based methods for learning vocabulary.

Commercial CALL program developers are of course not completely oblivious to criticism by pedagogists and language teachers, and program descriptions often pay lip service to the developments in language teaching methodology in stressing the communicative aspects of the exercises[2]. In many cases, however, this has led to a move away from the earlier structure drills towards the delivery and display of growing amounts of linguistic material. Unfortunately, this usually also means fewer attempts at trying to deal with learners' language

[1] It has to be mentioned here that technical aspects such as compatibility are among the first criteria that need to be taken into account when considering whether to buy some piece of software. But in an ideal world pedagogical aspects would override such purely technical considerations.

[2] One program criticized quite sharply in Nesselhauf and Tschichold (2002) has since appeared in a much improved version.

production, leading to an increasing lack of any type of interactivity that goes beyond mouse clicks. This is a safe strategy as it minimizes the danger of giving wrong or otherwise unsuitable feedback to the user, a danger difficult to underestimate from a didactic point of view (Schulze, 2003; Tschichold, 1999). Such CALL programs can be excellent for the delivery of practice material, but they avoid dealing with the productive skill of learners, an aspect which is increasingly being recognized as crucial for the language learning process. Restricting teaching to the presentation of linguistic material leads to a learnability problem, as simple exposure to such material is not sufficient to acquire many components of a language. CALL programs which do not invite the learner to produce utterances in the foreign language can thus not claim to be complete learning packages.

2.2. CALLware designed by practitioners

A number of enthusiastic language teachers have designed and implemented their own small CALL programs. As these are primarily meant to be used by the author's own learners, they are particularly well suited to the needs of these students. The downside is that the limited time and programming experience available to this group of CALL developers often results in the technological aspects of such programs being somewhat simple and not necessarily up-to-date with the latest hardware developments, a consideration that could take on increasing importance as more and more language learners will have grown up with technologically highly sophisticated computer games. The number of software functions in these teacher-developed CALL programs is also necessarily restricted. However, these shortcomings are often balanced by the considerable advantage of the fine-tuning and frequent up-dating such programs make possible, at least as long as the author is willing to continue working on the program.

What such programs lack in technological sophistication and innovation[3], they often make up in learner fit. Given that the authors know their students well, they can rely on their experience and adapt the feedback to the expected answers provided by their students. By doing this, they can at least partially compensate for one of the major difficulties when determining feedback on learner input, namely that fact that the language taught to language learners, especially to beginners and those taught in a communicative framework, is highly ambiguous simply because this is the most useful type of language, adaptable to all sorts of contexts. This property of human language makes it difficult for computer-assisted instruction, however, to predict all possible (correct and wrong) answers to any but the most simple types of questions.

2.3. NLP in CALL

The impossibility of providing adequate feedback to any type of question that goes beyond the complexity of multiple choice questions in traditional CALL has led to numerous attempts (see the projects described in *e.g.* Holland *et al.*, 1995; Jager *et al.*, 1998; Gamper and Knapp, 2002; Heift, 2003 and others in that special issue of the CALICO journal; and Dodigovic, 2005) by linguists to use techniques developed in computational linguistics in order to analyse the language produced by learners and generate feedback that would allow the learner to identify and correct the error found by the program. This group of CALL developers is not

[3] Teachers who want to write their own CALL materials do not need to learn a programming language any more; they can use so-called authorware that allows them to input their own customized language content with the help of templates, while providing the interface for a number of simple exercise types. Frequent updates of authorware can help to overcome the lack of technological sophistication.

primarily concerned with the use of the latest technology in order to come up with a good-looking interface either, so their programs will not be as eye-catching as the commercial programs. Despite the didactic and academic interest they offer, none of these NLP-based projects have made it into the commercial market so far, partly because of their small scale, but partly also because their most interesting aspect, error detection and customized feedback, does not work very reliably outside the context for which it has been developed. Coupled with the fact that only a minority of errors produced by language learners are detected at all, this is a disadvantage that is difficult to ignore from a didactic point of view.

Many of the NLP components used in ICALL projects were originally developed for processing correct native speaker language, and not for the purpose of error detection in learner language. On the level of morphology, the obstacles for adapting an NLP component (an electronic morphological lexicon in this case) to error-prone learner language are not insurmountable, but on the level of syntax it can become very difficult to balance the need for relaxing the grammar constraints in order to be able to parse a sentence at all, and at the same time limiting the proliferation of analyses caused by such a method. The alternative method of using 'error grammars' that look for specific types of errors is safer, but will obviously only lead to the detection of a limited number of errors. It will always remain impossible to imagine and plan for all possible errors language learners could make. In order for the latter approach to be effective at all, the program has to be specific to a particular learner group, with users sharing the same or very similar native language(s). A number of recent systems, such as those described in Heift (2003), Reuer (2003), Schulze (2003), and Vandeventer (2001), take this approach. Once an error has been detected by such a system, the developers are then faced with the task of translating the machine-like code produced by the NLP system into a text message that is fit for human consumption, *i.e.* for language learners who are neither computer programmers nor trained linguists in most cases. Given the vast distance between the way most language learners think about elements of a sentence and the methods used and output produced in NLP, this is far from trivial.

While morphology and syntax doubtlessly account for substantial numbers of errors found in learner language, problems on the levels of semantics and the lexicon also represent a significant proportion of all learner errors (Cutting, 2000), especially for the most frequently learnt foreign language, English, with its relatively poor morphology and vast vocabulary. There is no viable NLP system for dealing with this aspect of language as yet. Existing (commercial) systems are only capable of treating a small part of any language or a subregister where the amount of vagueness and ambiguity so typical for general language has been drastically reduced.

The problem for CALL then is twofold: on the one hand the foreign language lexicon is one of the biggest obstacles for language learners at all levels of proficiency (Nation, 2001), leading to significant numbers of errors and other divergences from native speaker language use, and – on the other hand – the lack of viable NLP components to deal with the semantic level of language.

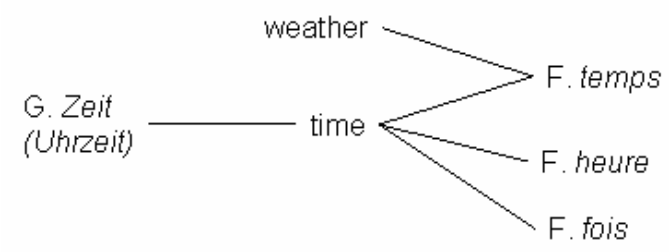
We find a corresponding twofold dilemma in computational linguistics: Along with the lexical bottleneck, *i.e.* the problem of scaling up small NLP systems to a sufficiently large lexicon, the semantic ambiguity in everyday language is one of the most persistent problems for computational linguistics. Those NLP components where this problem has been tackled to such a degree that viable products emerge are specific to highly specialized text genres, *e.g.* construction manuals rich in specialized terminology. In such text genres, the ambiguity inherent in most words can be reduced to such a degree that the NLP system can cope with the remaining problems to a satisfactory degree. This type of language, however, could not be

further away from the kind of language taught to and used by language learners. As the number of words language learners can be expected to learn within a few years is very limited compared to the total number of words in any natural language, the words that are chosen for beginners are necessarily those that will be most useful to them, in other words, those vocabulary items that are quite vague and ambiguous. In addition to this deliberately imprecise vocabulary, language learners are also often taught (and/or develop) strategies to paraphrase, *i.e.* to use even less precise words to describe the intended meaning. Briefly, the language produced by learners is about the worst imaginable type of language for NLP.

3. The dilemma for ICALL

Given that on the one hand, learners are usually taught highly ambiguous language because this is the most useful type of language for communicative purposes, and that on the other hand, NLP components can only deal reasonably efficiently with language genres where the natural ambiguity of everyday language has been massively reduced, we can see the fundamental dilemma for ICALL, a dilemma that remains even if other problems such as budget and time constraints and the quality of the interface and of the language content in CALLware are solved or disregarded. One possible conclusion CALL developers could draw from this situation is to concentrate on potentially more feasible and very specific subtasks rather than try and teach all aspects of the foreign language to all types of learners.

As computers are probably less than ideal to teach the truly communicative aspects of language, vocabulary learning could be seen as a subtask suitable for CALL. Learning hundreds and thousands of new words involves at least a certain element of drill and repetition for most learners, certainly those outside an immersion situation (Nation, 2001). While the vocabulary training programs on the market today have not necessarily been developed out of such motives, they can still serve as a suitable illustration of the argument. Most of them are specific to a single language pair, so one advantage in terms of NLP use is that the user group and the error types that can be expected are clearly more specific than they would be for CALL programs aiming to be L1-neutral. While there are vast differences in the didactic approach and quality of implementation of such vocabulary programs, most of them share one problematic aspect, the reduction of vocabulary to bilingual pairs of single words (Nesselhauf and Tschichold, 2002). This unfortunately supports the assumption by many beginner language learners that every text word in their native language has a one-to-one match in the foreign language, an assumption that has been called the naïve lexical hypothesis (Kesner Bland *et al.*, 1990). While such an assumption sometimes works reasonably well for some words from closely related languages (English *time* can very often be translated into German as *Zeit*), in most cases it is a dangerously simplistic view. Disregarding the most naïve level of this hypothesis, *i.e.* the assumption of complete isomorphism down to the level of morphology, we readily find examples of diverging polysemy to trap the learner. English *time*, for instance, has at least three common translations into French, with one of them, *temps* also being the translational equivalent of English *weather*.



Such complicated relations are the norm in the basic vocabulary of any language pair if we take into account the numerous combinations words enter into as soon as they leave the vocabulary list and are used in real utterances. Beginners' vocabulary lists are necessarily full of words that are quite highly polysemous, but unfortunately for both the language learner and the CALL developer, the polysemy of individual words in one language typically is not concurrent with the polysemy of one of the translational equivalents in the other language[4].

Linguists working on learner lexicography both inside and outside the field of CALL are beginning to address this challenge. Paper dictionaries like those in the Cambridge series "Word Routes" or the electronic learners' dictionary project Eldit (Abel and Weber, 2000) show a promising approach towards a possible solution of the dilemma described above. Computer-based lexicographic databases have a distinct advantage in this area as they are not bound to the linear presentation of data and allow for multiple links, offering almost limitless possibilities for presenting subsets of data to the user.

To conclude, I would like to argue in favour of a more localized approach to ICALL. The functional and communicative aspects of language are probably better left to human teachers for the time being, but this still leaves much room for CALLware. We should remember that lexical errors and lack of adequate vocabulary are the biggest hindrance to communication in a foreign language. It therefore makes sense for ICALL to concentrate on this aspect. But vocabulary needs to be taught in much more varied, complex, creative and interactive ways. Words need to be presented in several different contexts, gradually increasing in difficulty, so that learners can avoid the trap of the naïve lexical hypothesis. Words also need to be practised and revised through a range of exercises. These should be designed in such a way that the program can give intelligent feedback to the user. The aim should be for CALL activities to focus on the progression from controlled learning to automatic processing of linguistic forms, a step that is generally assumed to be achieved through practice and routinization. This fact would favour a role for CALL that is centred around vocabulary learning and a lexically centred approach to language teaching. We have the technology (large-scale lexicons and morphological analysers), the linguistic knowledge (research findings on vocabulary acquisition), and the language data (various corpora) necessary to design language learning tasks that put these principles into practice.

References

ABEL A., WEBER V. (2000). "ELDIT – A Prototype of an Innovative Dictionary". In U. Heid *et al.* (eds), *EURALEX Proceedings* II. Stuttgart: 807-818. http://www.eurac.edu/eldit.

BURSTON J. (2005). *Review of WordChamp*. http://calico.org/CALICO_Review/review/wordchamp00.htm .

CLIFFORD R.(1998). "Mirror, Mirror, on the Wall: Reflections on Computer-Assisted Language Learning". In *CALICO Journal* 16 (1): 1-10.

DODIGOVIC M. (2005). *Artificial Intelligence in Second Language Learning*. Multilingual Matters, Clevedon.

GAMPER J., KNAPP J. (2002). "A review of intelligent CALL systems". In *Computer-Assisted Language Learning* 15 (4): 329-342.

[4] The popular literature on so-called (true and) false friends illustrates the phenomenon for the relatively small group of words that also have a formal resemblance, but the phenomenon is common across the whole bilingual lexicon.

HEIFT T. (2003). "Multiple learner errors and feedback: a challenge for ICALL systems". In *CALICO Journal* 20 (3): 533-548.

HEWER S., RENDALL H., WALKER R., DAVIES G. (1999). "Introduction to computer assisted language learning (CALL)". In *Module 1.4 of ICT4LT*. www.ict4lt.org.

HOLLAND V.M., KAPLAN J.D., SAMS M.R. (1995). *Intelligent Language Tutors: Theory Shaping Technology*. Lawrence Erlbaum, New Jersey.

JAGER S., NERBONNE J., VON ESSEN A. (1998). *Language Teaching and Language Technology*. Swets & Zeitlinger, Esse.

KESNER BLAND S., NOBLITT J., ARMINGTON S., GAY G. (1990). "The Naive Lexical Hypothesis: Evidence from Computer-Assisted Language Learning". In *The Modern Language Journal* 74 (4): 440-450.

NATION I.S.P. (2001). *Learning Vocabulary in Another Language*. Cambridge University Press, Cambridge.

NESSELHAUF N., TSCHICHOLD C. (2002). "Collocations in CALL: an investigation of vocabulary-building software for EFL". In *Computer Assisted Language Learning* 15 (3): 251-280.

REUER V. (2003). "Error recognition and feedback with lexical functional grammar". In *CALICO Journal* 20 (3): 497-512.

SCHULZE M. (2003). "Grammatical errors and feedback: some theoretical insights". In *CALICO Journal* 20 (3): 437-450.

TSCHICHOLD C. (1999). "Grammar checking for CALL: strategies for improving foreign language grammar checkers". In K. Cameron (ed.), *CALL: Media, Design & Applications*. Swets & Zeitlinger, Lisse: 203–222.

VANDEVENTER A. (2001). "Creating a grammar checker for CALL by constraint relaxation: a feasibility study". In *ReCALL* 13 (1): 110-120.

The use of NLP tools for Basque in a multiple user CALL environment and its feedback

Itziar Aldabe, Bertol Arrieta, Arantza Díaz de Ilarraza,
Montse Maritxalar, Ianire Niebla, Maite Oronoz, Larraitz Uria

University of the Basque Country, Department of Computer Languages and Systems
{jibalari ; jiparkob ; jipdisaa ; jipmaanm}@si.ehu.es
iniebla001@ikasle.ehu.es

Résumé

Nous présentons un environnement d'apprentissage assisté par ordinateur (ALAO) pour le basque. Cet environnement a différents buts : d'une part, il offre aux utilisateurs (étudiants, enseignants, et linguistes informaticiens) différents outils et ressources permettant de clarifier des doutes linguistiques éventuels qu'ils peuvent avoir sur certains aspects de la langue et d'autre part, il permet de stocker des informations sur les apprenants (les déviations, les erreurs, etc.) pour permettre de nouvelles études en ALAO ou TAL. Le système est composé d'une plate-forme de base (Lentillak), de deux applications Web (Erreurs et Irakazi), de plusieurs outils TAL et de deux bases de données (Errors et Deviations). Il intègre en outre des corpus d'apprenants ainsi que des corpus de natifs. Nous présenterons également une expérience d'évaluation que nous avons réalisée pour mesurer l'efficacité des outils de TAL.

Mots-clés : ALAO, outils de TAL, application Web, basque.

Abstract

In this article, we present a Computer Assisted Language Learning (CALL) environment for Basque. The environment has different aims: on the one hand, to offer the users (teachers, learners and computational linguists) different tools and language resources to clarify the linguistic doubts they might have about the language, and on the other hand, to store information about language learners, deviations and errors as the basis for further studies in CALL and Natural Language Processing (NLP). The environment is composed of a workbench (Lentillak), two web applications (Erreus and Irakazi), several NLP tools and two databases (*Errors* and *Deviations*), and it takes in learner corpora as well as native corpora. In addition, we present the experiment we have carried out to evaluate the usefulness of the NLP tools.

Keywords: CALL environment, NLP tools, Web application, Basque.

1. Introduction

In the last years, there has been a growing interest in the NLP community in CALL, and vice versa. In fact, many new tools, applications and facilities are being constantly marketed. And it is true that, although there are still some limitations and difficulties when applying NLP tools in CALL, interesting and significant research is being carried out within this field. There are actually many projects, systems and research related to NLP in CALL; to mention some: Loiseau *et al.* (2005), L'haire and Vandeventer (2003), Nerbonne (2003), Heift (2003), Granger (2003), Granger (2004), etc.

As Nerbonne well points out, the central role for CALL is, or at least should be, to provide comprehensible and understandable materials for both teachers and learners. Indeed, NLP tools can be additional help resources within CALL software in order to enable language

learners and teachers to easily obtain information about the target language as well as to get some content materials. In fact, these tools can illustrate linguistic structures, make language comprehensible, provide varied exercise material, spot and correct errors, etc.

In this article, we present a CALL environment for Basque whose aims are to offer the users (teachers, learners and computational linguists) different tools and language resources in order to clarify the linguistic doubts they might have about the language as well as to store information about language learners, deviations and errors[1] as the basis for further studies in CALL and NLP.

When integrating our tools in this CALL environment, we took into account that *the use of NLP tools within a CALL software must be designed with care. Giving access to NLP tools is not enough, especially as the target user population is not already familiar with them. Therefore, careful integration of the NLP tools into the didactic concept of the CALL software is a prerequisite to benefit plainly from this innovative technology* (Linguistik online 17, 5/03).

Although Basque is a minority language, some robust NLP tools have been developed for the automatic treatment of the language in the last twenty years. In the IXA research group[2], we have created NLP tools and resources such as a morphosyntactic analyser, a shallow syntactic analyser, monolingual and multilingual dictionaries, a lexical database, a WordNet for Basque and some machine translation prototypes from English and Spanish into Basque. Some of these tools and resources have now been integrated in our CALL environment, and in addition, we have developed new ones specifically for this environment.

In sum, the CALL environment we present here is composed of Irakazi (Aldabe *et al.*, 2005), a teacher-oriented web application for analysing students' information, performances and progress; Erreus, another web application designed for storing information for errors' automatic treatment; two databases (*Errors* and *Deviations*) to store error and deviant instances with different purposes; and Lentillak, a workbench which offers several language resources and NLP tools to help Basque language learners in their tasks.

In the next section, we present the general architecture of the environment. The third section describes the NLP tools integrated in the environment. Section four describes the functionalities of the Lentillak workbench and section five deals with the experiment we have carried out with language learners in order to evaluate how they use and what they think about the integrated NLP tools. Finally, in section six, we outline some conclusions as well as our future lines.

2. The general architecture of the environment

The design and implementation of this environment involve an interdisciplinary approach claimed and followed in (Maritxalar and Díaz de Ilarraza, 1994), (Knutsson *et al.*, 2002), (Greene *et al.*, 2004), (Kraif *et al.*, 2004), and (Vitanova, 2004). According to this approach, learners, teachers and computational linguists work together in the creation of this environment that will be very useful for them. It is very interesting for learners because they can use several Natural Language Processing tools to clarify their linguistic doubts. And at

[1] We distinguish errors from deviations. We consider an error any ungrammatical output, and a deviation is, for us, the ungrammatical or inappropriate instances (such as avoidance or reiteration) made by language learners.

[2] http://ixa.si.ehu.es/Ixa

the same time, they enrich the environment with their data. The NLP tools integrated in this environment are also of great value for teachers to consult any linguistic information as well as to analyse the learning process of their students. In addition, teachers also provide the environment with learners' psycholinguistic information and deviant structures. Finally, computational linguists use the environment as a systematic and general framework to store linguistic and technical data to create new NLP tools for language learning purposes as well as for automatic error treatment.

In order to get the objectives we have mentioned before, computer engineers, computational linguists, psycholinguists and language teachers have been working together on the development of this environment. The specific goals are to develop tools for automatic error treatment and analysis as well as to collect learner corpora and store learners' psycholinguistic information.

All this will be essential in order to reach our goals. The environment offers in a systematic way a continuous feedback among the tools by means of the constant interaction among all the elements.

As figure 1 shows, we have implemented a workbench and two web applications: Lentillak, Irakazi and Erreus. In the **Lentillak** workbench some NLP tools have been integrated and they are available to clarify the linguistic doubts that arise when producing the target language. The texts written using this workbench are stored as learner corpora. By means of **Irakazi**, teachers store the information about the deviant structures found in learner corpora. It also offers teachers the chance to analyse the collected data in order to know more about the learning process of their learners and they can also specify the strategies to correct the stored deviant structures and use the integrated NLP tools that can be useful for them to correct learners' deviations. In fact, two databases, *Deviations* and *Errors*, are used to collect different aspects of learners' information. In addition, computational linguists add, by means of the **Erreus** application, the technical information corresponding to each error instance for their automatic treatment. Such information is stored in the *Errors* database.

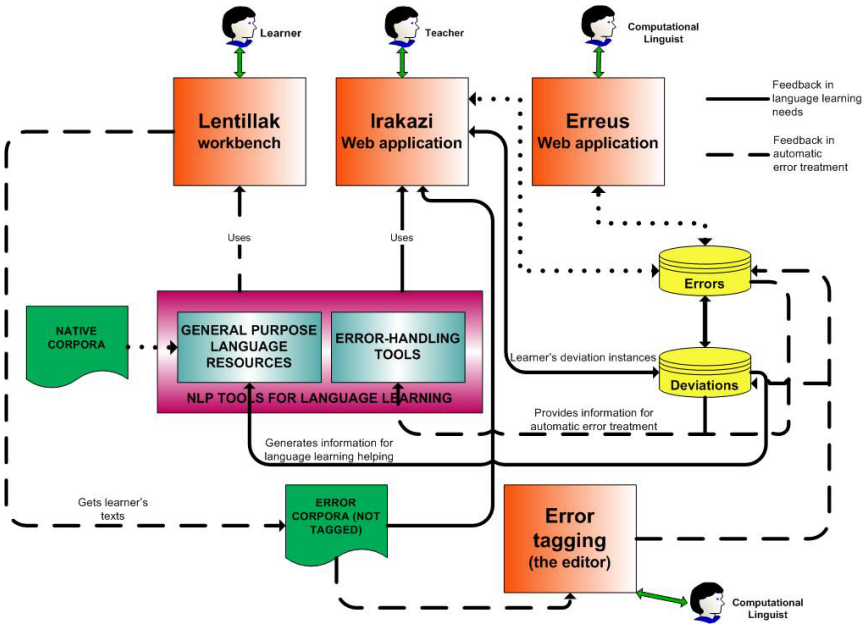


Figure 1. The CALL environment

In the near future, by means of the **error editor tool** (under construction), a computational linguist will manually tag errors and their possible corrections. The results of this tagging process will be used to automatically fill the two mentioned databases.

The **learner corpora** (named error-corpora in figure 1) we collect comprise the free-texts written by learners as well as their exercises. The results of these exercises offer information about the deviations that have been somehow induced by the teacher with pedagogical purposes. On the contrary, the free-texts provide us with the deviant structures students make unconsciously. Besides, it is important to mention that the **native** and learner corpora we make use of are organised according to the language levels set by the Basque language academies.

The **feedback** process in this environment can be defined in such a way that the deviant examples and the information stored in the *Deviations* and *Errors* databases are used to create new rules for automatic error treatment as well as to develop different language resources to respond to learners' and teachers' purposes. "General purpose language resources" and "Error-handling tools" are used in Lentillak and Irakazi, which provide us with learner corpora. And it is important to underline that the information to be introduced in the *Deviations* and *Errors* databases can be obtained by means of the error editor tool or introduced directly through Irakazi.

3. Developing and adapting NLP tools for language learning

Taking into account the needs detected and specified by Basque learners and teachers, we have developed and adapted some NLP tools. We want to remark again the robustness of the tools (morphological analyser, chunker, deep syntax analyser, etc.) used as the basis for the tools we described in this section.

We have grouped the integrated tools depending on whether they are for error detection or not. Therefore, in section 3.1 we describe the language resources concerning error-free data (general purpose language resources), and in section 3.2, we explain the work carried out in the field of automatic error detection and correction.

Basque is an agglutinative language where the constituents of the sentence are freely ordered. Because of this characteristic, linguistic tools such as a declension tool, a conjugation tool, etc. can indeed be very useful.

3.1. General purpose language resources

The adapted NLP tools we present below are based on error-free data and they have been integrated with pedagogical purposes.

- A **morphological information consulting tool** to consult the lemma entries stored in EDBL (a lexical database for Basque with approximately 85,000 entries)(Aldezabal *et al.*, 2001). In this database learners can view examples about how to use the lemmas and obtain morphological information about the entries, improving, in this way, their lexical and morphological competence. The student enters a word, the tool lemmatizes it (Aduriz *et al.*, 1992) and then consults it in the database. For example, suppose that the student wants to know if the word "dirudun" (*rich* or *who has money*) is an adjective or not. It may happen that the student does not know the lemma ("diru" or "dirudun") for consulting the word. This is not necessary as the tool obtains the lemma and it makes the query.

- A **conjugation tool** to consult verb conjugations. This tool makes use of the information stored in EDBL and it provides the verb declensions for all Basque verb forms. The users of this tool have the possibility i) to enter a person and number for each agreement marker and the mode/tense for the expected verb in order to get the corresponding auxiliary verb, or ii) to enter a person and number for each of the agreement markers, mode/tense and a root in order to get the corresponding synthetic verb form. This way, learners can get all the possible conjugations of a verb and enrich their knowledge about Basque verb system. For example, if she/he enters the pronouns "hura" (agreement marker=ABS, person=3, number=s) and "guri" (agreement marker=DAT, person=1, number=pl), and the verb root "etorri" (*to come*), the student will obtain the synthetic verb form "datorkigu" (*he comes to us*).
- A **sentence level structure helper**. This tool also makes use of the information stored in the mentioned lexical database. It is very useful for learners to get information about linguistic structures since it provides them with examples about how to use some grammatical structures. This is an interesting tool for learners to familiarise themselves with the linguistic structures of the target language as well as to enrich their grammatical competence. For example, if the student wants to know how to create "comparative" sentences, the tool will provide her/him with the suffix "-ago" (*more*) and examples of its use.
- A **declension generator tool** to find any declined form of a given word. The users have different choices such as i) to enter a word specifying its category (noun or adjective) in order to obtain all its possible declined forms, ii) to choose a word, its category and a declension case in order to get its possible variants (singular definite, plural definite and indefinite forms) of the given word, and iii) to specify a word, its category, a declension case and a declension form (singular definite, plural definite or indefinite) in order to get its declined form. This tool makes use of the morphological generator previously developed in the IXA group. For example, the student may want to know the inflection of the noun "itsaso" (*sea*) and its dative case in singular (choice iii). Introducing these data s/he will obtain the word "itsasoari" (*to the sea*).
- A **KWIC (KeyWord In Context) system** that provides access to authentic language use of different language levels. This system searches a word or a lemma in a corpus that has been clustered taking into account the different language levels specified at Basque language schools. The examples obtained from the corpus are displayed in KWIC form and they are a good training corpus for self-study.
- **Hiztegixa**, a web application where learners can look up a word in several dictionaries (monolingual and bilingual ones) as well as in a corpus, using the same interface.

At the moment, all these tools are available in the Lentillak workbench and their usefulness has already been evaluated (section 5).

3.2. Error-handling tools

Below we present the tools we have developed for the detection and correction of both language learners' deviant structures and native speakers' error instances:

- A **spelling checker** which warns users of their spelling errors.
- A **proposal tool** which offers correct proposals when the entered word is wrong. Besides, learners have the option to specify a number of proposals as well as to seek for the most typical error-types. For example, the insertion of the incorrect word "aizpa"

will provide us with correct forms “ahizpa” (*sister*) and, “aizka” (*to frighten off*), “zizpa” (*rifle*) . . . This tool is very useful when the students know more or less how to write the word but not its correct spelling. In the future, it will be also possible to ask for proposals specifying learners’ language level.

- **Grammar checkers**. For the detection of different error-types, several techniques must be used considering their linguistic requirements. Below, we explain the phenomena we have analysed and the tools we have created based on the collected errors and deviations. The first two techniques are rule-based while the last one makes use of machine learning techniques.
    - **Error detection rules using Constraint Grammar** (Karlsson *et al.*, 1983). We have created some Constraint Grammar based rules to detect errors in some grammatical expressions, determiners and postpositions. Now we are working on a proposal tool at syntactic level and have already got some results concerning postpositions.
    - **Error correction/detection rules applied to dependency trees**. We have designed and developed a system for the detection and correction of *agreement errors* in free texts based on the dependency trees of each sentence, or at least, part of it (Díaz de Ilarraza *et al.*, 2005). The system is composed of three main modules: i) a robust syntactic analyser, ii) a compiler that translates error processing rules, and iii) a module that coordinates the results of the analyser, applying different combinations of the already compiled error rules.
    - **Machine Learning**. Some machine learning techniques have been already applied to detect incorrect uses of the comma. Actually, we have tried to learn commas taking into account that no manual work had to be done, and supposing that commas were placed correctly in the chosen training corpora. In the near future, we will try some other approaches for the same purpose, and we are going to base the learning on clause splitting.

The proposal tool as well as the spelling checker are already integrated and evaluated by means of Lentillak. In the near future, we plan to integrate and evaluate the rest of the grammar checking tools.

4. The functionalities of the Lentillak workbench

Lentillak is the workbench we have used for our experiment in order to evaluate how the NLP tools help learners when making exercises. All the tools mentioned in section 3.1 as well as the spelling checker and the proposal tools described in section 3.2 are available.

The interface of Lentillak (figure 2) offers several functionalities to work with. Depending on their goals, such functionalities have been grouped in different menus. The implemented NLP tools are accessible in the *Helping tools* menu as well as in the right part of the interface. Apart from the specified menus, a notepad where users can take notes of their doubts is also accessible. The environment is implemented in Java, C++ and Perl, and it is based on a client server architecture and the communication with the user is made by means of an intuitive and easy to use web-based interface.

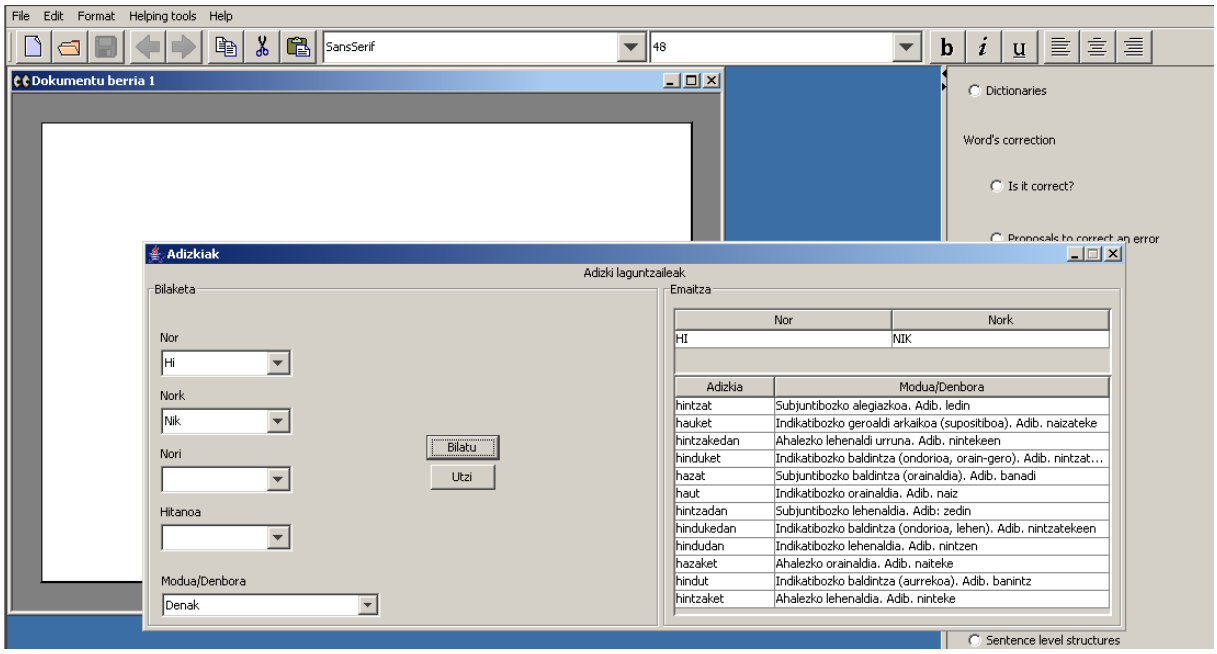


Figure 2. Lentillak's workbench with the verb conjugation tool displayed

5. An experiment with language learners

In this experiment, we wanted to evaluate with a questionnaire the usefulness of the NLP tools, *i.e.* whether the tools are helpful for learners to make exercises. An additional objective was to evaluate if the type of the exercise has influence on the usefulness of the tools (section 5.3). With this purpose, we prepared some exercises to be done by the students using Lentillak.

The experiment was carried out with twenty-five advanced learners of Basque and in three different sessions. Our objective was to measure the usefulness and comprehensibility of each tool integrated in this CALL environment and to check whether the interface of Lentillak was easy to use and intuitive.

5.1. The exercises and the questionnaire

We designed some exercises to provoke certain linguistic doubts to learners so that they could clarify them using the integrated tools. The type of exercises proposed were: i) try to find the error, ii) make a sentence with a concrete word, and iii) rewrite the sentence changing the verb tense.

In addition to the exercises, we prepared a questionnaire in order to find out i) whether they consider the interface easy to use and ii) whether they think that the NLP tools integrated in Lentillak are useful and easy to use. Respecting to this second point, we specifically made the following questions: a) did you use the tool?; a1) if so: has it been a helpful tool for you to make the exercise?; a2) if not: in spite of not using it, do you think it could be a useful tool?; b) did you understand easily the purpose and the use of the tool?

5.2. Results

As concerns the results of the first point, 88 % of the students think the interface is intuitive and easy to use. Apart from the interface's appropriate design, this high percentage may result

from the fact that the students are nowadays familiarised with this kind of computer based applications.

Table 1 shows the results of the comprehensibility and usefulness of the NLP tools.

	Comprehensibility	**Used by**	**Useful for**	**Not used but useful for**
Dictionaries	96 %	88 %	86 %	100 %
Is it correct?	100 %	72 %	94 %	86 %
Proposer	75 %	20 %	80 %	80 %
Word's declension	83 %	40 %	100 %	87 %
Verb	63 %	76 %	84 %	83 %
Sentence level structures	50 %	24 %	100 %	84 %
Gradding suffixes	50 %	4 %	100 %	79 %
Get examples	71 %	16 %	100 %	71 %
Morphological information	75 %	12 %	67 %	68 %

Table 1. Comprehensibility and usefulness of the NLP tools

With respect to *comprehensibility*, we observe that, almost all the tools are easily comprehensible for more than the 60 % of the questioned students. The tools that obtain the worst results are the ones to consult sentence level structures and gradding suffixes. We think this could be due to a higher complexity of the grammatical content presented through these tools.

Considering the *usefulness*, the second column shows how many learners have used each tool. The third column displays the percentage of those who, having used the tool, think the tool is useful. And in the last column we see the percentage of the users who, in spite of not having used the tool, think that it could be helpful.

We can also notice that *Dictionaries* (88 %), *Verb* (76 %) and *Is it correct* (72 %) are the most used tools, probably because they are very intuitive and already exist in environments of common use such as electronic dictionaries and spell checkers.

Most of the students who have used the tools consider that these linguistic resources are really interesting and useful (see column 3), and those who have not used them foresee they can be helpful (see column 4).

We consider very positive that the tools' usefulness tends to be over 80 %, even in those cases where learners have not used them to make the proposed exercises.

5.3. Influence of the exercise type for choosing the tools

Students will use some tools or others depending on the type of exercise they have to do. In order to compare the influence of the exercise type on the use of the tools, we additionally asked nine students to write an essay.

As a result of this experiment, we have deduced that the use of the tools is smaller when writing essays. In contrast to the seven tools students made use of to make the exercises, for this task they only used three of them (with different percentages of use): 33 % of the students used the *Is it correct?* tool (72 % of the students used it when making the exercises); 22 % of the students asked for correct proposals of errors (20 % in the exercise task) and 11 % of them consulted the *Dictionaries* (while 88 % used them when making the exercises).

It seems that students consider themselves quite capable of writing an essay without the help of any tool. In other words, if the exercises are not focused on provoking linguistic doubts, we foresee that advanced learners will rarely consult the tools.

Finally, it is also important to underline that the linguistic phenomena have an influence on the usefulness of the tools. table 1 shows that the tool for consulting *Gradding suffixes* has hardly been used, and this is because it involves very specific language phenomena. In any case, this would involve a deeper research we are not concerned with in this article.

6. Conclusions and Future Work

In this paper, we have presented a CALL environment where several NLP tools have been integrated (some after being adapted and some created precisely for this environment). The environment, which involves an interdisciplinary approach comprising psycholinguistics, computational linguistics and artificial intelligence, is an interesting means to collect learner corpora. Based on these data, we are able to develop new tools as well as to improve the already existing ones. All this continuous feedback is relevant in order to keep enriching the environment.

Within this environment, we find three frameworks: Lentillak, Irakazi and Erreus. The first one is mainly for language learners, the second one for teachers and the last one for computational linguists. And here, we have already integrated wide-coverage and robust tools such as a morphological information consulting tool, a conjugation tool, a sentence level structure helper, a declension generator, a KWIC system, some dictionaries, a spelling checker and a proposal tool at word level. Some more tools will be integrated in the near future. We think that the use of all these tools improves the autonomy of the student in the use of the language.

We have also carried out an experiment with language learners in order to evaluate the usefulness and the comprehensibility of the NLP tools integrated in the environment. For this experiment, we have made use of the Lentillak workbench. The results of the experiment are positive since students have really made use of the implemented tools to fulfil their learning tasks. Besides, they think most of the NLP tools are easy to use and helpful for learning Basque. However, we have demonstrated that the use of the tools depends very much on the assigned type of exercises. This experiment has been a first step in the evaluation of the tools, and in a second phase, we will evaluate what learners have learnt by using them.

A Basque language school is already in close cooperation with us to enrich our databases by means of the Irakazi web application. At the same time, thanks to their experience using the NLP tools integrated in this application, we will evaluate how helpful they are for teachers.

At the moment, the environment is used to develop different CALL systems for automatic essay evaluation and language self-study applications. We believe that pending on the availability of NLP tools for other languages, the environment could become multilingual.

7. Acknowledgements

This research is being partially supported by the Basque Government (SAIOTEK project, SPE04UN11) and the Spanish Ministry of Education and Science (HIZKING21 project, TIN2004-07918-C04-01). We would like to thank Koldo Gojenola and Aitor Sologaistoa for their help and collaboration in the development of some of the tools we have presented in this paper and to the Jakintza secondary school for their help when carrying out the experiment.

References

ADURIZ I., AGIRRE E., ALEGRIA I., ARREGI X., ARRIOLA J., ARTOLA X., DÍAZ DE ILARRAZA A., EZEIZA N., MARITXALAR M., SARASOLA K., URKIA M. (1992). "A morphological analyzer for basque based on two-level morphology". In *Proceedings of the 5th Int. Morphology Meeting*. Krems.

ALDABE I., AMOROS L., ARRIETA B., DÍAZ DE ILARRAZA A., MARITXALAR M., ORONOZ M., URIA L. (2005). "Irakazi: a web-based system to assess the learning process of basque language learners". In *Proceedings of a one-day conference Natural Language Processing in Computer-Assisted Language Learning*.

ALDEZABAL I., ANSA O., ARRIETA B., ARTOLA X., EZEIZA A., HERNÁNDEZ G. LERSUNDI M. (2001). "EDBL: a general lexical basis for the automatic processing of basque". In *IRCS Workshop on linguistic databases*. Philadelphia.

DÍAZ DE ILARRAZA A., GOJENOLA K., ORONOZ M. (2005). "Design and development of a system for the detection of agreement errors in basque". In *CICLing-2005, Sixth International Conference on Intelligent Text Processing and Computational Linguistics*. Mexico City.

DÍAZ DE ILARRAZA A., MARITXALAR A., MARITXALAR M., ORONOZ M. (1999). "Idazkide: an intelligent call environment for second language acquisition". In *Proceedings of a one-day conference Natural Language Processing in Computer-Assisted Language Learning.* ReCALL.

GRANGER S. (2003). "Error-tagged learner corpora and CALL: a promising synergy". In *CALICO* (special issue on Error Analysis and Error Correction in Computer-Assisted Language Learning) 20 (3): 465-480.

GRANGER S. (2004). "Computer learner corpus research: current status and future prospects". In U. Connor and T. Upton (eds.), *Applied Corpus Linguistics: A Multidimensional Perspective*. Rodopi, Amsterdam/Atlanta: 123-145.

GREENE C., KEOGH K., KOLLER T., WAGNER J., WARD M.,VAN GENABITH J. (2004). "Using NLP technology in CALL". In *Proceedings of InSTIL/ICALL2004*.

HEIFT T. (2003). "Multiple Learner Errors and Meaningful Feedback: A Challenge for ICALL Systems". In *CALICO* 20 (3): 533-549.

KARLSSON F., VOUTILANEN A., HEIKKILA J., ANTTILA A. (1983). *Constraint Grammar: Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter, New York-Berlin.

KNUTSSON O., CERRATTO PARGMAN T., SEVERINSON EKLUNDH K. (2002). "Computer support for second language learners' free text production - initial studies". In *European Journal of Open and Distance Learning (EURODL)*. Martin Valcke and Anne Bruce.

KRAIF O., ANTONIADIS G., ECHINARD S., LOISEAU M., LEBARBÉ T., PONTON C. (2004). "NLP tools for CALL: the simpler, the better". In *Proceedings of InSTIL/ICALL2004 Symposium, NLP and Speech Technologies in Advanced Language Learning Systems*.

LOISEAU M., ANTONIADIS G., PONTON C. (2005). *Third International Conference on Multimedia and Information & Communication Technologies in Education* (MICTE2005). Badajoz.

L'HAIRE S., VANDEVENTER A. (2003). *Using NLP tools in a CALL software: the FreeText project*.

MARITXALAR M., DÍAZ DE ILARRAZA A. (1994). "An ICALL system for studying the learning process". In *Computers in Applied Linguistics Conference*.

NERBONNE, J. (2003). "Natural Language Processing in Computer-Aided Language Learning". In R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*. Oxford University, Oxford.

VITANOVA I. (2004). "Evaluating integrated NLP in foreign language learning: technology meets pedagogy". In *Proceedings of In-STIL/ICALL2004*.

Le TALN au service de la didactique du français langue étrangère écrit

Isabelle Audras, Jean-Gabriel Ganascia

Université Pierre et Marie Curie – LIP 6
{isabelle.audras ; jean-gabriel.ganascia}@lip6.fr

Résumé

De nouveaux logiciels d'analyse textuelle tirent partie des progrès récents effectués en apprentissage symbolique et dans le traitement automatique des langues naturelles. Conçu au LIP6 par Jean-Gabriel Ganascia, le *Littératron* est l'un d'entre eux ; il extrait automatiquement des motifs syntaxiques[1] à partir de textes écrits en langage naturel. Plus exactement, le *Littératron* prend comme entrée un arbre d'analyse syntaxique et donne en sortie un certain nombre de motifs syntaxiques récurrents. Associé à un analyseur de textes, qui engendre l'arbre d'analyse syntaxique à partir de textes écrits en langage naturel, il révèle les singularités stylistiques de ces textes.

Nous allons voir qu'utilisé en sciences du langage, dans le domaine de l'acquisition du français écrit, le *Littératron* permet d'effectuer un diagnostic linguistique de l'apprenant, que celui-ci provienne d'une classe de langue hétérogène (différentes langues maternelles) ou homogène (une seule langue maternelle, en l'occurrence ici l'arabe). L'intérêt de cette approche concerne trois domaines : d'une part la didactique des langues, à titre éducatif ; d'autre part, la linguistique computationnelle, et enfin l'enseignement assisté par ordinateur.

Mots-clés : acquisition d'une langue étrangère écrit, didactique de l'écrit en langue étrangère, TALN, extraction de motifs récurrents, stylistique, diagnostic linguistique.

Abstract

New text analysis softwares issued from fields of research such as Machine Learning and Natural Languages Processing prove to be relevant tools for the language sciences. *Littératron* is a new data-processing tool for the automatic extraction of syntactic patterns, designed at LIP6 by Jean-Gabriel Ganascia. Associated with a linear text analyser, it reveals the stylistic peculiarities of a text.

We will see that *Littératron* carries out a linguistic diagnosis of learners if used in language sciences, especially in the field of acquisition of written French as a foreign language. The learner can be from a heterogeneous group (various language levels and various mother tongues) or from a homogeneous group (only one language level and one mother tongue, here, Arabic). The interest of this approach is related to three fields: first, language didactics, on a purely educational basis; next, computational linguistics; finally, computer-assisted learning.

Keywords: foreign-language acquisition, foreign-language written didactic, NLP, stylistics, extraction of recurrent patterns, linguistic diagnosis.

1. Cadre théorique

Les recherches sur l'acquisition de l'écrit en langue étrangère sont récentes. Cependant elles ont bénéficié des résultats des recherches concernant l'acquisition de l'écrit en langue maternelle. Prise de recul par rapport à la langue, aide à la mémorisation : les vertus cognitives du passage à l'écrit ne sont plus à démontrer (Mangenot, 1998).

[1] Un motif syntaxique est une association d'unités linguistiques cohérentes.

De plus les résultats des recherches en linguistique textuelle croisent également les intérêts de la didactique de l'écrit. En effet les notions de cohérence textuelle et de pragmatique, au centre de la linguistique textuelle, sont porteurs du développement chez l'apprenant d'une compétence textuelle qui lui rend disponibles des outils d'articulation en vue de construire un discours. Autrement dit, « les structures textuelles formelles [...] guident le scripteur dans la construction d'un texte et le lecteur dans sa compréhension » (Scardamalia et Bereiter, 1986)

Selon Tuffs (1993), travailler sur des genres textuels différents facilite l'acquisition des langues étrangères. De façon générale, l'écrit en classe de langue est associé à une consigne qui prévoit l'intention de communication, même à l'extérieur d'un genre. En effet, le cadre narratif choisi, par le genre ou la consigne, définit un objectif de communication précis. Celui-ci appelle des objectifs fonctionnels dont l'expression morphosyntaxique et lexicale est vue en classe. Ce contenu linguistique, découvert à l'intérieur d'une situation de communication, est automatisé lors de réemplois, et ceci est d'autant plus vrai si celui-ci se trouve dans un contexte similaire. Enfin, l'analyse des besoins communicatifs du cadre narratif aide l'apprenant à s'adapter face à une nouvelle situation de communication dans laquelle il doit réagir (Tagliante, 1994).

Par ailleurs, l'apprentissage du FLE est sanctionné par une certification appelée DELF (Diplôme d'Études en Langue Française) aligné sur le cadre européen commun de référence dans l'apprentissage des langues. Les épreuves écrites A1, A2 et A3 ont pour cadre narratif, respectivement : la carte postale, la lettre amicale, la lettre de motivation.
Dans ces deux types de production écrite en français langue étrangère, le niveau de l'apprenant est validé par rapport à sa capacité à exprimer un message à travers un modèle appris et reconnu et non simplement par rapport à ses compétences grammaticales.
Autrement dit, la production écrite en classe de langue est le reflet des compétences de l'apprenant lors du passage à l'écrit. Ses compétences se révèlent à la fois dans la fréquence des expressions observées, dans ses prises de risques et dans l'originalité de ses idées (Carroll et Stutterheim, 1997).

C'est pourquoi nous souhaitons repérer, grâce aux techniques actuelles du traitement automatique des langues, les erreurs écrites usuelles d'une population d'apprenants, ce qui permettra de mettre l'accent, au cours de l'enseignement, sur la correction de ces erreurs.

Ce repérage des erreurs peut se faire soit dans l'absolu, par détection des erreurs syntaxiques, soit par rapport aux usages, par une étude des tournures propres à une catégorie d'apprenants dans un cadre narratif précis, celles-ci se trouvant absentes ou peu usitées chez les locuteurs natifs. C'est cette seconde approche que nous avons adoptée, sachant que le rôle des enseignants de langue n'est pas d'enseigner une langue abstraite parfaite mais de transmettre les usages d'une langue.
Plus exactement, le travail présenté ici recourt à l'emploi d'outils d'analyse stylistique pour dégager les caractéristiques des apprenants, selon leur niveau, et les distinguer des locuteurs natifs. Des études empiriques conduites autour de trois populations d'apprenants, l'une à Paris, à l'Alliance Française, l'autre à l'université de Naplouse (Territoires Palestiniens), auprès d'un public arabophone et la troisième à l'École Normale de Port-au-Prince (Haïti) auprès d'étudiants créolophones valident l'approche proposée.

2. Présentation des outils informatiques utilisés

Un motif syntaxique se définit comme une association d'unités linguistiques cohérentes, par exemple : [préposition + pronom personnel réfléchi + verbe à l'infinitif]. Chaque motif, par exemple le motif précédent extrait automatiquement à partir des outils mis en œuvre, peut

recevoir un nombre plus ou moins grand de réalisation dans un texte donné. Voici un exemple des réalisations de ce motif dans des textes que nous avons traités : “de vous adresser”, “afin de vous donner”, “de m’investir”, “de vous donner”. Ces quatre réalisations ont été extraites ensemble d’un même groupe de scripteurs de lettres de motivation.

Deux outils informatiques sont nécessaires pour extraire les motifs syntaxiques caractéristiques des textes écrits par les différentes populations. Le premier outil informatique requis est un analyseur morphosyntaxique du français qui construit des arbres syntaxiques à partir de productions écrites. Nous avons eu recourt à l’analyseur linéaire avec dictionnaire partiel de Vergne qui a été élaboré par Jacques Vergne de l’Université de Caen, en 1998 (Vergne, 2001). Le second est l’analyseur stylistique *Littératron*, mis au point au LIP6 par Jean-Gabriel Ganascia (2001) qui dégage les motifs syntaxiques récurrents présents dans ces arbres.
Plus exactement, à chaque mot ou groupe de mots l’analyseur de Vergne associe une étiquette ; un arbre stratifié est donc une partition d’étiquettes dont les classes dépendent de la profondeur du nœud dans l’arbre d’analyse. Étant donnée une structure d’ASO, le *Littératron* calcule une mesure de similarité entre plusieurs ASO, fondée sur la notion de distance d’édition, et génère un graphe de similarité enregistrant les sous-arbres les plus proches de l’ASO donné en entrée.

C’est ce graphe de similarité qui sert ensuite d’entrée à l’algorithme de classification du *Littératron*, appelé ‘centre-étoiles’, qui construit des classes de motifs similaires et leur attribue un nom significatif. En effet, l’algorithme centre-étoile évalue d’abord l’ensemble des étoiles centrées sur les différents nœuds puis il prend, pour chacune, la somme des valeurs de similarité des nœuds de chaque étoile au centre. Une fois calculée la valeur de chaque étoile, l’algorithme ‘centre-étoiles’ prend celle qui à la plus forte évaluation. On marque ensuite, les nœuds qui appartiennent à cette première étoile, avant d’appliquer récursivement le même algorithme sur les nœuds non marqués, jusqu’à épuisement des nœuds non marqués.

En résumé, toute étoile est un sous-graphe du graphe de similarité centré sur un nœud. Pour chaque classe ainsi construite, l’algorithme choisit les motifs les plus similaires au centre de l’étoile, pour illustrer la signification de l’étoile. Il indique aussi le texte source couvert par chacun des motifs.

Voici l’exemple d’un centre d’étoile, illustré par la figure 1 :

[PREP [“de”]] + [GN [ART [“la”]] + [NOM [“forêt”]]] (texte : “de la forêt”), auquel sont associés les 5 motifs syntaxiques suivants :

- [PREP [“à”]] + [GN [ART [“l”]] + [NOM [“auberge”]]] (texte : “à l\’auberge ”) ;
- [PREP [“d\”]] + [GN [ART [“un”]] + [NOM [“hiver”]]] (texte : “d\’un hiver”) ;
- [PREP [“dans”]] + [GN [ART [“le”]] + [NOM [“monde”]]] (texte : “dans le monde”) ;
- [PREP [“avec”]] + [GN [ART [“les”]] + [NOM [“chiens”]]] (texte : “avec les chiens”) ;
- [PREP [“depuis”]] + [GN [ADJ [“quelques”]] + [NOM [“jours”]]] (texte : “depuis quelques jours”).

Ceci signifie que la mesure de similarité entre le premier motif (‘de la forêt’) et l’un des arbres dérivés des arbres syntaxiques de chacun de ces cinq groupes nominaux est supérieure à un certain seuil. Ces cinq motifs font partie de la même étoile dont le centre est de la forme : [PREP [“de”]] + [GN [ART [“la”]] + [NOM [“forêt”]]].

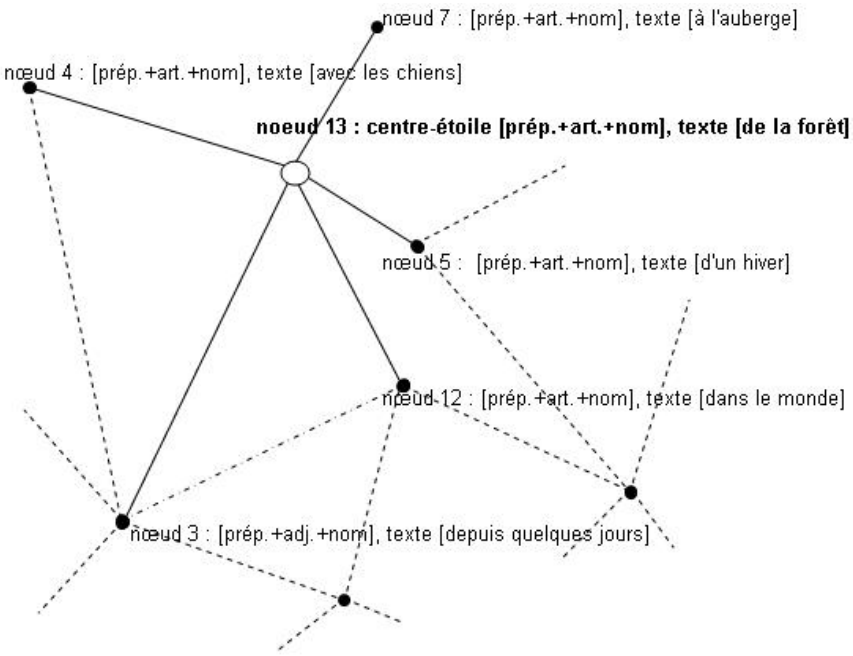


Figure 1. Graphe du centre-étoile présenté en exemple

Outre la construction d'étoiles et l'extraction de motifs, le *Littératron* procède à un second type d'opérations qui consiste à comparer les étoiles issues de plusieurs textes afin de repérer les étoiles présentes dans l'un et absentes dans l'autre. Ceci permet de discriminer, parmi les motifs présents dans une production, ceux qui le distinguent d'autres productions. C'est à partir de ce type de discrimination que l'on construira les tournures caractéristiques de populations d'apprenants.

Par exemple, voici l'une des sorties du *Littératron* analysant trois groupes de scripteurs distincts à partir de lettres de motivation (LM app : lettres de motivation d'apprenants de langues maternelles diverses, LM appa : lettre de motivation d'apprenants arabophones, LM frcph : lettres de motivation de francophones).

Patron N°46

- Indépendante

  S
  P p 1 s
  V
  COD
  N c f s

- Exemples:

  Fichier LMapp: Je parle la langue anglaise et française
  Fichier LMfrcph: Je maîtrise la mise en place de l'organisation de l'archivage
  Fichier LMappra: J'apprends la presse à l'université de Naplouse

Le patron décrit la structure syntaxique du motif extrait. Ici il s'agit d'une proposition indépendante de forme sujet S + verbe V + COD. Il y est précisé que le sujet est un pronom personnel à la première personne du singulier (Pp 1 s) et que le COD est un nom commun féminin singulier (Nc f s).

Les exemples extraits de chaque groupe de scripteurs donnent un aperçu des différents textes en langage naturel que le *Littératron* détecte comme étant proches de cette structure centrale.

3. Premier type d'expérience : les apprenants sont de langues maternelles diverses

L'idée de cette recherche est de comparer des productions écrites en classe de FLE de différents niveaux avec des productions de francophones répondant aux mêmes consignes. Les scripteurs francophones sont des natifs français de niveau d'étude au moins équivalent à bac+4. Une autre étude, qui pourrait se révéler intéressante, travaillerait avec des francophones d'un niveau d'études moins élevé, voire des débutants scripteurs adultes (Morais et Kolinsky 2001). L'idée de cette perspective est de montrer que le critère du niveau d'éducation n'est pas négligé dans cette approche.

3.1. Présentation des productions écrites et méthodologie expérimentale

Quatre types de production ont été choisis : la carte postale (CP), la lettre amicale (LA), la lettre de motivation (LM), la description (Des). Chaque production correspond à un niveau d'apprentissage du français langue étrangère. Quant à la description, chaque apprenant, tout niveau confondu, est soumis à l'observation puis à la description écrite d'un même dessin en couleurs de format A3 (place de village, art naïf).
Toutes les productions d'apprenants ont été faites en classe, entre le mois d'avril et le mois de juin 2002. La plupart se sont déroulées à l'Alliance Française de Paris. Certaines descriptions ont été réalisées dans une formation en FLE et en alphabétisation dans le Foyer de travailleurs Pinel, à Saint Denis.

Le tableau 1 a une double fonction. Premièrement, il récapitule les expérimentations réalisées par genre textuel. Par exemple : en ce qui concerne la 'carte postale' (CP), vont être introduits simultanément dans les analyseurs les productions d'apprenants débutants et de francophones. Deuxièmement, il détaille le nombre total de production de chaque type.

Concernant la description, les productions des 4 groupes de scripteurs sont introduites en même temps dans les analyseurs.

	apprenants			
	débutants (niveau A1 du CECR[2])	intermédiaires (A2 niveau du CECR)	avancés (A3 niveau du CECR)	francophones
carte postale (CP)	6			6
lettre amicale (LA)		4		4
lettre de motivation (LM)			6	6
Description (Des)	5	5	5	5

Tableau 1. Tableau récapitulatif des productions et leur nombre

[2] CECR : Cadre Européen Commun de Référence.

3.2. Résultats et commentaires

Les résultats obtenus sont de nature statistique, auxquels nous ajoutons des commentaires linguistiques sur les motifs extraits.

	CP déb.	CP frcph.	LA inter.	LA frcph.	LM av.	LM frcph.	Des deb.	Des inter.	Des. av.	Des frcph
nb étoiles	6	10	2	5	6	6	2	3	3	13
% texte	50	50	60	30	25	17	33	33	35	14

Tableau 2. Nombre d'étoiles et pourcentage de texte représenté par celles-ci

Le tableau 2 ci-dessus donne les résultats numériques des calculs statistiques effectués par l'analyseur. Il indique, pour chaque classe de scripteurs (francophones : frcph ; apprenants débutants : deb ; apprenants intermédiaires : inter ; apprenants avancés : av) et pour chaque type de production, le nombre d'étoiles détectées par le *Littératron* ainsi que le pourcentage de texte représenté par ces étoiles. Les paramètres d'entraînement du *Littératron* sont identiques sur tous ces ensembles de productions, en particulier les seuillages de l'algorithme centre étoile et du graphe de similarité. Autrement dit, l Le nombre d'étoiles détectées est donc un bon indicateur de la richesse stylistique : plus il y a d'étoiles, plus le style est riche, c'est-à-dire moins les automatismes prévalent. Il en va de même pour le pourcentage de texte couvert par les étoiles détectées : plus celui-ci est faible, plus les patrons varient, ce qui signifie que le style est plus riche.

Notons que cette notion de richesse stylistique doit être relativisée ; en effet, un grand écrivain pourrait se caractériser par la singularité d'un style qui déclinerait une palette restreinte de patrons, tandis qu'un écrivain sans style les déploierait tous. En dépit de ces quelques réserves, dans le cas particulier de la didactique qui nous intéresse, nous assimilons la richesse d'un texte (ou d'un ensemble de textes) au nombre de figures syntaxiques employées.

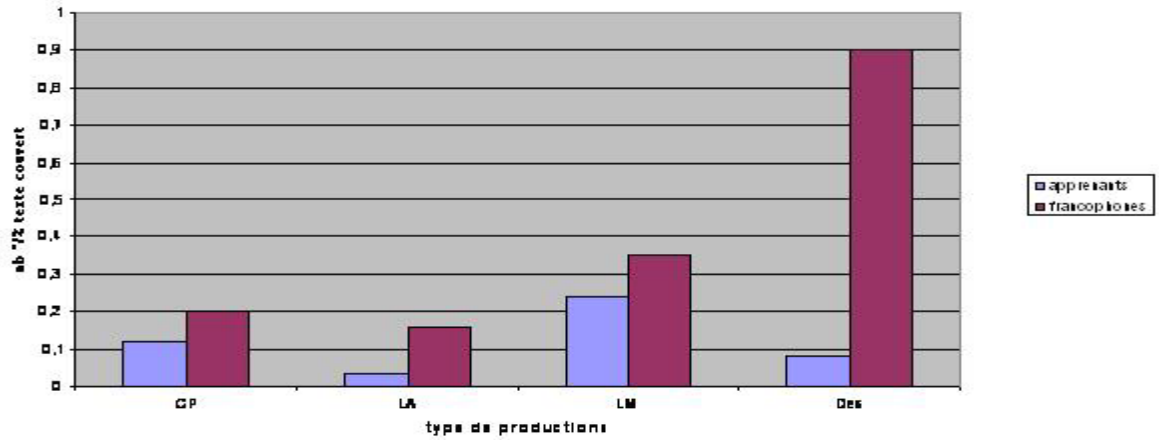


Figure 2. Indice de variabilité en fonction du type de production

Sur le graphe de la figure 2, nous définissons un indice de variabilité qui est, pour chaque type de texte, le rapport du nombre d'étoiles détectées sur le pourcentage de texte utilisé par l'application.

Que conclure du nombre de motifs syntaxiques récurrents et du pourcentage de texte recouvert par ces motifs ? D'une part, les résultats statistiques représentés par l'indice de variabilité nous montre que pour un même genre de production écrite, les motifs syntaxiques retenus par l'application sont plus nombreux, divers et dans une proportion de texte plus petite chez les francophones que chez les apprenants. De plus, la partie de texte non recouvert par les motifs syntaxiques récurrents varie dans un rapport 2 (pour les CP, LM et LA) à 9 (pour la Des) fois plus important chez les francophones que chez les apprenants, même les plus avancés. Cette partie de texte, où le *Littératron* n'a pas détecté de motifs récurrents, pourrait être utilisée pour définir l'originalité du scripteur.

Cette analyse a révélé des automatismes de l'écrit à l'intérieur de certains types de production. Ces automatismes concernent aussi bien des textes d'apprenants du français que ceux des francophones. Il y a donc des matrices d'écriture de cartes postales, de lettres d'invitation ou de lettres de motivation. Pour ce qui concerne les descriptions, la comparaison entre les différents niveaux fait apparaître des fréquences de motifs qui évoluent vers une complexification dans la composition et les liens de dépendance, donc une aisance d'écriture qui s'installe au fur et à mesure que la compétence morpho-syntaxique s'acquiert.

Enfin, concernant la description, nous sommes en mesure de rajouter quelques commentaires sur la structure syntaxique des motifs extraits. Les motifs de base extraits en qualité de syntagme nominal et en qualité de syntagme verbal ont, respectivement, la composition suivante : préposition + substantif + adjectif qualificatif et pronom sujet + verbe + adverbe. Ces motifs de base s'enrichissent progressivement en fonction de la maîtrise du français écrit. Par exemple, le syntagme nominal « des paysages variés » issu d'une production d'un scripteur débutant évolue en « un paysage bien vert » chez un scripteur natif francophone.

4. Deuxième type d'expérience : les apprenants sont de même langue maternelle

L'hypothèse est de déceler une spécificité de la syntaxe française au sein d'une classe homogène.

4.1. Population arabophone

4.1.1. Présentation des productions écrites

L'ensemble des productions analysées correspond aux examens d'histoire et de civilisation du 1er semestre 2004 d'étudiants en 3e année du département de français de l'Université An-Najah de Naplouse (Territoires Palestiniens). Les étudiants de l'université sont tous de langue maternelle arabe, l'anglais est leur première langue étrangère, le français leur deuxième.

Les productions de chaque type sont au nombre de dix.

4.1.2. Résultats et commentaires

Trois mêmes motifs syntaxiques ressortent systématiquement des productions arabophones. Ces motifs recouvrent ¼ du texte analysé. Il s'agit de deux motifs nominaux et d'un verbal. Les deux motifs nominaux sont de construction : DE + adjectif + nom, comme dans les exemples : « de choses magnifiques » et « d'autres villages ». :

Motifs présents dans la constellation issue de ::atelier:civilisatioarabph.prs
et absents de la constellation issue de ::atelier: civilisatioarabfrph.prs
Motif N°151 (étoile N°7) :
V s1
P k ne
C V s1 fatigue
P l pas
Expression couverte : ne fatigue pas
Motif N°184 (étoile N°11) :
N fp3
I o de
C S fp3 choses
P E fp3 magnifiques
Expression couverte : de choses magnifiques
Motif N°183 (étoile N°13) :
N ms3
I o d'
P a ms3 autre
C S ms3 village
Expression couverte : d' autre village

Figure 3. Sortie du Littératron

Ce motif syntaxique révèle une utilisation massive de groupes nominaux compléments du verbe de forme adjectif + nom commençant par DE, au détriment d'autres articles et d'autres prépositions. L'étudiant, à défaut de connaître la bonne rection d'un verbe, directe ou indirecte, ou le bon emploi de l'article partitif sur le défini ou l'indéfini, va utiliser systématiquement la préposition DE pour introduire ses compléments d'objet.
L'apprenant ne semble pas maîtriser une bonne utilisation des prépositions et des articles. Cette expérience révèle pour une classe homogène d'apprenants arabophones, une spécificité de la syntaxe française dont l'acquisition nécessite un accompagnement particulier.

4.2. Population créolophone

Une expérimentation en cours, sur des productions de même consigne, auprès d'étudiants de l'École Normale de Port-au-Prince (Haïti) semble converger vers la même conclusion, avec une flagrante récurrence de motifs extraits de la forme : de + adjectif + nom.

5. Conclusion

Utilisé en sciences du langage dans le domaine de l'acquisition en langue étrangère du français écrit, le *Littératron* est en mesure d'effectuer un diagnostic linguistique de l'apprenant sur des productions au cadre narratif contraignant. Ce diagnostic sert à l'accompagnement à l'acquisition de l'écrit.
À terme, ce travail doit faire l'objet de deux types de développements complémentaires, sur les plans technique et expérimental.

D'un côté, nous nous sommes limités ici à une décomposition en syntagmes, et à une étude de la structure de la phrase relativement à cette décomposition. Cela restreint assez fortement le type de motifs détectés. Nous allons faire appel à une décomposition plus riche qui prendra en compte la structure propositionnelle. L'algorithme d'extraction de motifs est identique, mais

l'analyse syntaxique diffère ; surtout, l'arbre résultant de cette analyse doit être considérablement enrichi.

D'un autre côté, les résultats obtenus auprès d'étudiants arabophones et créolophone nous encourage à poursuivre plus loin l'étude des différences spécifiques auprès d'apprenants venant de différentes régions du monde, et de langues maternelles diverses, sur des productions de DELF.

Références

ADAM J.-M. (1992). *Les textes : types et prototypes. Récit, description, argumentation, explication et dialogue*. Nathan, Paris.

AUDRAS I. GANASCIA J.-G. (2005). « Analyses comparatives de productions écrites d'apprenants du français et de locuteurs francophones, à l'aide d'outils d'extraction automatique du langage ». In *Apprentissage des Langues et Système d'Information et de Communication (ALSIC)* 8 : 81-94. http://alsic.u-strasbg.fr/v08/audras/alsic_v08_16-rec10.htm

BESSE J.-M. (dir.) (2003). *Qui est illettré ? Décrire et évaluer les difficultés à se servir de l'Écrit*. Retz, Paris.

CARROLL M., STUTTERHEIM Ch. (1997). « Relations entre grammaticalisation et conceptualisation et implications sur l'acquisition d'une langue étrangère ». In *Acquisition et Interaction en Langue Étrangère (AILE)* 9 : 14-19.

GANASCIA J.-G. (2001). « Extraction automatique de motifs syntaxiques ». In *Actes de Traitement Automatique du Langage Naturel 2001* (*TALN 2001).*

GANASCIA J.-G. (2001). « Extraction of Recurrent Patterns from Stratified Ordered Trees". In *Actes de Machine Learning : 12th European Conference of Machine Learning 2001 (ECML 2001)* : 167-179.

GANASCIA J.-G. (2004). "Detection of Statistically Abnormal Patterns from Stratified Ordered Trees". In Milutinovic, Vujovic Milutinovic (dir.) *Advances in the Internet Technology, Concepts and Systems.*

GIGUET E. (1998). *Méthode pour l'analyse automatique de structures formelles sur documents multilingues*. Thèse de doctorat en informatique, Université de Caen. http://users.info.unicaen.fr/ ~giguet/ these/

KAHN G. (dir.) (1993). *Des pratiques de l'écrit*. Numéro spécial *Le Français dans le Monde*. Hachette., Paris.

MANGENOT F. (1998). « Outils textuels pour l'apprentissage de l'écriture en L1 et en L2 ». In M. Souchon (dir.). *Pratiques discursives et acquisition des langues étrangères*. *Actes du Xe colloque international* « *Acquisition d'une langue étrangère : perspectives et recherches* » : 515-525.

MOIRAND S. (1990). *Une grammaire des textes et des dialogues*. Hachette FLE, Paris.

MORAIS J., KOLINSKY R. (2001). « The literate mind and the universal human mind ». In E. Dupoux (ed.) *Language, brain and cognitive development : Essays in Honor of Jacques Mehle.* MIT, Cambridge, Mass : 463-480.

PERY-WOODLEY M.-P. (1993). *Les écrits dans l'apprentissage. Clés pour analyser les productions des apprenants*. Hachette FLE, Paris.

SCARDAMALIA M., BEREITER, C. (1986). « Research on written composition ». In M.C. Wittrock (dir.). *Handbook of research on teaching*, McMillan, New York : 778-803.

TAGLIANTE C. (1994). *La classe de langue.* CLE International, Paris.

TUFFS R. (1993) « A genre approach to writing in the second language classroom : the use of direct mail letters ». In *Revue belge de philologyphilologie et d'histoire* 71 : 691-721.

VERGNE, J. (1999). *Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur. Analyse syntaxique non combinatoire Synthèse et résultats. Habilitation à Diriger des Recherches*, Université de Caen. http://users.info.unicaen.fr/~jvergne/#HDR.

VERGNE, J. (2001). « Analyse syntaxique automatique de langues : du combinatoire au calculatoire ». In *Actes de Traitement Automatique du Langage Naturel 2001 (TALN 2001)*. http://users.info.unicaen.fr/~jvergne/Taln2001FR_JV.pdf

Automated Analysis of Students' Free-text Answers for Computer-Assisted Assessment[1]

Matthieu Hermet[1], Stan Szpakowicz[2], Lise Duquette[3]

[1] University of Ottawa, School of Information Technology and Engineering
[2] Polish Academy of Sciences, Institute of Computer Science
{mhermet,szpak}@site.uottawa.ca

[3] University of Ottawa, Second Language Institute
lduquett@uottawa.ca

Résumé

Nous présentons une approche dans le domaine de la Correction Assistée par Ordinateur qui repose sur une analyse symbolique des entrées. La théorie qui soutient cette approche repose sur un travail préalable de mise au point d'un tutoriel d'aide à la comprehension en lecture du français, *Didalect*. Cette théorie doit rendre possible le traitement de segments de texte libre à fins de correction automatique sans l'aide d'une base de réponses pré-encodées. Une étude basée sur un ensemble réduit de réponses étudiantes à plusieurs types de questions nous a permis de justifier notre approche, d'établir une méthodologie et d'élaborer un prototype.

Mots-clés : analyse symbolique, correction assistée par ordinateur, enseignement des langues assisté par ordinateur, tutoriels intelligents, traitement automatique du langage.

Abstract

We present an approach to Computer-Assisted Assessment of free-text material based on symbolic analysis of student input. The theory that underlies this approach arises from previous work on *DidaLect*, a tutorial system for reading comprehension in French as a Second Language. The theory enables processing of a free-text segment for assessment to operate without precoded reference material. A study based on a small collection of student answers to several types of questions has justified our approach, and helped to define a methodology and design a prototype.

Keywords: computer-assisted-assessment, computer-assisted language learning, intelligent tutoring systems, natural-language processing, symbolic analysis.

1. Introduction

The literature on Computer-Assisted Language Learning (CALL) sometimes presents this field in the more general context of Computer-Assisted Learning. This may be fitting, because learning occurs mainly through language. Viewed as a computing and engineering problem, CALL requires contributions in such areas as Virtual Learning Environments (VLE), Computer-Assisted Assessment (CAA) and Intelligent Tutoring Systems (ITS), perhaps controlled by Learning Management Systems (LMS). The success of these complex tools may rely on adaptability and feedback. This calls for solutions motivated by Artificial Intelligence (AI) and Natural Language Processing (NLP).

[1] This work has been partially supported by the Social Sciences and Humanities Research Council of Canada, the program « Initiative on the New Economy ».

CALL is a relatively new field, balanced between the somewhat opposing goals of reusability and didactic specificity. There is a tension between the need to design generic modules (SCORM/ADL, 2006) and very specific systems (for example: Chen and Tokuda, 1999). The former enables the application of sound software engineering principles (without losing sight of commercial potential), while the latter makes didactically attractive solutions possible. Examples of CALL systems with different pedagogical concerns are presented in Graesser (2005), VanLehn (2002), Vitanova (2004) and Michaud (2001). Perez *et al.* (2004, 2005) give an overview of existing CAA solutions and systems as a background in the presentation of their own system.

A prototype VLE *DidaLect* (Desrochers *et al.*, 2004 ; Balcom *et al.*, 2006) has been constructed at the University of Ottawa. It is meant to help students of French as a Second Language (FSL) improve reading comprehesion. It is designed to meet a number of didactic and cognitive goals. Part of the *DidaLect* project is a study of the feasibility of automatic recognition and assessment of free-text answers to certain types of open test questions that the system uses to evaluate the student's progress. This is an autonomous project. Its first prototype will soon be completed, to help assess the validity of the approach. The novelty is that answer assessment does not rely on a comparison with precoded reference material. In this paper we present a theoretical framework of the project and a central procedure. Section 2 briefly discusses *DidaLect*, sections 3 and 4 present our method of assessing free-text answers, and section 5 lists future work and conclusions.

2. DidaLect

DidaLect is an adaptive ITS designed to enhance the reading skills of intermediate and advanced FSL students who work without teacher supervision. The system has a solid grounding in theories coming from the fields of education, cognition and psycholinguistics. Its VLE includes a placement test, a collection of tutorial texts, and resources that assist in the acquisition of reading skills: dictionaries, comprehension tests and *bloc-notes* which help activate prior knowledge. Duquette and Desrochers (2006) present a detailed description of *DidaLect*'s tutoring content.

Text comprehension consists in understanding communication goals expressed in language and accessible through cognitive operations of sense acquisition (understood as sets of *propositions*) and the student's conscious use of such operations (Denhière and Baudet 1992). These, in turn, depend on declarative and procedural knowledge. The former includes syntactic and lexical facts, either encyclopaedic (searchable in a knowledge base) or inferred (by detecting and using the required material in the text). The latter involves mainly iteration and integration: the first as a means of decomposing the steps for structure-building, the second as a means of guaranteeing coherence in the building process. This theory, which underlies *DidaLect*'s implementation, helps delimit the nature of questions that can be asked. We believe that having firm theoretical grounds for the question types is a large part of the job of implementing free-text answer analysis in a CAA module.

Exercises are necessary to assess and help develop comprehension skills indirectly. In language learning, various types of texts are applied, with various types of questions. There are two objectives: variable degree of text and question difficulty, and high adaptability to a student's learning history. *DidaLect* uses informative texts in four categories: cause-effect, problem-solution, comparison and description (Richgels *et al.,* 1987). Questions built upon this textual material range over two categorization scales based on the relation to text structure and cognition, and on the relation to information. The first scale divides questions

into Text-Explicit (TE), based on one sentence, Text-Implicit (TI), which require inference between two or more text segments, and Script-Implicit (SI), where inference is to be drawn based on the learner's world knowledge (Wixson, 1983). The second scale categorizes questions as verification, enumeration, comparison, identification, and possibly more. These scales turn the student's interest to the vocabulary and the communication goals on the one hand, and lexical relations on the other. *DidaLect*'s present prototype is built around multiple-choice questions. Our project aims to extend that to some types of free-text answers.

3. Assessment of free-text answers

In CALL systems, assessment of free-text answers is usually understood as a supervised classification problem. The learner's answer is compared to a set of reference answers, matched fully or partially based on syntactic and semantic similarity. We take advantage of the fact that valid answers to topical questions must be extracted from a reference text, which deals with a specific domain. We plan to perform assessment with limited supervision: the only reference answers to be considered are fragments of the text from which the question is created, and reformulations are not precoded as reference material, but left to evaluation through analysis.

The restriction on the forms of valid answers – and therefore on the reformulation possibilities – is guaranteed at the theoretical level by the pedagogical framework in which we conceive the questions. Only TE and TI questions may allow automatic assessment. A TE question comes from a single sentence of the text. We assume that a TI question arises from at most three sentences via an inference restricted to resolving co-reference, so we can still say that all that is needed for an answer is contained in the text. (SI questions require drawing conclusions from world knowledge, which in most cases is absent from the text.) Therefore, we claim that our method of assessing answers can rely entirely on symbolic linguistic analysis.

The didactic emphasis is on second-language acquisition rather than domain lexicons or lexical relations as in Graesser (2005), VanLehn (2002) or Vitanova (2004). This means that all admissible lexical relations are contained in the text fragment used to create the question. As these relations are not part of a domain lexicon, there is no need to precode the semantics of a domain in a knowledge base. A dictionary of synonyms is enough to assess the validity of lexical expressions being reformulated. (This is not the case in closed domains, where polysemy is rare.) Our goal is to consider lexical relations in a reference answer as manifested in surface-syntactic constituents. We do not deal with concepts in any general ontology, that is, we do not work with word meanings recorded in a knowledge base. In the end, lexical relations in a student answer are reformulated using synonyms of the words that participate in these relations.

With this assumption, assessing the correctness of the student's answer consists in verifying the correspondence of the lexical content with the reference answers, with respect to possible role shifting when the answer is reformulated syntactically, such as shifting the patient and agent around the predicate in passivization – an approach already defended in (Rosé *et al.*, 2003). Therefore, leaving aside the question of feedback, the procedure is close to that of Question-Answering techniques, comparing textual fragments for a match. Here, however, any divergence of segments must be either recorded as supplementary content or recognized as a given known type of error.

The resources required include a robust parsing module and a good dictionary of synonyms coupled with a derivational dictionary. In case of parser failure, a list of part-of-speech-tagged words could be a backup resource. We used the Xerox Incremental Parser, licensed to our lab (Ait-Mokhtar *et al.*, 2001; XIP, 2003). It produces several types of output, such as syntactic trees or lists of lexicalized roles and adjuncts, possibly in XML to facilitate post-processing. We also applied the 540,000 word dictionary of synonyms and the derivational dictionary included in *Dictionnaire Intégral* (2006).

4. Processing

Three word lists are produced from the student's input, based on a comparison between the reference (Ref) and the student's answer (S). L1 is a list of words present in S and absent in Ref. L2 contains words absent in S and present in Ref, L3 – words common to both. As the question to be answered is based on a single sentence, or a maximum of three adjacent sentences, reference material comes from a small segment of the text. The idea is to compare the two segments first for lexical and then syntactic similarities. The lists L1, L2, L3 trigger processing based on how they are populated. Typically, empty L1 and L2 with L3 containing all of words present in the segments would signal a correct student's answer, or at least correct words. Often, however, students tend to reformulate sentences, either to express their knowledge or as they feel that the purpose of the exercise is to show reformulation capability. That is why processing usually begins with both L1 and L2 non-empty.

Once these lists have been built, processing continues with the parsing of the sentences Ref and S. Figure 1 shows part of a XIP output for a real student answer, complete with errors. One of the advantages of XIP is its ability to proceed with limited or incomplete knowledge (hence « incremental parsing »), so it produces a parse even in the presence of errors in the input.

```
SUBJ(<approche^approcher:53>,<cardio-vasculaire^cardio-vasculaire:48>)
OBJ(<approche^approcher:53>,<personne^personne:56>)
VMOD_POSIT1(<approche^approcher:53>,<une^un:55>)
NMOD_POSIT1(<cardio-vasculaire^cardio-vasculaire:48>,<rat^rat:51>)
PREPOBJ(<une^un:55>,<à^à:54>)
PREPOBJ(<rat^rat:51>,<d'^de:49>)
DETERM(<rat^rat:51>,<un^un:50>)
```

Figure 1a. XIP output for an erroneous student's answer: « Le cardio-vasculaire d'un rat s'approche à une personne humain – note that the analysis is incomplete

Based on lists L1 and L2, four cases are possible, depending on the degree of reformulation that the student applied:

1. both L1 and L2 are non-empty: reformulation (figure 1);
2. only L1 is empty: S is a minimal expression of Ref (it contains less than what Ref expresses);
3. only L2 is empty: S builds over a maximal expression of Ref (it contains more than what Ref expresses);

4. both L1 and L2 are empty: no reformulation, S is a maximal expression of Ref (it contains no more and no less than what Ref expresses).

SUBJ(<est^être:2>,<rat^rat:1>)
OBJ_SPRED(<est^être:2>,<animal^animal:4>)
OBJ(<possède^posséder:6>,<système^système:8>)
COREF_REL(<animal^animal:4>,<qui^qui:5>)
NMOD_POSIT1(<système^système:8>,<cardio-vasculaire^cardio-vasculaire:9>)
NMOD_POSIT1(<système^système:8>,<semblable^semblable:11>)
NMOD_POSIT1(<celui^celui:13>,<humain^humain:16>)
ADJMOD(<semblable^semblable:11>,<celui^celui:13>)
PREPOBJ(<humain^humain:16>,<de^de:14>)
PREPOBJ(<humain^humain:16>,<de^de:14>)
PREPOBJ(<celui^celui:13>,<à^à:12>)
DETERM(<système^système:8>,<un^un:7>)
DETERM(<animal^animal:4>,<un^un:3>)
DETERM_DEF(<rat^rat:1>,<le^le:0>)
CONNECT_REL(<possède^posséder:6>,<qui^qui:5>)

Figure 1b. XIP output for a reference sentence: « Le rat est un animal qui possède un système cardio-vasculaire très semblable à celui de l'humain »

Naturally, the above is only valid in what concerns lexical reformulation, with the exception of 4: perfect correspondence of word sets should signal the identity of syntactic structures, though errors in S could occur in agreement and the use of function words. This, however, will trigger different procedures in order to assess the general correctness of S with respect to Ref and general linguistic principles. Cases 2 and 3 call for similar processing, though they are opposite in terms of content sets. In Case 2, the words present in S are a subset of the words in Ref, and conversely in Case 3. We examine them all in turn, in the following combinations:

- Case 1: Synonymy Detection + Building Tree + Syntactic Assessment,
- Cases 2 and 3: Building Tree + Syntactic Assessment,
- Case 4: Syntactic Assessment.

Pre-processing

Prior to parsing, an answer is pre-processed. This involves sentence-level anaphora resolution, part-of-speech (PoS) tagging of the words, and verb tagging. Perez *et al*. (2005) has shown that co-reference resolution is actually a crucial part in processing of language for comparison between a reference fragment and a candidate fragment. For now, anaphora resolution in our system is manual: questions are encoded together with the reference passage in which all co-reference is resolved. While this process is not trivial to automate, we consider it less important in the phase of prototyping other elements of the procedure.

PoS and verb tagging consist in tagging words with (1) PoS information to counter and correct possible parser failures due to tagging errors – a separate list from that of the parser is used in the process – and (2) verb information, namely categories assigned to verbs, which are tagged as action, state (« être ») or attribute (« avoir ») verbs.

Parsing with XIP produces a tree, a set of dependency relations (such as those shown in figure 1) and an XML file containing gender and number information at the word level. After parsing, XIP output is decomposed in order to catch agreement values and therefore check agreement correctness provided via an independent XML file that records gender and number values for each word. The XML format of the file makes this operation easy to implement, and processing agreement can take place independent of (before or after) other steps.

Synonymy Detection

In Case 1, which represents the majority of situations, the procedure begins with searching for synonymy relations between words in L1 and L2. Words from L1 are derived using the derivational dictionary in all categories of words appearing in L2. This is done to detect possible synonymy relations across PoS categories, such as synonymy between the verb « s'approcher de » from S and the adjective « semblable » from Ref in the example in figure 1. An advantage of the synonymy module of *Dictionnaire Intégral* is that, for a given meaning, it shows prepositions which follow verbs. The synonyms, as detected, are recorded and used as anchors to put S and Ref in parallel in the subsequent process of tree building. Any word from L1 that has not been recorded as a synonym is supplementary and therefore signals information not contained in Ref, while any such word from L2 signals information missing from S with respect to Ref.

Tree Building

This step varies depending on the cases. It employs the XIP relations such as those shown in figure 1. In Cases 2 and 3, it amounts to first putting in correspondence the maximal and minimal phrases expressed in S and Ref, starting from the word(s) contained in the non-empty list (L2 in Case 2, L1 in Case 3), up to a match with a word contained in both sentences. Then, complete trees for the whole S and Ref are built, starting from the XIP relations in which the previously detected matches appear. This step enables us to record the missing (Case 2) or supplementary (Case 3) lexemes from S under their syntactic dependency.

In Case 1, the starting points are the synonyms. If several synonyms have been detected, the synonym that appears in most dependency relations is selected to initiate tree building. Recursion in the process takes place on the most promising words. Those words are identical in form and sense. They are, however, far enough in the two different syntactic structures to maximize the possibility of discovering the location of different lexical content in the process of agglomerating the dependency relations that associates these words.

This tree-building operation works on the two structures S and Ref, of which only Ref is known to be correct. The process will naturally halt once all relations inclusive of the synonyms have been extracted and incorporated together with modifiers (_MOD in figure 1). It results in agglomerated relations (<S>) and (<Ref>), such as the temporary relations in figure 2 built from the analysis in figure 1. After agglomerating all relations in which synonyms appear, further lexemes are needed to continue tree building. These come from the list L3 as the words contained in (<S>) or (<Ref>), but not in both, (« rat » and « humain » in the example). This makes the process of tree-building parallel in that (<S>) and (<Ref>) builds over lexical elements of each other.

Eventually, supplementary words from L1 and L2 will appear in the complete trees as modifiers or heads in relations incorporated based on common words or synonyms. We can therefore detect supplementary or missing material within the frame of dependency relations in the agglomerated relations. Controlling the semantic content of S with respect to Ref amounts to comparing their respective constituents, as words, and their bindings. At the same time, role positions occupied by those words may shift when S is a reformulation of Ref. Comparing the constituents amounts to a search for similarity between the two sentences, keeping a record of all variations. Hence the idea of starting with the synonyms, and continuing with common words kept in different syntactic positions due to syntactic reformulation. Besides, we know from experience that reformulated words are usually semantically the most promising: they make the core of a sentence's postulate. Two trees correspond if they contain the same lexical elements, perhaps as synonyms, within equal or equivalent role-predicate structures. There are four categories of such corresponding structures, for constituents centred on a verb, noun, attributive adjective or predicative adjective.

(S) SUBJ(<OBJ(<approcher>,<personne>)>,<NMOD(<cardio-vasculaire>,<rat>)>)
(Ref) OBJ(<posséder>,< NMOD(<NMOD(<système>,<cardio-vasculaire>)>,
<ADJMOD(<semblable>,<NMOD(<*système*>,<humain>)>)>)>)

Figure 2. Agglomerated relations based on synonyms and modifiers. A solved co-reference appears in italics

Building the trees based on lexical commonality between the two sentences (identity or synonymy) also enables error detection. Stopping the recursive process of building trees signals either the student's mistake or a parsing error. In our example, « humain » does not appear in the XIP analysis of S: « personne » has been encoded in the lexicon as a pronoun. This is a parser error. We have devised a number of heuristics to tackle the issue, shown in the Heuristics subsection.

Syntactic Assessment

This step verifies the correctness of structure S once tree building has been performed for Cases 1-3. The step rests on a set of correct structures as well as typical ill-formed structures. These ill-formed structures come from a collection of student dissertations as hand-written answers to Script-Implicit questions as well as from literature on second-language learning. Questions of this kind, not assessed automatically, call for longer answers containing multiple sentences and thus are a rich source of error material, though our set of such errors is far from complete. Correct structures are understood both in terms of form and equivalence (rules of syntactic reformulation), so they serve a double purpose of verifying the correctness of a syntactic structure and controlling lexical content and lexicon position between two corresponding structures (that of Ref and S). Simplifying, an active sentence structure of the form S[VP[NP1[V]]NP2] is to be read both as a correct structure in itself and as a reformulation of a passive sentence structure of the form S[VP[NP2[V]]PP[P[NP1]]].

The two sets of rules (for wrong and correct structures) have different forms. Intermediate FSL students' typical mistakes tend to be lexical: wrong articles, prepositions and conjugation. Therefore, ill-formed structures work for phrases rather than sentences. They record typical lexical mistakes together with their correction. Correct structures come from various answers. We currently have 29 questions, each with four syntactically different forms of answers. This material might be helpful if chunked into phrases, but insufficient if each sentence is treated as a fixed pattern.

Heuristics

As often in NLP systems, a number of heuristics are used to solve problems arising from errors. Increasing the number of heuristics is an on-going effort. The various heuristics cover the issues that arise from errors introduced by the parser, errors introduced by the student or reformulations at the syntactic and semantico-syntactic level.

Errors introduced by the parser can be corrected when they have lexical reasons (such as « personne » encoded as a pronoun). Therefore, words from Ref and S are pre-processed using a parallel list for PoS tagging with all possible forms for a noun. If an error occurs, the lexicon is automatically augmented with the missing information (XIP offers this possibility), and sentence is sent back for parsing.

Student errors at the lexical level are tackled through heuristics when parsing is impossible. Students tend to mix word forms and use forms inappropriate enough to prevent parsing. This can be addressed partially. We can detect the point of the parsing failure and either send information to the student or derive the word's alternative forms, reshape the sentence for proper parsing and continue processing, keeping a record of the error. The last approach has been tested and is successful in cases where an erroneous word stands as a wrong reformulation of a word present in Ref (a student mistook present participle *semblant* for adjective *semblable*, which halted parsing).

Heuristics for the assessment of syntactic structures actually call for the rules briefly presented for Syntactic Assessment, and are part of that processing step. Verifying semantico-syntactic reformulations is another matter. This corresponds to semantic variation out of the reach of synonymy analysis. Due to the simplicity of the exercises and the reference text, however, we can address this issue at least partially. We have observed it empirically at the verbal expression level (in our example in figure 1, attribution expressed by the verb « posséder » is an instance of this problem: « le rat possède un système cardio-vasculaire » and « le système cardio-vasculaire du rat » have equal senses but different syntactic values). In this respect, lists of current verbs encoded as action, attribute or state verbs have been produced and put to use to facilitate detection of verbs which act in sentence as discourse connector instead of sense conveyer. This is treated as a position problem under the reformulation rules, and a constituent phrase containing one such verb is to be put in relation of equivalence with other forms equal in sense. The latter rules are part of the set of rules used for Syntactic Assessment, so this aspect of verbal expressions is understood as « semantico-syntactic ».

Assessment

Properly speaking, assessment consists of keeping record of all errors in agreement, lexicon, syntax and content. These categories of assessment are of different pedagogical importance and are not treated as equal. Typically, agreement and syntactic judgements have an absolute value, either right or wrong. Lexicon errors are evaluated with respect to a typology of lexical errors, making distinctions ranging from mistake to error, usually from orthography to misconceptions (as having mistaken *semblant* for *semblable*). Finally, the content of S, as minimum in relation to the content of Ref, may or may not yield an error depending on the nature and function of the missing part(s). In our example, having mistaken « cardio-vasculaire » for a noun is both a lexical and syntactic error. Having taken « personne humain » for « celui (*système cardio-vasculaire*) de l'humain » in the enunciation of comparison is a content error leading to a semantic error, not mentioning agreement. Having left out superlative « très » or subordinate « le rat est un animal qui » are not errors.

Error qualification with respect to minimal validity lies in the distinction between phrase modifiers and sentence modifiers (complements or subordinates). Phrase modifiers are evaluated under their PoS categories, which are considered as necessary or superfluous. For example, an adverb is considered less important than an adjective or a noun. Subordinates are evaluated with respect to the nature of the contained verb. Within a sentence, a subordinate with a state or attribute verb is considered less semantically (or informatively) important than one containing an action verb. Sentence complements, such as temporal complements, are considered as superfluous, at least in the present state of our research.

Feedback can vary in size and content depending on the correctness of the student's answer. Several points are salient.

- General validation over Ref to assess if the student's sentence answers the question.
- Syntactic correctness. In case of error, the system should show the correct structure. If, however, most syntactic errors are due to lexical mistakes, the system should show correct lexical forms, explaining *why* an error occurred. Reasons include PoS mistagging, wrong preposition selection or erroneous verb tense.
- Lexicon. Given a typology of lexical errors, feedback gives the category of error plus a correction when possible. This implies, not q uite realistically, that the system *knows* what the student intended to say.
- Range of semantic content with respect to Ref (see the beginning of this section, minimal and maximal coverage).

Support

A fair number of CAA systems give good results, perhaps better than ours, but those systems work under tight knowledge control. For a question, a number of answers have to be encoded, either in the form of referential isolated sentences or as expressed in an annotated corpus. In both cases, the encoding task is substantial. The advantage of our system is that no encoding of reference material is necessary. The question arises as to whether answer fragments in the text should be included in the formulation of a question, or left to a tracing module, in charge of automatically detecting proper segment. This question has been left aside for now, as implementing a tracing module is of little difficulty, especially due to the nature and size of the texts. We nevertheless posed this question as a possibility of providing the means of detecting wrong text segment selection in the student's effort to build an answer (knowing what the student is talking about).

Therefore, all that is needed to feed the system with didactic material is a set of texts and questions, and a slight degree of expertise on the admissible types, TE and TI, when creating the questions.

5. Future Work and Conclusions

We have presented the design of a CAA module for the unsupervised validation of free-text answers to open questions.[2] This module is rooted in the theoretical framework deployed in *DidaLect*, a tool for the improvement of text comprehension. It is an autonomous tool, though it functions as a self-learning system. The present resources are directed to the acquisition of French, but the core of the system is portable from one European language to another if backed with adequate resources. The module is being implemented. The design centred

[2] This is the topic of the first author's PhD thesis in Computer Science, to be completed in early 2007.

around the proper assessment of a small set of 12 TE and TI answers, specifically gathered for this purpose from intermediate-level FSL students.

Future work includes the enlargement of the answer corpus for testing and validation purposes, as well as the inclusion of a word frequency list in order to support the students with difficult and rare words, and a co-reference resolution module. Our main objective for the coming months remains the implementation of the system based on solid software engineering principles.

References

AIT-MOKHTAR S., CHANOD J.-P., ROUX C. (2001). "A Multi-Input Dependency Parser". In *Proceedings of the Seventh IWPT (International Workshop on Parsing Technologies)*. Beijing.

ANCTIL D. (2005). "Characterization of the Notion of Lexical Error in the Framework of Explanatory Combinatorial Lexicology". OLST-RALI, Université de Montréal.

BALCOM P., COPECK T., SZPAKOWICZ S. (2006). "DidaLect: Conception, implantation et évaluation initiale". In *Technologies langagières et apprentissage des langues. Actes du congrès de l'ACFAS* 105. Montréal (to appear).

CHEN L., TOKUDA N. (1999). "A New Diagnostic System for J-E translation ILTS". In *Proceedings Machine Translation Summit*: 608-616.

G. DENHIÈRE G., BAUDET S. (1992). *Lecture, Compréhension de Texte et Sciences Cognitives*. PUF, Paris.

DESROCHERS A., DUQUETTE L., SZPAKOWICZ S. (2004). "Adaptive Courseware for Reading Comprehension in French as a Second Language: The Challenges of Multidisciplinarity in CALL". In *Proceedings of 11th International CALL Conference*. Antwerp: 85-91.

DESROCHERS A., L. DUQUETTE L. (2006). "Appuyer la Compréhension en Lecture à l'Aide d'un Logiciel Adaptatif". In *Technologies langagières et apprentissage des langues: Actes du congrès de l'ACFAS* 105. Montréal (to appear).

Dictionnaire Intégral (2006). http://www.memodata.com.

GRAESSER A.C., PERSON N., LU Z., GEON M.J., B. MCDANIEL B. (2005). "Learning while Holding a Conversation with a Computer".In *Technology-Based Education: Bringing Researchers and Practitioners Together*. Information Age Publishing: 143-167.

MICHAUD L.N., MCCOY K.F., STARK L.A. (2001) "Modeling the Acquisition of English: an Intelligent CALL Approach". In *Proc 8th International Conference on User Modeling*. *Lecture Notes in Artificial Intelligence* 2109. Springer: 14-23.

PEREZ D. (2004). *Automatic Evaluation of User's Short Essays by Using Statistical and Shallow Natural Language Processing Techniques*. Advanced Studies Diploma, EPS, Universidad Autonoma de Madrid.

PEREZ D., POSTOLACHE O., ALFONSECA E., CRISTEA D., RODRIGUEZ P. (2005). "About the Effects of Using Anaphora Resolution in Assessing Free-Text Student Answers". In *Proceedings of RANLP-2005*: 380-386.

RICHGELS D., MCGEE L., LEMAX R., SHEARD C. (1987). "Awareness of Four Text Structures: Effects on Recall of Expository Texts". In *Reading Research Quarterly* XXII: 177-197.

ROSÉ C.P., GAYDOS A., HALL B.S., ROQUE A., VANLEHN K. (2003). "Overcoming the Knowledge Engineering Bottleneck for Understanding Student Language Input". In *Proceedings of AI in Education*.

SCORM/ADL (2006). http://www.adlnet.org.

VANLEHN K., JORDAN P., ROSÉ C.P. (2002). "The Architecture of why2-atlas: a Coach for Qualitative Physics Essay Writing". In *Proceedings of Intelligent Tutoring Systems Conference*.

VITANOVA I. (2004). "Evaluating Integrated NLP in Foreign Language Learning: Technology Meets Pedagogy". In *Proceedings of InSTIL/ICALL symposium*.

WIXSON K.K. (1983). "Postreading Question-Answer Interactions and Children's Learning from Texts". In *Journal of Education Psychology* 5: 413-423.

XIP Parser (2003). *Xerox Research Center Europe Technical Document*.

Capitalisation d'une ressource en or : le dictionnaire

Michael Zock

Laboratoire d'Informatique Fondamentale (LIF) – CNRS
michael.zock@lif.univ-mrs.fr

Résumé

Notre objectif est de montrer comment l'ajout de certaines fonctionnalités à un dictionnaire électronique existant pourrait aider des êtres humains à apprendre ou à traiter la langue. Ces extensions concernent l'accès aux mots, leur mémorisation et l'acquisition d'automatismes pour produire des structures syntaxiques fondamentales d'une langue. Pour atteindre ce dernier objectif, nous proposons un générateur de phrases paramétrable, basé sur la notion de formulaires : l'utilisateur indique son *intention de communication* et les *mots* à utiliser (apprendre), et le programme construit les phrases correspondantes. Quant à l'aide à l'accès lexical, nous proposons d'ajouter, à un dictionnaire électronique existant, un index basé sur la notion d'association. L'index est construit à partir des mots co-occurrents d'un corpus qui, lui est supposé représenter les connaissances du monde de l'homme de la rue. La recherche lexicale se fait par navigation : partant d'un mot ou d'une idée, on s'approche progressivement du candidat lexical recherché.

Mots-clés : extensions aux dictionnaires électroniques, générateur d'exercices linguistiques, mémorisation de mots, acquisition de réflexes linguistiques, outils d'aide à la navigation dans un espace lexico-conceptuel, index basé sur la notion d'association.

Abstract

The goal of this paper is to explore extensions to electronic dictionaries. Adding certain functions could considerably extend the range of tasks for which they could provide support. Putting the needed information at the distance of a mouse click would allow for *active* reading. This would require tight coupling of the dictionary with a text editor: all the information in the dictionary should be accessible via a mouseclick. Dictionaries combined with a flashcard system and an exercise generator could support the *memorization* and *automation* of words and syntactic structures. Finally, structuring the dictionary in a way akin to the human mind (associative network) could help the writer to find new ideas, and if needed, the word he is looking for. In sum, rather than considering the dictionary just as another component of the process of language production or comprehension chain, we consider it as the single most important resource, provided that one knows how to use it.

Keywords: extensions to electronic dictionaries, exerciceexercise generator, memorisation of words, acquisition of linguistic skills (habits), tools for assisting navigation in a conceptual-lexical network, association-based index.

1. La problématique

L'objectif de ce papier est triple : (a) montrer comment au prix de quelques extensions à un dictionnaire électronique existant, on pourrait aider des êtres humains à apprendre ou à traiter la langue ; (b) convaincre les chercheurs du monde du TAL qu'il vaut mieux construire des petits modules, mais ouverts que des systèmes complets et fermés ; (c) informer les étudiants en langues de l'existence de certains outils disponibles sur le web pour envisager l'apprentissage différemment.

Si les méthodes de langue ont sûrement des qualités, elles ont également toutes au moins un défaut : elles sont fermées. Or, les besoins des apprenants sont variables, souvent

imprévisibles et de toute façon sujets à des changements. Aucune méthode ne pourrait jamais convenir à tout le monde à tout moment.

Il y a maintes manières d'apprendre une langue, celle de l'école est artificielle qu'on le veuille ou non. Pourtant, il y a d'autres manières d'apprendre bien plus naturelles, en l'occurrence l'apprentissage incident. En effet, c'est beaucoup plus naturel de lire un journal ou d'écouter des nouvelles que de faire des exercices comme on vous le demande à l'école. Pourtant, en se livrant à ce type d'activités extra scolaires on apprend également, mais le résultat est un effet de bord et non pas l'objectif premier[1]. Enfin, si on apprend pour la (sur)vie, c'est qu'on apprend également pendant toute la vie, surtout à notre époque où tout change si rapidement. C'est ici que le TAL pourrait rendre d'énormes services, en assistant la formation continue (ou l'auto-formation) et l'apprentissage incident.

Dans ce qui suit nous allons montrer comment on pourrait assister le lecteur en intégrant les dictionnaires à un éditeur de texte, si bien qu'en cliquant sur un mot on verrait apparaître les informations contenues dans les différentes rubriques (traduction, informations grammaticales), etc. Pour aider l'étudiant à acquérir une certaine aisance verbale (fluidité) nous présenterons deux outils dont les buts respectifs sont la *mémorisation* de mots et l'acquisition d'*automatismes* concernant les structures fondamentales d'une langue. Enfin, pour aider les rédacteurs à trouver les mots cherchés, nous présentons une méthode pour les mettre sur la bonne voie. En particulier, nous proposons d'ajouter à un dictionnaire électronique existant un index basé sur la notion d'association. L'index étant construit à partir de mots co-occurrents issus d'un corpus représentatif des connaissances du monde du citoyen moyen. Un dictionnaire enrichi de ce type d'information permettrait donc d'initier la recherche avec des mots quelconques pour s'approcher progressivement du candidat idéal.

2. FonctionalitésFonctionnalités assistant le *récepteur :* mettre l'information recherchée au bout des doigts

Étant habitués à des langues comme l'anglais, l'allemand ou l'espagnol, nous avons tendance à oublier qu'il y a des langues dont l'écriture est très différente. Outre le problème de la morphologie (les entrées des dictionnaires sont généralement des lemmes et non pas la forme fléchie), il y aura donc en plus celui du déchiffrage de caractères. Supposons qu'on veuille apprendre une langue dont l'écriture ne se fait pas en termes de caractères latins (grec, russe, coréen, japonais, chinois). Comment consulter le dictionnaire dans ce cas là? là ? C'est pratiquement impossible, pourtant il existe une solution relativement simple et pour la plupart des cas satisfaisante.

Prenons comme exemple un texte en japonais でんわほどこですか. L'apprenant est confronté ici à au moins deux problèmes : celui d'identification des frontières lexicales (les mots ne sont pas forcément séparés par un blanc), et celui de la conversion des symboles (kana) en phonèmes/graphèmes (で → de)[2]. Ces deux informations sont bien entendu capitales. Ainsi, on aimerait voir apparaître sur écran la prononciation correspondante à la chaîne de caractères でんわ, à savoir, denwa. Même si cela ne nous dit pas pour autant que ce mot signifie 'téléphone', on aurait pu l'apprendre ensuite, car, grâce à la translittération, on pourrait

[1] Bien entendu, il ne s'agit pas de bannir les cours de langue. Le point ici est plutôt de montrer qu'on peut construire, à coup raisonnable, des béquilles fort utiles pour l'utilisateur de la langue, qu'il soit déjà expert ou encore apprenant.

[2] La situation est encore un peu différente dans une langue comme le chinois qui utilise des idéogrammes.

consulter un dictionnaire (à condition que les entrées soient en lettres romanes), ou demander à une personne sachant parler cette langue.

Trouver des informations dans un dictionnaire n'est donc pas toujours chose aisée, d'abord cela demande souvent beaucoup de temps, et il peut y avoir des problèmes liés à l'écriture et à la morphologie. Pourtant, les problèmes mentionnés pourraient trouver une solution simple et satisfaisante pour bon nombre de cas. Il suffirait d'ajouter au dictionnaire un translittérateur[3] et un lemmatiseur[4]. En intégrant ce type de fonctionnalité dans des applications courantes (traitement de texte, butineur), on pourrait désormais lire et comprendre des textes écrits dans une langue étrangère (*lecture active*). Il suffirait alors de cliquer sur un mot pour voir apparaître un menu "pop up" permettant au lecteur de choisir parmi les informations celle qui l'intéresse à cet instant (traduction, définition, information grammaticale, etc.). Par exemple, on pourrait imaginer l'interface suivante (cf. figure 1) : la traduction concerne ici le mot « shite », forme conjuguée de l'entrée lexicale « suru » (verbe à l'infinitif)

À noter qu'on commence à voir des programmes capables de faire cela, comme par exemple, GLOSSER (Nerbonne et Smit, 1996), un prototype développé au GETA par Mathieu Lafourcade, et les produits de la société *Transparent Language*. Hélas, tous ces systèmes sont fermés, et il n'y actuellement aucun moyen d'y ajouter d'autres fonctionnalités. Enfin, les urls suivantes méritent considération, surtout pour ceux souhaitant apprendre à lire en chinois ou en japonais : (http://language.tiu.ac.jp/tools_e.html ; http://www.rikai.com/perl/HomePage.pl?Language=Ja; http://www.popjisyo.com/ ; http://www.popjisyo.com/WebHint/Portal_e.aspx ; http://www.newsinchinese.com/. La même remarque vaut pour les sites suivants : http://www.animelab.com/anime.manga/translate ; http://www.animelab.com/anime.manga/dictionary/ ; http://www.eloquentsw.com/livedictionary.html, mais cette fois-ci pour des problèmes liés aux dictionnaires.

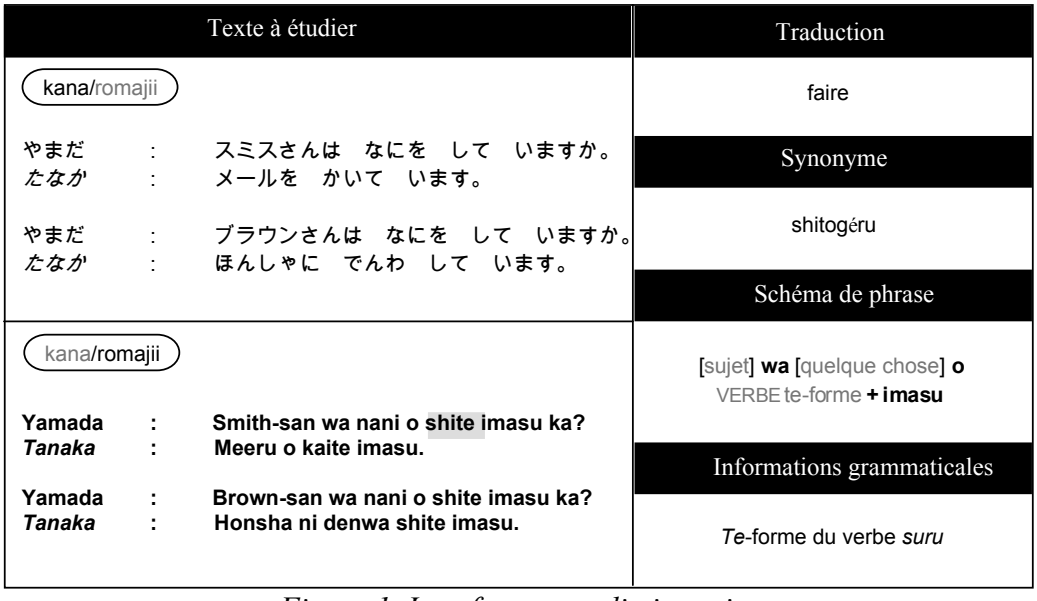


Figure 1. Interface texte-dictionnaire

[3] La translittération est l'opération consistant à transcrire les graphèmes d'un alphabet ou d'un syllabaire (comme le japonais) dans les graphèmes d'un autre système d'écriture (généralement un alphabet), de telle sorte qu'à un même graphème (ou suite de *graphèmes*) de la langue de départ corresponde toujours un même graphème (ou suite de graphèmes) du système d'écriture d'arrivée, et ce indépendamment de la prononciation (http://fr.wikipedia.org/wiki/Translitération).

[4] Un lemmatiseur est un programme qui permet de passer d'un mot portant des marques de flexion (pluriel, forme conjuguée d'un verbe...) à sa forme de référence (entrée lexicale, lemme) ou inversement.

3. Fonctionnalités assistant les *producteurs* de langue

Les besoins sont bien entendu différents selon que l''on est expert ou étudiant, en train d'apprendre une langue. Commençons par ce dernier.

3.1. Assister l'étudiant de langues

Pour pouvoir s'exprimer dans une langue, il faut avoir non seulement énormément de connaissances, mais également posséder un savoir-faire non négligeable. Ainsi, un locuteur doit-il pouvoir trouver le mot exprimant sa pensée[5], l'insérer au bon endroit de la phrase, l'adapter morphologiquement, tout en continuant à planifier l'idée suivante, et tout ceci en un très court laps de temps. Si jamais une de ces étapes tarde ou échoue, on assiste à des lapsus, bafouillages, sons de remplissage, ou, des pauses plus ou moins prononcées, pouvant aller jusqu'au silence total. L'apprentissage du *vocabulaire* (mémorisation) et des *structures syntaxiques*, tout comme l'*acquisition* d'*automatismes* est donc indispensable pour pouvoir produire des phrases à un débit *normal*.

*3.1.1. Apprentissage de vocabulaire : **mémorisation** de mots*

Dire que l'apprentissage de mots est fondamental est trivial, ce qui l'est moins c'est de dire comment, car, même appris, les mots semblent avoir la fâcheuse tendance de vouloir s'échapper de notre mémoire.

Se basant sur les travaux des psychologues étudiant la mémoire, Leitner (1972) a proposé une application intéressante. Le principe est simple. L'auteur propose de ranger dans une boîte à cinq compartiments des cartes contenant d'un côté la question (par exemple, un *mot à traduire*) et de l'autre la réponse (*mot traduit*). Toutes les cartes se trouvent dans le premier compartiment au début de l'apprentissage, pour passer successivement au compartiment suivant (ou précédent), tout dépend de la qualité de la réponse (bonne/mauvaise). La leçon est considérée acquise lorsque toutes les cartes sont dans le dernier compartiment. L'idée sous-jacente à cette méthode est simple : consacrer un maximum de temps aux éléments récalcitrants. L'idée ne date pas d'hier, en fait elle est connue sous le nom de la *loi de Jost*, selon laquelle un apprentissage fractionné et espacé dans le temps est plus efficace qu'un apprentissage regroupé (Kekenbosch, 1991 : 7-13).

Bien sûr une version informatique n'a de sens que si elle apporte quelque chose par rapport à la version papier. Or, c'est clairement le cas. La souplesse et la puissance informatique sont au rendez-vous. Les données sont faciles à acquérir, à échanger et à mettre à jour. Par exemple, on peut imaginer la construction de ponts mnésiques (associations) entre la question et la réponse. L'ergonomie, l'ouverture, la gestion des performances (décompte automatique) sont toutes des facteurs apportant un confort indéniable à l'utilisateur. Quant aux paramètres de présentation des données (couleur, taille, nombre de présentations, vitesse de défilement, etc.) il y a de très nombreuses possibilités pour adapter l'outil à son goût et ses besoins. Enfin, on peut même imaginer la mise au point de stratégies pour tester leurs avantages respectifs : *précision* vs *rapidité*.

[5] Ce qui veut dire, qu'il doit chercher dans un stock énorme (les chiffres avancés varient selon les auteurs entre 30 à 60 000 mots.) un élément particulier. La performance est impressionnante, équivalent à la consultation d'un dictionnaire comme *Le Grand Robert* trois fois par seconde, et ceci pendant plusieurs heures.

*3.1.2. **Mémorisation** et **automatisation des** structures fondamentales **d'une langue***

Posséder un grand vocabulaire (même actif) ne signifie pas encore savoir produire des phrases. Pour cela il faut savoir (et savoir-faire) bien plus de choses. Il y a différentes manières de produire une phrase : (a) on peut la produire pièce par pièce (mot par mot) en recourant à une grammaire formelle ;; (b) on peut faire appel à des fragments de taille variable, allant d'expressions toutes faites jusqu'à la phrase ;; (c) on peut utiliser des schémas de phrase que l''on remplit ensuite avec des données lexicales.

Cette dernière solution est une sorte d'heureux compromis entre la génération incrémentale et celle consistant à réutiliser des pièces toutes faites. La première étant basée sur une grammaire, donc souple et puissante, mais lente et assez complexe, tandis que la seconde est rigide, mais rapide et extrêmement simple. Ce n'est donc pas étonnant de constater que la méthode représentant le meilleur compromis soit celle retenue dans l'apprentissage naturel d'une langue et dans l'enseignement des langues.

Nous allons esquisser ici comment, partant d'un dictionnaire, on peut construire ce type de générateur de phrase. En fait, comme nous allons voir, ce n'est pas seulement un générateur de phrases, mais un générateur d'exercices. L'idée est toute simple. On indexe l'ensemble des structures à apprendre en termes de buts, puis on demande à l'utilisateur d'en choisir un et de remplir les variables de la structure correspondant au but avec des données lexicales. Ceci étant fait, le système a tout ce qu'il faut pour construire des phrases[6] (Zock et Quint, 2003).

Prenons un exemple. Supposons qu'on veuille exprimer le but suivant : *définition (animal)*. Pour cela il y a plusieurs schémas en français, par exemple : (a) «un **X** est un **Y** qui **ACTION**» (b) «un **X** est une espèce de **Y** vivant en **Z**», où **X**, **Y**, **Z** et **ACTION** sont des variables (**X** : nom d'animal; **Y** : hyperonyme ; **Y** : lieu) pour lesquelles il faudrait préciser la valeur lexicale. Ceci étant fait, le système pourrait alors produire des phrases du type:type : (a) un **perroquet** est un **oiseau** qui **parle**, ou (b) un **koala** est une espèce de **marsupial** vivant en **Australie**.

Prenons un autre exemple. Supposons qu'on veuille apprendre en japonais l'équivalent du français « *où est* **x** ? », but ou intention qu'on pourrait communiquer soit en choisissant dans un menu, soit en ayant recours à un langage de requête : `lieu (x), x` étant l'objet pour lequel on demande la localisation.

Connaissant désormais l'*intention* de communication, le système présenterait alors la ou les structures correspondantes, en l'occurrence « x-*wa doko desu ka* », attendant qu'on lui précise la ou les valeurs de « x ». Supposons qu'elles soient « *arrêt de taxi*, *banque* et *hôpital* ». Le système aura donc désormais tout ce qu'il faut pour produire les phrases contenant ces éléments, mais, comme il s'agit d'un exercice dont le but est d'aider l'étudiant à produire ces phrases, on incite ce dernier à essayer d'abord lui-même, avant que le système ne produise sa version. Ainsi, le système présente une amorce, attendant que l'élève l'insère au bon endroit de la structure correspondant à son intention. C'est en effectuant les opérations requises tout en comparant les résultats (la sienne et celle du système) que l'étudiant apprend.

Utilisateur *:*	*Lieu x ?*	(intention de communication)
Système :	x wa doko desu ka?	(structure correspondante)
	x = ?	(demande de précision de valeur)
Utilisateur :	x : *arrêt de taxi*, banque, hôpital	(valeurs lexicales)

[6] Bien sûr, ce type de générateur est d'autant plus limité que le système n'a pas de composant morphologique. Certes, on pourrait l'inclure, mais le point ici était justement de montrer qu'on pourrait créer un générateur de phrase (ou d'exercice) avec un minimum d'informations. Celles-ci se trouvent pratiquement toutes dans le dictionnaire. Il n'y a que les buts qui n'y figurent pas.

Système :	*Takushii-noriba* wa doko desu ka?	(insertion dans la structure de la phrase)
Système :	*banque*	(amorce)
Utilisateur :	*Ginko* wa doko desu ka?	(réponse étudiante)
Système :	Ginko wa doko desu ka?	(confirmation système)
Système :	hôpital	(amorce suivante)
		etc.

3.1.3. Discussion

Le type d'exercice que nous venons de proposer existe depuis fort longtemps. Ce sont les fameux pattern-drills (*exercices structuraux* en français), très en vogue pendant les années 50 et 60, à l'époque où les laboratoires de langue et les méthodes inspirées des idées behavioristes (notamment les méthodes audio-orales) avaient le vent en poupe. Pourtant, cette technique avait également ses détracteurs.

Si le behaviorisme de Skinner (1968) a néanmoins tant inspiré le monde éducatif, malgré les très nombreuses critiques de la part de *linguistes* (Chomsky, 1959), de *didacticiens* (Besse 1975) et de *psychologues* (Chastain, 1969; Le Rouzo, 1975), c'est qu'on trouve à sa base deux principes fondamentaux de l'apprentissage : celui de la ***rétroaction***, information concernant la qualité d'une réaction (réponse) à un problème (stimulus) et celui de la ***répétition***[7]. S'ajoute à cela un troisième élément, celui de la ***structure***, nommée jadis Gestalt, ou, selon les périodes, patron (pattern), schéma, cadre, frame. Ce que l''on cherchait à capter par ces termes, c'étaitest l'idée, que derrière une masse de formes variables il y a un invariant, la structure sous-jacente. Vu la généralité et la complémentarité de ces principes, il n'est donc pas étonnant de les trouver à la base de certaines théories comme le structuralisme ou l'apprentissage, ou encore à la base de certaines pratiques didactiques comme l'enseignement programmé (Pocztar, 1971) ou les *exercices structuraux*.

Certes, cette forme d'exercices n'est pas une panacée, mais utilisée à bon escient elle peut s'avérer utile, ne serait-ce que pour mémoriser et automatiser les mots dans le contexte de la phrase. Ainsi faisant, elle libère l'apprenant des aspects élémentaires de la langue (éléments mécaniques et de bas niveau) pour lui permettre d'accéder aux niveaux supérieurs, ceux des idées (planification de messages). Aussi, qu'on le veuille ou non, la répétition des formes (structures) est le prix à payer pour acquérir la maîtrise d'une activité aussi complexe que la production du langage.

Cependant, comme tous les praticiens le savent, les exercices structuraux souffrent de certaines faiblesses évidentes. Ils sont rigides et ils engendrent rapidement une certaine lassitude, ce qui est partiellement lié au caractère fermé des supports (livre, magnéto). Tout doit être prévu, et rien ne peut être changé après impression et/ou enregistrement. Or, ceci a complètement changé avec l'arrivée des ordinateurs. Désormais on peut changer les données à tout moment, pour les adapter en fonction des besoins du « client ». Or, ceux-ci varient non seulement d'une personne à l'autre, mais aussi intra-individuellement. Nos besoins changent à tout moment, d'où l'intérêt de construire des outils ouverts, adaptables en fonction des besoins du moment.

Enfin, il y a d'autres manières d'apprendre ces structures. Nous avons montré ici une façon parmi d'autres pour construire sa base. La méthode est interactive et fait appel à un générateur

[7] Que l'apprentissage (mémorisation), l'adresse (habileté) et la perfection demandent de l'exercice est connu depuis fort longtemps (practice makes perfect), si bien que les débuts de la psychologie expérimentale coïncident pratiquement avec leur étude. En effet, Ebbinghaus a étudié dès 1885 le rôle de la répétition (nature, espacement, etc.) dans l'apprentissage. Pour une revue de la question, voir Hilgard et Bower (1975).

(rudimentaire, certes). Cependant, on pourrait également constituer ce type de base en fouillant un corpus comme le web. Bien sûr, un débutant ne connaît pas forcément les schémas réalisant une intention de communication, mais le système les connaît (à condition de le lui avoir appris). Ayant fourni au système votre intention de communication, celui-ci peut vous indiquer le ou les schémas permettant sa réalisation. Désormais on pourrait donc lancer une recherche en prenant ce schéma comme filtre et fouiller le web pour trouver des instances (exemples).

3.2. Assister le rédacteur à trouver le mot recherché

La production du langage peut être vue comme une forme de réécriture de sens ou d'idées par des mots. L'hypothèse sous-jacente étant que les idées[8] précèdent les mots. Ayant à l'esprit un sens il y a plusieurs cas de figure lors de la consultation du dictionnaire. En particulier, on peut distinguer la qualité d'entrée (ce que le locuteur sait à cet instant précis : sens, et/ou forme) et la qualité de la sortie, proximité *formelle* et *sémantique* entre le mot source (MS, celui qu'il est capable de produire) et le mot cible (MC), le mot recherché. Autrement dit, l'information fournie au moment de la requête peut être très variable (sens, mots), tout comme la distance entre le MS et le MC qui peut être plus ou moins grande : (a) le MS et le MC peuvent avoir des rapports de sens ("chaud-froid"; "jaune-banane"; "fruit-banane", "dog-chien", etc.) sans qu'il n'y ait de rapport formel ; (b) ils peuvent avoir des rapports formels, sans avoir de rapports sémantiques (vin vs vingt) ;; (c) ils peuvent avoir et des rapports formels et des rapports sémantiques (chat-rat) ;; (d) le MS peut être formellement proche du MC (reléguer vs dé*léguer*).;

Seul le premier cas de figure nous intéresse ici. La réalisation d'une méthode d'accès par la forme a été décrite dans (Zock et Fournier, 2001). Nous ne nous y attardons donc pas ici. Nos efforts actuels sont concentrés sur l'accès par le sens ou plutôt sur l'accès par les mots liés à l'idée à exprimer (carnaval-Brésil). En clair, nous nous intéressons aux rapports associatifs, l'hypothèse étant, que le dictionnaire mental est un vaste réseau dont les mots sont les noeuds et les liens, les associations.[9] L'accès au mot s'effectuera par navigation. On entre dans le réseau en donnant un mot (MS) proche du MC pour recevoir en sortie tous les mots associés à ce dernier. En choisissant parmi ces éléments un candidat prometteur pour le soumettre à nouveau on s'approche progressivement du mot cible.

Prenons un exemple. Supposons qu'on cherche le mot ***infirmière*** (*MC*), alors que le seul mot qui nous vienne à l'esprit (*MS*) est ***hôpital***. Le système prendra alors celui-ci comme noyau pour présenter tous les mots (satellites) ayant un rapport direct avec lui, par exemple, les *employés/méronymes* (médecin, infirmière), des *sous-types* (clinique, sanatorium), etc. C'est à l'utilisateur de décider dans quelle direction continuer la recherche, car il sait généralement quel mot de ceux présentés par le système est le plus prometteur. Celui-ci deviendra donc à son tour le noyau (point de départ), susceptible de produire d'autres candidats. Le tout s'arrête lorsqu'on a trouvé le mot recherché.

À noter que le graphe de la figure 2a n'est qu'une *représentation interne* du système, l'utilisateur ne voit successivement que desque des paquets de mots, paquets qui sont créés

[8] Bien qu'en absence d'un support matériel, c'est moins trivial qu'il n'y paraît. Voir tous les débats tournant autour des questions concernant les rapports langage-pensée.

[9] Cette hypothèse est à nuancer cependant, selon qu'on parle d'une simulation des fonctionnalités de la mémoire mentale (comme nous le faisons ici) ou du cerveau lui-même. En lisant les travaux de certains psychologues, et en regardant leurs simulations on a l'impression que le cerveau est plutôt une usine "fabriquant" des mots qu'un lieu de stockage de mots (Zock, 2005).

grâce aux liens connus par le système (figure 2b). C'est justement pour éviter de noyer l'utilisateur dans un amas d'information qu'on a recours à ces liens qui serviront alors comme critères de sélection ou aides de navigation.

Bien entendu, pour pouvoir fonctionner selon la manière esquissée ici (navigation par association) il faut d'abord créer le réseau, typer les liens et déterminer leur poids. Nous avons commencé ce travail (Zock et Ferret, (2006) en utilisant l'extracteur de collocation de Ferret sur un corpus un peu particulier, le journal *Le Monde*. Bien entendu, le choix du corpus est très important dans la mesure où il est censé représenter les connaissances du monde de l'homme de la rue. Qui plus est, les poids devraient être pondérés en fonction du thème. Ce dernier point n'est pas un problème simple. Enfin, si l'intuition selon laquelle le dictionnaire mental (d'aucuns préféreraient parler d'encyclopédie) serait un vaste réseau, dont les *noeuds* sont des mots (et/ou des concepts), et les *liens* essentiellement des associations ne date pas d'hier, et si elle est partagée par de nombreux chercheurs[10], il n'y a à notre connaissance aucun inventaire exhaustif (ou de classification) de ces liens. Or, connaître leur nature est l'une des conditions préliminaires pour indexer un tel dictionnaire, et c'est là un de nos objectifs des années à venir. Pour une feuille de route, voir (Zock et Bilac, 2004).

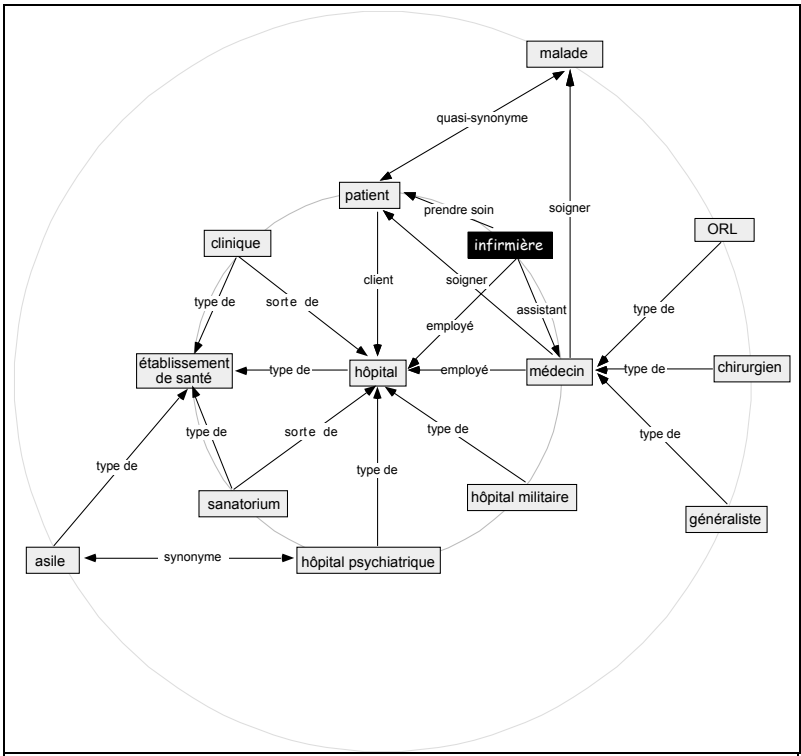


Figure 2a. Représentation interne

sorte de/type d'
établissement de santé
hôpital
asile
hôpital psychiatrique
hôpital militaire
sanatorium
clinique

méronymes
médecin
infirmière
malade

type de
ORL
Chirurgien
généraliste

Figure 2b. Candidats proposés

4. Conclusions

Les composants esquissés ci-dessus tournent tous autour de l'idée d'une meilleure utilisation des ressources lexicales, l'accent étant mis sur la *réutilisation* (capitalisation) des informations stockées dans le dictionnaire et sur l'aide à la navigation. Il est clair qu'un dictionnaire est un composant fondamental pour tout système de traitement de la langue. Mais un dictionnaire ne vaut que par l'information qu'il contient et par les moyens qu'il offre pour

[10] En effet, cette intuition se trouve déjà chez Aristote (« De memoria et reminiscentia »), chez des *psychologues* (Galton, 1880) et *psycholinguistes* (Deese, 1965). Enfin, cette idée est sous-jacente à WORDNET (Miller, 1990). Pour des synthèses en psycholinguistique voir (Hörmann, 1972 ; chapitres 6-10).

accéder à l'information. À l'heure actuelle, il y a un fossé énorme entre les dictionnaires papier, les dictionnaires électroniques et le dictionnaire mental. L'architecture particulière de ce dernier lui confère un énorme pouvoir en termes d'organisation et de souplesse d'accès. Contrairement à une hiérarchie avec une seule voie d'accès, dans ce réseau hautement interconnecté il y a presque toujours un moyen d'accéder à l'information recherchée. De ce fait, le dictionnaire mental constitue un excellent modèle en termes de stockage et d'accès d'informations. Si les dictionnaires traditionnels sont passifs et assez limités en termes d'accès, les dictionnaires électroniques ont un potentiel considérable, susceptible de présenter rapidement et sous des formes diverses l'information recherchée. Les idées présentées ici sont une première tentative allant dans ce sens, mais il est clair, que beaucoup de travail reste encore à faire, notamment au niveau des liens (associations).

Références

BESSE H. (1975) « De la pratique aux théories des exercices structuraux ». In *Études de Linguistique Appliquée* 20 : 8-30.

CHASTAIN K. (1969). « The audio-lingual habit learning theory vs. the code-cognitif learning theory ». In *IRAL* 7, 2 : 97-107.

CHOMSKY N. (1959). « Critique de *Verbal Behavior* de B.F. Skinner », dans *Language* 35 : 26-58.

DEESE J. (1965). *The structure of associations in language and thought*. Baltimore.

GALTON F. (1880). « Psychometric experiments ». In *Brain* 2 : 149-162.

HILGARD E., BOWER G. (1975). *Theories of learning*. Englewood Cliffs, N.J.

HÖRMANN H. (1972). *Introduction à la psycholinguistique*. Larousse, Paris.

KEKENBOSCH C. (1991). *La mémoire et le langage*. Nathan Université, Paris.

LE ROUZO M.L. (1975). « Y a-t-il une justification psychologique à la pratique des exercices structuraux ? ». In *Études de Linguistique Appliquée* 20 : 37-51.

LEITNER, S. (1972). *So lernt man lernen.* (Voici comment-on apprend). Heider, Freiburg.

MILLER G.A. (éd.) (1990). « WordNet: An On-Line Lexical Database ». In *International Journal of Lexicography* 3(4).

NERBONNE J., SMIT P. (1996). « GLOSSER: in Support of Reading ». In *COLING '96*, Copenhagen : 830-835.

POCZTAR J. (1971). « En enseignement programmé, quoi de nouveau? nouveau ? ». In Revue française de pédagogie 15 : 5-14. (voir aussi du même auteur: *Théories et pratique de l'enseignement programmé*).

SKINNER B.F. (1968). *La révolution scientifique de l'enseignement*. Dessart, Bruxelles.

SPOLSKY B. (1966). « A Psycholinguistic Critique of Programmed Foreign Language Instruction ». In *IRAL* 4, 2.

ZOCK M. (2005). « Le dictionnaire mental, modèle des dictionnaires de demain? demain ? ». In *Revue Française de Linguistique Appliquée*, Vol. X, 2005-2.

ZOCK M., FERRET O. (2006). « Enhancing electronic dictionaries with an index based on associations » (en préparation).

ZOCK M., QUINT J. (2004). « Why have them work for peanuts, when it is so easy to provide reward? Motivations for converting a dictionary into a drill tutor ». In *Papillon*, *5th workshop on Multilingual Lexical Databases*, Grenoble.

ZOCK M., BILAC S. (2004). « Word lookup on the basis of associations :associations: from an idea to a roadmap ». In *Proceedings of Coling workshop : Enhancing and using dictionaries*, Genève : 29-34.

ZOCK M., FOURNIER J.-P. (2001). « How can computers help the writer/speaker experiencing the Tip-of-the-Tongue Problem ? Problem? » In *Proceedings of RANLP*, Tzigov Chark : 300-302.

TAEMA : Traitement Automatique de l'Écriture de Mots Affectifs

Pierre-André Buvet, Fabrice Issac

Université Paris 13 – Laboratoire de Linguistique Informatique, CNRS
{pierre-andre.buvet ; fabrice.issac}@lli.univ-paris13.fr

Résumé

Nos recherches portent sur l'élaboration et l'utilisation de ressources linguistiques de type dictionnaire. Les dictionnaires utilisés sont capables d'apporter une aide à la rédaction. Ils sont bien plus que des dictionnaires « classiques » puisque l'acte de produire met en œuvre de nombreuses compétences liant la morphologie, la syntaxe et la sémantique. Nous présentons le prototype d'une application qui génère des phrases simples du français en se basant sur des critères sémantiques. Nous utilisons le modèle des classes d'objet qui est un modèle linguistique permettant de décrire le lexique du point de vue de ses propriétés syntactico-sémantiques.

Mots-clés : apprentissage des langues, dictionnaires électroniques, lexique, XML.

Abstract

Our research is concerned with the development and use of dictionary type resources. Such dictionaries are of assistance in the composition of texts and are considerably more pragmatic than classic dictionaries. The act of writing requires many integrated skills: syntax, semantic, morphology. We present the prototype of an application enabling the generation of simple French sentences based on semantic criteria. We use the model of object classes which is a linguistic model capable of managing the lexicon, subjacent syntax and semantics simultaneously.

Keywords: language learning, electronic dictionaries, lexicon, XML.

1. Introduction

Les techniques de TAL sont indispensables à chaque fois que l'objet à traiter informatiquement est de nature linguistique. Il est donc nécessaire d'appliquer ces techniques dans le cadre d'environnements d'apprentissage des langues. L'aide apportée est multiple. Elle peut prendre la forme d'évaluation de productions, de mise à disposition de ressources ou d'incitation à la production via des mises en situations diverses. Nous nous plaçons dans une autre optique. Nous proposons un environnement d'aide à la rédaction basée sur l'utilisation d'une ressource linguistique riche permettant de prendre en compte les descriptions morphologiques, syntaxiques et sémantiques.

Dans un premier temps, nous établirons un parallèle entre l'apprentissage des langues orienté vers le lexique et le traitement des données linguistiques fondé sur des dictionnaires électroniques pour justifier la spécificité du système d'aide à la rédaction que nous présentons. Dans un deuxième temps, nous présenterons la théorie linguistique qui a été prise en compte pour structurer les dictionnaires électroniques utilisés et nous discuterons des particularités de ces dictionnaires. Dans un troisième temps, une fois décrit le fonctionnement de notre système du point de vue de l'utilisateur, nous ferons état des choix effectués pour rendre compte sur le plan informatique des particularités du modèle.

2. Lexique, apprentissage et TAL

Les nombreux travaux existant dans le domaine de l'EAO et plus précisément dans le domaine de l'ALAO montrent que l'ordinateur apporte une aide efficace à l'apprentissage. Cet apport est d'abord d'ordre matériel, lié à la nature même du média informatique, mais aussi cognitif, c'est-à-dire lié à la manière dont l'homme et la machine interagissent.

L'apport principal, le plus visible, concerne les données et les types de données manipulés. En effet, l'ordinateur offre un accès rapide et facile à l'information tant au niveau qualitatif que quantitatif. De plus, le flux d'information est maîtrisé, l'utilisateur peut en effet filtrer l'information suivant des critères qui lui sont personnels. Les interfaces, dont le rôle est de présenter des informations à un utilisateur et de recevoir des ordres de celui-ci, contribuent également à ce que la vigilance de l'apprenant ne soit pas prise en défaut.

Un effort est fait, et cela semble nécessaire dans le cadre de l'apprentissage des langues, pour intégrer des techniques de TAL. Ainsi, les projets Free-Text (L'Haire *et al.*, 2003) et AlexiA (Chanier *et al.*, 1997) proposent des modules évolués d'analyse syntaxique. D'autres projets, tels les plateformes MIRTO (Antoniadis, 2005) et EXILLS[1], regroupent, outre des ressources, des outils classiques d'analyse textuelle. De ce point de vue, la ressource linguistique proprement dite n'est pas le centre des préoccupations puisque l'objectif est bien souvent de proposer un retour des productions de l'apprenant. D'autres modules sont mis en place : ils permettent l'accès à une ressource riche, l'objectif étant ici d'aider l'apprenant dans sa recherche d'information linguistique. Ainsi AlexiA et le DAFLES (Selva *et al.*, 2003) proposent des dictionnaires dont le contenu et la structure ont été pensés spécifiquement en fonction d'un objectif d'apprentissage.

Nos recherches s'orientent dans cette dernière direction. En effet nous disposons de dictionnaires suffisamment riches pour pouvoir générer des phrases simples (cf. infra). Après avoir présenté l'intérêt d'une orientation lexique pour l'apprentissage des langues et le TAL, nous indiquons nos vues quant à la mise en œuvre d'un système d'aide à la rédaction utilisant les ressources élaborées à partir du modèle des classes d'objets.

2.1. Lexique et apprentissage

L'apprentissage du vocabulaire n'est pas une tâche simple et les informations attachées aux mots sont multiples. Nous allons très rapidement exposer les motivations théoriques pour un apprentissage des langues orienté vers le lexique. Ces motivations sont d'ordre sémantique, syntaxique et pragmatique.

Tout d'abord du point de vue sémantique, on se rend compte que la signification d'un mot n'est pas monovalente, c'est-à-dire que le mot a une propension naturelle à la polysémie. C'est le contexte qui précise les différents sens des mots. Par ailleurs, syntaxe et lexique ne peuvent pas être considérés indépendamment. Même s'il est possible d'utiliser correctement un mot sans utiliser la syntaxe qui lui est propre, cette dernière devient nécessaire dès qu'il s'agit de communiquer à un niveau plus sophistiqué. L'aspect pragmatique en rapport avec le lexique implique la prise en compte de nombreux facteurs qui régissent la sélection du vocabulaire dans un discours : le type de discours, le statut de l'interlocuteur et du locuteur, les motivations, etc.

Un dernier argument en faveur d'un apprentissage orienté vers l'acquisition lexicale est que celle-ci est notoirement insuffisante. Le taux d'acquisition du vocabulaire en langue seconde

[1] http://www.exills.com/

est assez faible. Ainsi le nombre de mots réutilisables en production est de 1500 après cinq ou six ans d'apprentissage scolaire (Bogaards, 1995).

En résumé l'apprentissage d'un mot fait intervenir de nombreuses connaissances : savoir dans quel contexte on l'utilise à l'oral ou à l'écrit (fréquence d'utilisation, cooccurrences associées), appréhender les limitations de son usage selon les variations de situations et connaître ses comportements morphologiques, syntaxiques et sémantiques.

2.2. Lexique et TAL

La polysémie et la polymorphie, d'une part, le figement, d'autre part, sont sources de nombreuses difficultés en TAL. Nous indiquons en quoi ces deux catégories de phénomènes langagiers posent problème.

2.2.1. Polysémie et polymorphie

Il est courant de traiter le lexique en distinguant trois niveaux d'analyse : la morphologie, la syntaxe et la sémantique. Un traitement différencié des mots se heurte à deux difficultés : la polysémie et la polymorphie.

La possibilité pour une forme donnée de recevoir des sens différents (<sentiment appréciatif> et <réaction appréciative> en ce qui concerne admiration) est associée à ses particularités syntaxiques (les deux sens de admiration donnent lieu à des complémentations et des constructions différentes : Luc éprouve de l'admiration (pour + *devant) Luc / Luc éprouve de l'admiration (*pour + devant) ce tableau). Par conséquent, il vaut mieux recourir à une approche qui tienne compte conjointement des propriétés morphologiques, syntaxiques et sémantiques des mots pour faire état de leur caractère polysémique.

La polymorphie est une particularité des seuls prédicats. Un même emploi peut avoir différentes formes. Ainsi, Luc a du mépris pour Léa, Luc méprise Louise et Luc est méprisant envers Léa sont trois phrases équivalentes. Elles sont constituées d'un même prédicat (mépris-) associés aux deux mêmes arguments (Luc et Léa). Chaque forme du prédicat est caractérisée par une construction spécifique. De nouveau, il est clair que seul un traitement conjoint de la morphologie, de la syntaxe et de la sémantique est adéquat pour rendre compte des différentes formes d'un prédicat.

L'importance de la polysémie et celle de sa contrepartie, la polymorphie, ont été sous-estimées en TAL. Ce qui explique de nombreuses insuffisances des systèmes qui opèrent sur des données linguistiques. Une amélioration qualitative des systèmes implique que les variations de sens rattachées à une forme ou les variations de formes rattachées à un sens soient systématiquement traitées. Pour ce faire, nous défendons le point de vue suivant : il faut implémenter dans les systèmes des dictionnaires électroniques à large couverture lexicale qui permettent de rendre compte de la polysémie et de la polymorphie ; autrement dit, des dictionnaires qui intègrent des descripteurs de nature morphologique, syntaxique et sémantique.

2.2.2. Figement

Les expressions figées correspondent à des séquences de mots dont l'agencement prête peu à la variation et n'est jamais régi par les règles combinatoires usuelles. Il s'ensuit que leur interprétation n'est généralement pas déductible de celle de leurs constituants. Ainsi, le fait que la suite être bouche bée devant est un synonyme de admirer n'est pas directement explicable à partir du sens des mots qui la constituent. Par ailleurs, de nombreuses expressions figées peuvent correspondre à une séquence libre (typiquement prendre le taureau

par les cornes) de telle sorte qu'il est difficile de déterminer quelle interprétation privilégiée sur la seule base de l'assemblage de leurs formes.

Nous considérons que la prise en compte des expressions figées en TAL implique qu'elles fassent l'objet de recensements exhaustifs et systématiques. La seule notion de collocation (au sens de grande fréquence de proximité entre au moins deux mots) est insuffisante car elle rassemble des données très hétérogènes. Il faut notamment distinguer les expressions figées selon qu'elles procèdent de la signification lexicale ou de la signification grammaticale (Blanco et Buvet 2004).

Dans le premier cas, elles fonctionnent comme des adjectifs, des noms ou des verbes ; il est impératif alors de les répertorier dans des bases de données et d'y indiquer leurs particularités morphologiques, leurs propriétés distributionnelles et les classes sémantiques qui les caractérisent. Par exemple, la séquence prendre la poudre d'escampette est catégorisée comme une locution verbale dont le domaine d'arguments unaire se rapporte nécessairement à un humain et est rattachée à la classe sémantique <déplacement humain>.

Dans le second cas, il s'agit principalement de séquences déterminatives (comme un éclair de dans Luc a eu un éclair de génie) ou adverbiales (par exemple comme un pinson dans Luc est gai comme un pinson). Leur signification peut être rapportée à un petit nombre de valeurs ('quantité forte', 'quantité faible', 'intensité forte', 'intensité faible', 'mélioratif', 'péjoratif', 'fréquentatif', etc.). Généralement, la portée distributionnelle de ces séquences est réduite puisqu'elles se combinent avec un nombre très limité de noms, d'adjectifs ou de verbes (*Luc a eu un éclair de esprit ; *Luc est content comme un pinson). Pour autant, leur association avec d'autres mots ne correspond pas à des expressions figées puisque ces mots peuvent apparaître sans les séquences en question (Luc a (un éclair de + du) génie ; Luc est gai (E + comme un pinson). Les séquences déterminatives et adverbiales figées doivent être également toutes recensées et il convient d'expliquer les particularités de leur mode de fonctionnement (notamment avec quels items elles sont compatibles et quelles sont leurs valeurs).

2.3. Mise en place d'un SAR

L'importance de la part du vocabulaire dans l'apprentissage d'une langue et la possibilité en TAL d'améliorer les systèmes en implémentant des dictionnaires électroniques à large couverture constitue un point de convergence. Cela nous a conduits à développer un système d'aide à l'apprentissage qui met en avant la dimension lexicale des langues. D'autant plus que le FLE (Français Langue Étrangère) et le TAL ont des problématiques communes vis-à-vis des mots : le traitement de la polysémie, de la polymorphie et du figement sont les mêmes pour un apprenant ou un système opérant sur des données linguistiques. La nécessité de prendre en compte conjointement les propriétés morphologiques, syntaxiques et sémantiques des mots pour expliquer leur mode de fonctionnement est valable dans l'un et l'autre cas.

Le système d'aide à la rédaction développé s'intitule TAEMA (Traitement Automatique de l'Écriture de Mots Affectifs). Il s'agit de produire des phrases centrées sur le vocabulaire affectif du français pour des apprenants en langue seconde ou pour des apprenants natifs. Un affect est défini comme un état psychologique (par exemple, la 'joie', la 'peur', 'l'amour', le 'regret') qui est ressenti et qui n'est pas inhérent à un individu (cf. infra). La particularité du dictionnaire des prédicats d'<affect> utilisé permet au système non seulement d'indiquer tout le vocabulaire en rapport avec un type d'affect donné (par exemple une émotion ou un sentiment particulier) mais aussi toutes les constructions associées avec ce vocabulaire. Au final, après avoir fait une requête relative à un type d'affect, l'utilisateur a la possibilité de choisir parmi un ensemble de phrases équivalentes celle qui lui semble la plus adéquate pour sa production écrite.

Avant de présenter TAEMA, nous discutons des particularités linguistiques des ressources lexicales utilisées dans le système.

3. Les données linguistiques : théorie et applications

Le modèle des classes d'objets a donné lieu à la réalisation de dictionnaires électroniques destinés aux divers systèmes opérant sur des données linguistiques. Ces dictionnaires visent à une couverture exhaustive du français, entre autres langues. Il s'agit de décrire le lexique avec des propriétés syntactico-sémantiques explicites et reproductibles qui sont susceptibles de faire l'objet de procédures informatisées.

Nous présentons en premier lieu le modèle des classes d'objets. Nous illustrons ensuite la description du lexique qui en résulte avec les prédicats d'<affect> dans la mesure où ce sont ces données lexicales qui sont implémentées dans le système TAEMA. Nous terminons par une présentation du mode de structuration des dictionnaires électroniques.

3.1. Le modèle des classes d'objets

Le modèle des classes d'objets (Gross 1995, Gross 1996, Le Pesant et Mathieu-Colas, 1998) postule que toute phrase est constituée d'un prédicat et de ses arguments et que les autres unités linguistiques ressortissent à l'actualisation. Les prédicats prédominent structurellement les arguments et les conditions d'apparition des différents actualisateurs sont subordonnées aux différentes relations entre les prédicats et les arguments constitutives de phrases.

Le modèle a comme conséquence que la partition des unités linguistiques en fonction de leur statut de prédicat, d'argument élémentaire, ou d'actualisateur ne recoupe pas celle qui a trait aux parties du discours traditionnelles. Ainsi, les prédicats peuvent correspondre, entre autres, à des verbes (chérir dans Luc chérit Léa) ; des adjectifs (épris dans Luc est épris de Léa) ; des noms (béguin dans Luc a le béguin pour Léa). De même, les verbes peuvent être soit prédicatifs (gifler dans Léa a giflé Luc) soit supports (donner dans Léa a donné une gifle à Luc).

Le modèle des classes d'objets permet de prendre en compte la polysémie. Ainsi, pour ce qui est des prédicats verbaux, il est fait état de leurs schémas d'arguments. Cela a conduit en particulier à factoriser les arguments en les typant sémantiquement en termes de classes. Pour chaque emploi identifié sans ambiguïté, la description gagne en pertinence à tous les niveaux d'analyse : la conjugaison (certains emplois se distinguent par des particularités de temps, de nombre ou de personne), la dérivation (le lien entre les formes est dépendant du sens), la synonymie et les restructurations (passif, extraction, etc.). Par exemple, la forme conduire a autant d'emplois que de compléments nominaux définis soit en intension (par le biais de traits ou de classes d'objets) soit en extension : conduire un ami à l'hôpital (conduire1 = emmener) ; conduire le troupeau à l'abreuvoir (conduire2 = mener) ; conduire l'autocar (conduire3 = piloter) ; conduire la forêt (conduire4 = aménager) ; conduire la chaleur (conduire5 = véhiculer) ; etc.

Les classes d'objets constituent des ensembles d'items sémantiquement homogènes définis à l'aide de propriétés syntaxiques. On distingue les classes d'arguments, d'une part, les classes de prédicats, d'autre part. Les premières résultent d'une sous-catégorisation syntactico-sémantique des substantifs correspondant aux arguments élémentaires. Les secondes ont trait essentiellement à des adjectifs, des noms et des verbes. Ce classement est pris en compte dans la description des prédicats en termes de classes. Il s'agit cependant d'une catégorisation effectuée sur la base de leurs propriétés syntaxiques et sémantiques remarquables. De ce fait, la caractérisation syntactico-sémantique d'un prédicat prime sur ses particularités

morphologiques lorsqu'il recouvre deux de ces formes, voire les trois. Ainsi, les phrases Luc a désiré cela, Luc a été désireux de cela et Léa a eu le désir de cela sont considérées comme strictement équivalentes dans la mesure où les arguments Luc et cela se rapportent à un même prédicat qui recouvre soit la forme verbale (désirer) soit la forme adjectivale (désireux) soit la forme nominale (désir). Le modèle rend donc également compte de la polymorphie.

Depuis de nombreuses années, les séquences figées du français ont fait l'objet de recensements exhaustifs. Plus de 100 000 noms composés de la langue générale ont ainsi été recueillis. Parallèlement, un travail similaire est en cours pour les langues de spécialité. D'autres travaux portent sur les adjectifs et les verbes composés (à ce jour 10 000 locutions adjectivales et 30 000 locutions verbales ont été listées). Les recensements exhaustifs et systématiques entrepris recouvrent une grande partie des constructions figées du français. Leur spécification dans les dictionnaires électroniques est une contribution majeure au traitement du figement. Ainsi, le système INTEX (Silberztein, 1993) intègre l'ensemble de ces dictionnaires.

3.2. Les prédicats d'<affect>

Nous présentons rapidement les prédicats d'<affect> en termes de classes d'objets dans la mesure où c'est le vocabulaire qui a été implémentée dans le système TAEMA.

La langue française dispose de toutes sortes d'adjectifs, de noms et de verbes pour exprimer ce qui se rapporte à l'intériorité mentale des êtres humains, une grande partie de ce lexique étant considérée par les linguistes comme des termes psychologiques.

L'intériorité mentale se scinde en deux selon sa nature cognitive ou bien psychologique. Il est possible de répartir les prédicats de la seconde catégorie comme suit : (i) les prédicats de <disposition d'esprit> (colérique) ; (ii) les prédicats d'<appréciatif psychologique> (désobligeant) ; (iii) les prédicats d'<affect> (amoureux).

Une justification rapide de cette sous-catégorisation tient en deux points, plus précisément deux oppositions. La première opposition se manifeste comme suit : les prédicats de la première sous-catégorie correspondent à des états psychologiques inhérents, les prédicats des deux autres sous-catégories à des états psychologiques contingents. Le caractère permanent ou occasionnel de l'état auquel se rapporte un prédicat explique l'incompatibilité ou la compatibilité du prédicat avec des adverbes à valeur durative. Ainsi, constamment se combine difficilement avec le prédicat dans le premier cas, la caractérisation aspectuelle imputable à l'adverbe est superfétatoire par rapport au trait permanent du prédicat, mais beaucoup plus facilement dans le second cas, l'adverbe annihile le trait occasionnel du prédicat en indiquant un aspect duratif : ? Luc est constamment peureux : prédicat du type (i) ; Luc est constamment sévère avec Max : prédicat du type (ii) ; Luc est constamment apeuré : prédicat du type (iii).

La deuxième opposition résulte du fait que les prédicats de la troisième sous-catégorie ont comme particularité d'être en rapport avec un 'ressenti intérieur' alors que les prédicats des deux autres sous-catégories ont trait à un 'jugement extérieur'. Les verbes d'<opinion>, *e.g.* trouver, rendent compte de la distinction en admettant plus facilement les complétives formées à partir d'un prédicat de <disposition d'esprit> ou de <comportement> que celles formées à partir d'un prédicat d'<affect> : Je trouve que Luc est (avenant + acariâtre) : prédicats du type (i) ; Je trouve que Luc est (bienveillant + désobligeant) avec Max : prédicats du type (ii) ; Je trouve que Luc est (amoureux de Léa+ dégoûté de la vie) : prédicats du type (iii).

En résumé, les prédicats d'<affect> sont des adjectifs, des noms et des verbes en rapport avec des états psychologiques contingents qui sont de l'ordre du ressenti. Le champ d'étude lexicale une fois précisé, une nomenclature a été établie.

3.3. Typologie et structuration des dictionnaires du LLI

L'exploitation des dictionnaires traditionnels (typiquement Le Robert) en TAL pose de nombreuses difficultés étant donné qu'ils ne sont pas complets (seuls les sens les plus fréquents sont traités), que les descriptions linguistiques sont insuffisantes (toutes les informations nécessaires à la construction des mots ne sont pas spécifiées et celles relatives à l'aspect inhérent ou contextuel sont totalement ignorées), que les classifications sémantiques sont peu cohérentes (on peut difficilement retrouver les mots qui appartiennent à la même classe). Il est plus aisé d'implémenter dans les systèmes qui opèrent sur des données linguistiques des dictionnaires électroniques correspondant à des bases de données où une entrée lexicale est associée à des descripteurs normalisés (Courtois et Silberztein 1990).

Les dictionnaires électroniques pris en compte ici ont les particularités suivantes : (i) ils rapportent les unités lexicales à la distinction prédicats de premier ordre/arguments élémentaires (cela a pour conséquence que les différentes formes d'un même prédicat sont décrites de la même façon) ; (ii) les descripteurs associés à chaque entrée ne sont pas hiérarchisés (ils sont de nature morphologique, syntaxique et sémantique) ; (iii) les descripteurs se rapportent au modèle des classes d'objets.

Le recours à des dictionnaires électroniques dans un système d'aide à la rédaction permet de produire toutes sortes de phrases afin d'aider les rédacteurs à améliorer leur production.

4. Système d'aide à la rédaction de mots affectifs

L'aide à l'écriture, d'un point de vue informatique, peut être envisagée à partir de plusieurs fonctions :

- la planification : recherche d'idées, dictionnaires, banques de textes, générateur de textes ;
- l'édition : dactylographie, couper/copier/coller ;
- la correction : recherche/remplace, correcteur orthographique et grammatical ;
- la présentation : mise en page, maquettage, impression ;
- la collaboration : écriture en chaîne, écriture collaborative, discussion et document partagé.

Par ailleurs, la notion d'interactivité est primordiale (Mangenot 2000). Elle fait partie des critères d'évaluation des outils et peut être plus ou moins développée. Dans cette optique, on peut proposer des activités d'écriture basées sur le dialogue (Caviglia *et al.*, 1994). Cependant le dialogue repose sur les types du discours (narration, argumentation, etc.) et non pas, comme nous le proposons, sur un modèle linguistique permettant de rendre compte à la fois de la morphologie, de la syntaxe et de la sémantique.

Nous allons maintenant présenter l'interface du système et la structure interne utilisée pour rendre compte de la richesse du modèle linguistique.

4.1. L'interface de TAEMA

Le prototype développé est susceptible de s'adresser aussi bien à des apprenants L1 que des apprenants L2 pour peu que ces derniers aient un niveau de langue minimum. En effet, l'utilisation de l'interface nécessite une connaissance sommaire du français. L'usage du

logiciel présuppose que les utilisateurs sachent exprimer au moins la phrase prototypique associée à la notion et l'environnement spécifié (par exemple *Luc aime Léa* pour l'affect <amour>). La finalité du logiciel est de lui proposer toutes les paraphrases possibles (*Luc a le béguin pour Léa, Luc en pince pour Léa, Luc est fou de Léa*, etc.). Il est cependant à noter que nous présentons un prototype non scénarisé et qu'il conviendrait de l'adapter à l'apprenant. La seule contrainte concerne le niveau minimal de l'apprenant.

Du point de vue de l'utilisateur du système, il s'agit d'indiquer un type d'affect donné et des éléments contextuels afin d'obtenir toutes les phrases canoniques du français qui sont en rapport avec les indications fournies. Le système développé permet à un utilisateur de sélectionner (sous forme de menus déroulants ou en les spécifiant en langue naturelle) les différents éléments d'une phrase à produire. Nous détaillons les différents écrans qui permettent à l'utilisateur de formuler sa requête.

Le premier écran présente : (i) la finalité du projet ; (ii) un mode d'emploi ; (iii) un bouton 'entrée' qui permet d'accéder au système.

Le deuxième écran invite l'utilisateur à choisir un type d'affect dans un champ de saisie à l'aide d'un menu déroulant. Il s'agit de l'appellation de l'une des sous-classes de prédicats d'<affect> qui figurent dans le dictionnaire électronique utilisé par le système. Cette page comporte également trois autres champs de saisie qui ne sont pas activés à ce niveau d'utilisation. Le premier est dit 'expérienceur' : il s'agit de la personne qui ressent l'affect. Le second est dit 'bénéficiaire/objet' : il s'agit de l'éventuelle personne ou entité sur qui porte l'affect. Le troisième est dit 'temps' : il permet de conjuguer le prédicat à un temps grammatical simple.

Le troisième écran rend actifs en partie ou en totalité les champs de saisie autres que celui déjà rempli. Selon que le prédicat est monadique ou dyadique (s'il comporte un ou deux arguments), les seuls champs 'expérienceur' et 'temps' ou ces deux champs ainsi que le champ 'bénéficiaire/objet' sont disponibles. Les champs 'expérienceur' et ''bénéficiaire/objet' présentent deux options : d'une part, si l'utilisateur souhaite que les arguments correspondent à des pronoms, il fait appel à un menu déroulant où il spécifie la personne, le nombre, le genre du pronom désiré ; d'autre part, s'il veut que les arguments soient des expressions (groupes nominaux, complétives, infinitives ou entités nommées), il est invité à remplir lui même le champ tout en spécifiant le nombre et le genre de l'expression.

Le quatrième écran donne tous les résultats relatifs à la requête de l'utilisateur de telle sorte qu'il puisse choisir une des phrases relatives à l'affect spécifié et les éléments contextuels qu'il a précisé. Le prototype basé sur environ 60 sous-classes permet de générer à peu près 3000 phrases simples du français.

4.2. Nécessité de structuration des données

Les ressources élaborées sont structurées via XML. Ce choix permet d'utiliser une gamme très large d'outils mais surtout de définir une ressource très riche qui peut servir de format pivot. Cette ressource pouvant être déclinée en fonction du support ou de l'objectif à atteindre.

Il faut noter que nous utiliserons plusieurs DTD plutôt qu'une seule pour affiner au maximum le balisage. Jusqu'à présent, nous avons établi que le dictionnaire des prédicats nécessitait au moins cinq DTD (chacun repéré par un espace de nom spécifique).

1. En-tête : contient les informations portant sur l'utilisateur ayant fait la dernière modification, mais aussi la date de la dernière modification, le nombre de révisions... On s'inspirera pour cet en-tête du travail réalisé pour la XCES (XML Corpus Encoding Standard). Cette partie permet d'envisager un développement coopératif d'une part et de gérer l'historique de l'évolution des dictionnaires ;
2. Le document maître : son rôle est de regrouper les différentes parties du dictionnaire ;
3. Les prédicats ;
4. Les arguments ;
5. Hiérarchie : décrit le lien qui existe entre les classes et les hyperclasses.

La figure 1 décrit la manière dont les différentes informations se combinent. Deux DTD supplémentaires sont envisagées pour l'instant et concerneraient la structuration des déterminants et des prépositions.

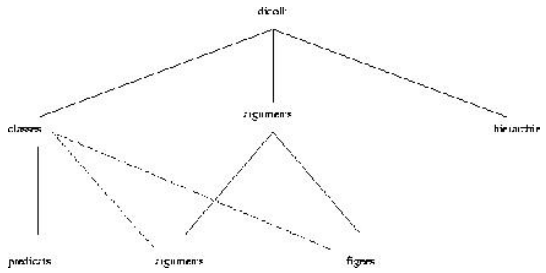


Figure 1 : organisation des différents composants des dictionnaires

Le nombre de DTD peut sembler important, mais ce découpage permet de résoudre un certain nombre de problèmes. En effet, un argument pouvant appartenir à différentes classes d'argument de par la polysémie, il est nécessaire que cette information soit déportée pour éviter un phénomène de redondance. Pour les mêmes raisons, l'hyperclasse d'une classe d'argument, qui selon le schéma d'argument du prédicat peut varier, doit être déportée. Ces possibilités doivent donc apparaître au sein de la structure du dictionnaire. Un autre argument en faveur du découpage concerne la mutualisation des informations par les différents concepteurs des dictionnaires. En effet, la constitution des dictionnaires de prédicats nécessite de connaître l'ensemble des classes d'arguments afin de rédiger le schéma ; de même la hiérarchie des classes doit être mutualisée.

Exemple de description du nom commun artère dans le document XML décrivant les arguments :

```
<elt type = "arg">
        <nom>artère</nom>
        <liste_classe>
                <classe id="voie"/>
                <classe id="partie interne corps"/>
        </liste_classe>
</elt>
```

Le substantif artère est un nom commun polysémique. Ce nom a deux sens et appartient donc à deux classes d'argument : <les voies de communication>, notée ici <voie> et les <éléments internes du corps d'un animal ou d'un humain>, notée ici <partie interne corps>.

Exemple de description du nom commun visage :

```
<elt type = "arg">
        <nom>visage</nom>
        <liste_classe>
                <classe id="npc"/>
        </liste_classe>
</elt>
```

Exemple de description du schéma d'argument du prédicat verbal porter :

```
<schema>
     <predicat>porter</predicat>
     <comp type ="sujet">
            <hyperclasse id="humain"/>
     </comp>
     <comp type ="COD">
            <hyperclasse id ="inc"/>
            <classe id = "vetement"/>
     </comp>
</schema>
<schema>
        <predicat>porter</predicat>
        <comp type ="sujet" pos="0">
               <hyp erclasse id="humain"/>
        </comp>
        <comp type ="COD" pos="1">
               <hyperclasse id ="ina"/>
               <classe id = "appellation"/>
        </comp>
</schema>
```

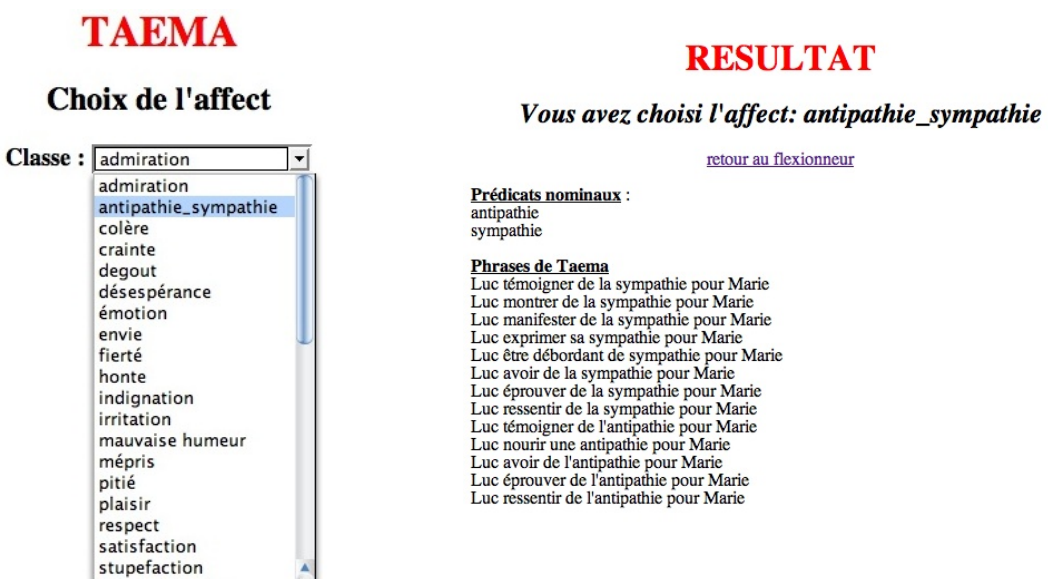
Le prédicat verbal porter est polysémique. Chaque sens possède un schéma d'arguments spécifique. Ainsi, le premier schéma d'argument présenté ici correspond à une phrase du type *Jeanne porte une jupe* tandis que la seconde phrase correspond à une phrase du type *Emeric porte un nom breton*.

5. Conclusions/Perspectives

Nous avons présenté le prototype d'une application permettant de générer des phrases simples du français. L'objectif étant de fournir à un apprenant un outil lui permettant d'appréhender le contexte associé au lexique. L'organisation des informations dans le cadre de ce prototype est primordiale. Nous avons utilisé le modèle des classes d'objet qui est un modèle linguistique permettant de rendre compte, comme nous le souhaitons du lexique du point de vue de ses propriétés syntactico-sémantiques.

Notre objectif initial était de valider le modèle linguistique, tant du point de vue théorique que fonctionnel, dans le cadre d'un outil d'aide à la rédaction. Il nous a donc fallu définir les limites de ce qui pouvait être utilisé, sans que le modèle en soit affecté. Ce travail a été réalisé de paire avec l'élaboration d'un modèle informatique qui permet de représenter toute la richesse de l'information linguistique. Suite à ce travail théorique, il était nécessaire de valider les ressources proprement dites. Les copies d'écran ci-dessous montrent un exemple d'utilisation du prototype[2].

[2] Le prototype présenté ici n'intègre pas encore le module de flexion morphologique. La version finale permettra d'obtenir par exemple « *Luc témoigne de la sympathie pour Marie* ».



La réalisation de ces trois aspects du projet s'est fait en parallèle et de manière itérative puisque toute incohérence ou difficulté nous obligeaient à adapter les modèles linguistiques et informatiques sous-jacents. Bien évidement, le prototype réalisé reste très incomplet en ce qui concerne sa couverture (il permet néanmoins de générer l'ensemble des phrases en rapport avec le lexique de l'affect).

Le travail que nous devons maintenant entreprendre concerne tout à la fois la couverture linguistique et l'interface avec l'utilisateur. Nous avons en effet focalisé nos efforts sur la pertinence des ressources. La présentation de ces ressources, qui sont par nature très complexes, reste un point de réflexion important. Couplé à ce module d'aide à la rédaction, il nous faut envisager à présent une scénarisation pédagogique propre à guider les apprenants.

Références

ANTONIADIS G., ECHINARD S., KRAIF O., LEBARBÉ T., PONTON C. (2005). « Modélisation de l'intégration de ressources TAL pour l'apprentissage des langues : la plateforme MIRTO ». In *ALSIC* 8 : 65-79.

BLANCO X., BUVET P.-A. (2004). « Verbes supports et significations grammaticales. Implications pour la traduction espagnol-français ». In *Lingvisticae Investigationes* 27 (2) : 327-342.

BOGAARDS P. (1995). *Le vocabulaire dans l'apprentissage des langues étrangères*. LAL, CREDIF, Hatier/Didier.

BUVET P.-A., GIRARDIN C., GROSS G., GROUD C. (2005). « Les prédicats d'<affect> ». In *LIDIL* 32 : 125-143.

CAVAGLIA F., FERRARIS M. (1994). « Scrivere in collaborazione con il computer ». In D. Camilleri, F. Mangenot et G. Poletti (éds), *La Plume et l'Écran. Actes des Journées d'écriture créative*. Petrini Editore, Turin : 113-119.

CHANIER T., FOUQUERÉ C., ISSAC F. (1997). « AlexiA : un environnement d'aide à l'apprentissage lexical du français langue seconde ». In P. Fiala, P. Lafon et M.-F. Piguet (éds), *La locution : entre lexique syntaxe et pragmatique*. Klincksieck, Paris : 105-118.

COURTOIS B., SILBERZTEIN M. (1990). « Dictionnaires électroniques du français ». In *Langue française* 87.

GROSS G. (1996a). « Les expressions figées en français : noms composés et autres locutions ». Ophrys, Paris-Gap.

GROSS G. (1996b). « Prédicats nominaux et compatibilité aspectuelle ». In *Langages* 121 : 54-72.

L'HAIRE S.,VANDERVENTER FALTIN A. (2003). « Diagnostique d'erreurs dans le projet Free-Text ». In *ASLIC* 6 (2).

LE PESANT D., MATHIEU-COLAS M. (1998). « Introduction aux classes d'objets », In *Langages* 131 : 6-33.

MANGENOT F. (2000). « Aide à l'écriture ou environnements d'écriture ? ». In J. Anis et N. Marty (coord.), *Lecture-Écriture et Nouvelles Technologies*. La collection de l'ingénierie éducative, CNDP, Paris : 59-68.

MEJRI S. (1997). « Le figement lexical : descriptions linguistiques et structuration sémantique », *série linguistique X*, Publications de la Faculté des lettres de la Manouba.

SELVA T.,VERLINDE S., BINON J. (2003). « Vers une deuxième génération de dictionnaires électroniques ». In *Revue TAL* 44 (2) : 177-197.

SILBERZTEIN M. (1993). *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*. Masson, Paris.

A Computational Approach to the Discovery and Representation of Lexical Chunks

David Wible[1], Chin-Hwa Kuo[2], Meng-Chang Chen[3], Nai-Lung Tsao[3], Tsung-Fu Hung[2]

[1] Department of English, Tamkang University, Taipei
wible45@yahoo.com

[2] Computer and Network Lab, Tamkang University, Taipei

[3] Institute of Information Science, Academia Sinica, Taipei

Résumé

La connaissance des « chunks » (tronçons) lexicaux est maintenant reconnue comme une compétence essentielle pour l'apprentissage d'une seconde langue. Nous étudions deux des principaux problèmes que les « chunks » posent en lexicographie et nous présentons des méthodes de résolution informatiques. Le premier problème est celui de l'apprentissage de connaissances lexicales, c'est-à dire la nécessité de définir quelles suites de mots constituent des « chunks » utiles à l'apprenant. Le deuxième problème est celui de la représentation, c'est-à-dire comment mettre cette connaissance à la disposition de l'apprenant. Pour résoudre le premier problème, nous proposons un algorithme glouton exécuté sur un corpus de 20 millions de mots du BNC qui reproduit des mesures d'associations de mot sur des n-grams de plus en plus longs. Cette approche donne la priorité à un rappel élevé et tente d'isoler les faux positifs à l'aide de mécanismes de tri. Pour résoudre le problème de la représentation, nous nous proposons d'associer cet algorithme à un navigateur en tant qu'extension de notre outil de détection de collocations.

Mots-clés : « chunks » (tronçons) lexicaux, lexicographie computationnelle, association de mots, apprentissage de langues étrangères.

Abstract

Lexical chunks have in recent years become widely recognized as a crucial aspect of second language competence. We address two major sorts of challenge that chunks pose for lexicography and describe computational approaches to addressing these challenges. The first challenge is lexical knowledge discovery, that is, the need to uncover which strings of words constitute chunks worthy of learners' attention. The second challenge is the problem of representation, that is, how such knowledge can be made accessible to learners. To address the first challenge, we propose a greedy algorithm run on 20-million words of BNC that iterates applications of word association measures on increasingly longer n-grams. This approach places priority on high recall and then attempts to isolate false positives by sorting mechanisms. To address the challenge of representation we propose embedding the algorithm in a browser-based agent as an extension of our current browser-based collocation detection tool.

Keywords: lexical chunks, computational lexicography, word association, foreign language learning.

1. Introduction

There has been a recent growing trend in foreign language education research to recognize words as inextricably entwined with a range of syntagmatic contexts and contextual patterns as opposed to viewing them as discrete units that can be mastered in isolation. With this has come the recognition that multiword expressions are a central rather than peripheral aspect of lexical knowledge (See Wray 2002 for an extensive review). Among the range of multiword phenomena, certain types, such as phrasal verbs and collocations, have enjoyed a long tradition of attention in lexicography and language pedagogy. Other sorts of multiword

expressions such as lexical chunks and formulaic sequences, however, have only recently begun to attract this sort of attention. Knowledge resources concerning these expressions are thus correspondingly scarce. The work reported here is part of a larger project aimed at filling this gap in resources for the learning of lexical chunks.

2. Lexical Chunks and the Two Challenges to Lexicography

Lexicography traditionally has faced two kinds of challenges: lexical knowledge discovery and lexical knowledge representation. We want to suggest that lexical chunks introduce novel requirements to both sorts of challenge. In what follows we consider these two in turn and present our approach to each as we go. The literature on chunks includes a variety of criteria for defining what constitutes a chunk (see Weinert 1995 for an overview) from semantic and syntactic criteria such as (non-)compositionality and (non-)productivity to processing criteria, the most commonly cited of the latter being that users store and retrieve chunks as single multiword units rather than by rule-governed composition in real time. Implementing any of these criteria computationally for real world applications involves serious problems of scalability. The most promising trait of chunks in this respect is frequency of occurrence, yet the implementation of frequency must be appropriately nuanced. Taking inspiration from the achievements of statistical approaches to collocation extraction (Church et al 1991; Smajda 1993, inter alia), we choose as a practical starting point to operationalize the notion of lexical chunk statistically. Our specific implementation of a statistical approach is elucidated in what follows as part of our approach to lexical chunk discovery.

3. Lexical Chunks and Knowledge Discovery

A fundamental task of the lexicographer is to uncover facts concerning the behavior and meaning of words. These facts constitute the substance of any dictionary. One of the most important developments in this work over the past few decades has been the rise of machine-readable corpora and computational tools that aid in their analysis. Learner dictionaries in particular have benefited from the lexicographers' extended capacity to distill patterns and nuances governing the use of specific words made possible by access to massive collections of texts and tools for mining them (Sinclair 1987, inter alia). Certain multiword expressions have proven particularly susceptible to discovery with such tools. Inspired by the work of Church and his colleagues in the late 1980's (Church *et al.*, 1991), researchers have found, for example, that collocations are detectable by well-understood statistical measures of word association strength such as mutual information (MI). There are certain characteristics of collocations and, say, phrasal verbs as well, that render them especially vulnerable to detection by these statistical tools. Word association measures can detect the association strength between two event types. Construing a pair of lexemes occurring in close proximity to each other to be a pair of events in running text, lexicographers can use these word association measures to determine the strength of association between these two events and thus detect which word pairs have sufficiently strong association to be collocations.

Lexical chunks as a class of multiword expression, however, are much less homogeneous in this respect. A chunk could consist of two words (*e.g.*, *in fact*) or five words (*as a matter of fact*). In fact, there would seem to be no principled limit to the size of a chunk. What we want to suggest is, first, that this property of chunks creates a non-trivial challenge to computational lexicography and, second, that it is tractable.

We have developed a computational approach to the extraction of lexical chunks from very large corpora. Our algorithm takes as its kernel a simple word association measure (any such

measure can serve as this kernel, for example, mutual information or simple conditional probability, inter alia) and iterates its application. Unlike traditional word association measures, then, our algorithm is greedy. As a consequence, our approach is insensitive to chunk size. The same algorithm that detects *in fact* also detects *as a matter of fact*.

The current interface is designed for lexicographers, not for learners (the design of a representation for learners is the topic of the next section). The algorithm takes as its input a single word/POS pairing. A sample input would be *fact/N* (*i.e.*, the noun *fact)*. The current version of the algorithm runs both a conditional probability measure and MI for the target word paired with each possible candidate collocate occurring within a five-word window of the target word in a 20-million-word portion of BNC (See Wible *et al.*, 2004a for details). The user selects a minimum association score to set as a threshold for each of the two measures; a default threshold is used if the user makes no selection of score threshold. Word association scores are then tabulated for every word that has tokens which occur within the five-word span of any token of the target word (*e.g.*, *fact*). These pairings are sensitive to linear order; for example, tokens of the pair *in…fact* and *fact…in* are scored separately. Those pairings with association scores that satisfy the threshold are considered hits and are highlighted in the results display.[1] To this point, this process resembles closely the extraction of collocations. Since chunks are not limited to word pairs, however, we iterate this process, and all (ordered) pairings with the target word (*e.g.*, *X fact* or *fact X*) serve as the input to the next iteration. We therefore refer to these inputs as bigrams. Word association measures look at the strength of association between two events x and y. In traditional collocation detection, as in our first iteration, x and y are individual words. At the second iteration, however, we use the same measures but change the definition of the events x and y. Here x is a word pairing (or bigram) from the first iteration (rather than just a word) and y is a word. This iteration then is measuring the association between each bigram (or a word pair) on the one hand and each word appearing within the five-word span of that bigram on the other. The output of this next iteration would consist of a set of ordered trigrams (potentially discontinuous due to our five-word window). Those trigrams that achieve the threshold word association score at this iteration are considered hits and highlighted in the results representation. All trigrams from this iteration in turn serve as input to the next iteration, and so on, in greedy fashion until no candidate strings meet the threshold association score. A hit for this algorithm, then, is a string of any number of words that results from satisfying the threshold score at the last iteration of any number of iterations. The system then creates a link from each such string to a display of all of the BNC sentences that instantiate that string.

Our searches provide two levels of results, which we will refer to as string types and string tokens. An example of a string type would be the string *point of view* whereas string tokens would be specific attested instances in the corpus that instantiate this string type. One reason this distinction becomes important is that our algorithm allows a five-word span for co-occurrence at any iteration. Thus, the string type *point of view* includes not only tokens where the three words are contiguous (*e.g.*, *According to her point of view…*), but also tokens where the same three words are non-contiguous (*e.g.*, *The point of mentioning this view is to…*).

There is no way of us predicting a priori which of these patterns within the tokens of one string type constitutes a true positive, so we cast our net wide in this way for the sake of recall at the cost of introducing noise. In the case of *point of view*, for example, it is the completely

[1] The results display distinguishes three conditions: word pairs that meet the conditional probability threshold only are highlighted in red, those meeting the MI threshold only are displayed in blue, and those meeting both in purple. This makes it easy to compare the performance of the two measures or the value of using both.

contiguous string tokens that are the true positives while the non-contiguous example (*the point of mentioning this view…*) illustrates a false positive. There are cases, however, where the reverse holds, that is, where discontinuity is a true property of a particular chunk. For example, our algorithm detects the longer string type *from point of view* as well. Notice, however, that for this case, in the true positives, *from* and *point* are non-adjacent (*from a logical point of view*; *from their point of view*, etc).

Thus, while our use of a five-word window at each iteration is motivated, it also introduces substantial noise in the results. To address the noise, we apply to these results some sorting mechanisms intended to help distinguish noise from the true positives. While this sorting falls short of actually affecting the precision of the results, it is aimed at organizing the results into groups of patterns for the hand inspection of the lexicographers at the user end.

We illustrate our approach to the sorting with a specific string type: *tell the time*. This string token is extracted by taking the noun *time* as the query target. At the second iteration, the trigram *tell the time* is detected as satisfying both the conditional probability and the MI thresholds. Moreover, 54 tokens of this type are found in the 20-million-word portion of BNC that we use. While the string type *tell the time* would appear at first glance to be a true positive, a look at the 54 tokens shows the noise. Our original version of the system left these 54 tokens unsorted and displayed them in the order they were detected in BNC. A sample of these 54 tokens is provided here:

She told herself sternly that the time had passed when sympathy
You tell me lies all the time
and one half could not tell the time or correctly select a medicine bottle
He found messages telling him that the time was not ripe
observation of the sun was a useful way of telling the time
I'm telling you for the last time, Harvey
I was told at the time that this system had been adopted because…
No mechanical indicator can tell you the right time to strike
…activities of daily living (such as counting money, telling the time, reading…)
observation of the sun was a useful way of telling the time

Our current version of the algorithm adds two stages of sorting to these 54 examples. The first step sorts the examples according to the patterns of contiguity of the string members. Thus, at this stage, all tokens where the string members (*tell*, *the*, *time*) are contiguous (*tell the time*) are grouped together, then all tokens where the first and second word (*tell* and *the*) are separated by one word are grouped together (*tell him the time*; *tell of the time*), and those with two words intervening there (tell him that the time), and so on. This stage of sorting, thus, is sensitive to the existence and location of any gaps separating the words in the string type and to how many words appear in those intervening gaps. These patterns are displayed in order of frequency. Thus, in the case of *tell the time*, the most frequent pattern is the one with no intervening words, such as *So what if you can't tell the time?* These comprise 16 of the 54 tokens. The second most common is the pattern where three words separate *tell* from *the time*, as in *She told herself sternly that the time has passed*. There are 8 tokens of this pattern. A look at these 8 tokens, however, will illustrate the motivation for adding a second stage to this sorting. Specifically, there is no coherent pattern shared by these 8 tokens. The fact that *tell* and *the time* are separated by three words in all eight examples follows from nothing interesting.

1. The court was told that teenagers were made to suck dummies and wear nappies , were bathed like babies and told to regress to the time they were last happy.

2. But , I told myself , by the time you are standing at the airport terminal (not the train or the bus station) , you have burned off the top ones and , come on , lad , you can afford to relax a little.
3. She told herself sternly that the time had passed when sympathy , hope , tender care , even love could have anything to do with the figure on the bed.
4. You tell me lies all the time !
5. Once the commotion had died down , he told them to break the time pencils and get to work.
6. I told her I thought The Times would probably have a man on the spot and it was late , and I prised my Toshiba away from her grasping hands.
7. Junior accountants at Price Waterhouse , one of the big six firms , have been told that now is the time to take their once-in-a-lifetime world tour or perhaps a summer stint as a yacht deckhand and that applications for extended unpaid leave bquo will be looked on favourably.
8. Carter had told the police at the time they had a row over money and his wife had walked out , never to return.

While this group is the second most common pattern for *tell the time* string type, it is a false pattern, one that represents no interesting regularity. In this case, then, the first stage of sorting does not contribute directly to chunk detection. Part of the motivation of the second stage is to bring us closer to detecting such results as noise automatically, eventually enabling us to not only sort examples but also filter out noise. To help achieve this, the second stage examines the POS of the intervening words in the string tokens within a pattern. The POS tags are from the CLAWS tagset used to tag BNC[2]. Listed here are the eight different POS trigrams representing the three words separating *tell* from *the time* in this group.

1.	TO0_VVI_PRP	3. PNX_AV0_CJT	5. PNP_TO0_VVI	7. CJT-DT0_AV0_VBZ
2.	PNX_PUN_PRP	4. PNP_VVZ_DT0	6. PNP_PNP_VVD	8. AT0_NN2_PRP

The fact that the eight sentences, while all sharing a 3-word gap between *tell* and *the time*, each has a different POS trigram in that gap can serve as readily detectable indication that this group of eight tokens does not reflect an interesting regularity concerning *tell the time*. Eventually our goal is to represent results to learners and teachers, not only to lexicographers. For this reason, such strategies for automatically detecting false positives and increasing precision will be important to ensure such examples are not presented to learners as instances of the chunk tell the time. In what follows, we describe our approach to representing our lexical chunk knowledge for learners and teachers.

4. Lexical Chunks and Knowledge Representation

A second challenge for lexicography concerning chunks is how to represent them to learners. We propose embedding the representation of chunk knowledge within the contexts where learners encounter chunks. As learners encounter the target language in contexts of authentic use, they are exposed to chunks (whether they recognize them as such or not). Our aim is to identify these chunks in real time directly within digital contexts. We illustrate how we have implemented this approach with collocations and describe the challenges in extending this approach to chunks. Our collocation tool (called Collocator) is a browser-based tool that can

[2] http://www.natcorp.ox.ac.uk/what/garside_allc.html

detect collocations in real time within the web pages that the user browses (Wible *et al.*, 2004a).

The core component of Collocator that detects collocations in web pages in real time works much like the first iteration of our chunk detecting algorithm. The collocation-extracting scheme is part-of-speech sensitive, which means we have to detect the part-of-speech information of each word in browsed web pages in real time. We train a Markov Model-based POS tagger (Brants, 2000) and use British National Corpus (BNC)[3] as our training data. The internal evaluation shows this tagger has 93 % precision including identifying unknown words. After part-of-speech tagging, the agent uses the following equation from (Wible *et al.*, 2004b) to measure the word association score for all candidate word pairs:

$$normMI(x, y) = \log_2 \frac{P(x, y)}{\left(P(x)/sn(x)\right) \cdot \left(P(y)/sn(y)\right)}$$

where x, y mean the word with specific part-of-speech and *sn* means the number of distinct senses for that word listed in WordNet. This adaptation of traditional MI takes into account the polysemy of the words x and y by normalizing for the number of senses of x and y, helping overcome traditional MI's under-extraction of collocations that contain high frequency words. For example, traditional MI does not detect *take* as a collocate of the noun *temperature* (*The nurse took the patient's temperature*), but our normalized MI does detect *take* in this case.

Similar to the first iteration in our chunk detector, Collocator takes as collocation candidates all possible pairings of POS-specific words (x and y above) in which the two words appear within a five-word window of each other in our 20 million-words of running text of the BNC. Using the above measure, each x,y ordered pair yields a word association score. Collocations are word pairs that show a sufficiently strong word association between the two words in the pair. Thus, a minimum score threshold is used to select which of the candidate word pairs constitute collocations. This threshold can be lowered or raised to adjust the agent's precision and recall. The collocation knowledge thus extracted from our POS-tagged BNC feeds our browser-based Collocator, enabling it to detect and highlight collocations that appear in the web pages that the user browses (See figure 1).

Figure 1. Toolbar's dropdown menu with detected collocations and link to examples

[3] http://www.natcorp.ox.ac.uk/

Using Collocator as our reference point, we can now describe how our lexical chunk extraction is to be applied to learners. In brief, our purpose is to enrich Collocator so that it detects and represents not only collocations (typically, word pairs) but longer strings, that is, chunks. This tool will then detect and highlight chunks in real time within the web pages that the user browses. The approach to lexical chunk discovery described above is intended, then, as the knowledge source supporting this tool. The fundamental challenge this application poses is that of precision.

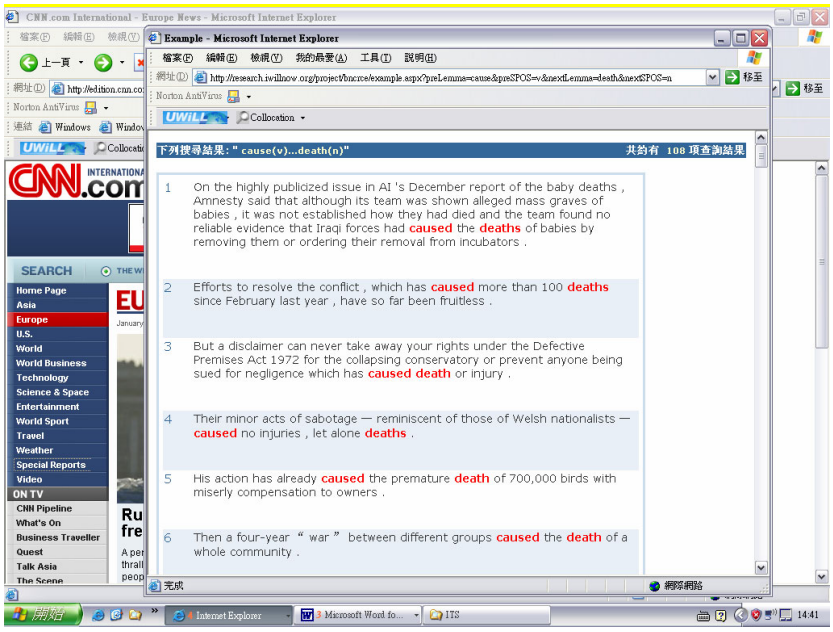


Figure 2. Example sentences

As shown above, each iteration of the chunking algorithm introduces noise and leads to more false positives with each iteration. A ubiquitous browser-based version of the chunker would detect strings in the web page that match any string types that have achieved a threshold word association measure at our 'discovery' stage. Recalling the example of *tell the time*, this means that a browser-based chunk detector would treat as a match not only the true positive *He hasn't learned to tell the time*, but also false positives like *Tell him that unfortunately the time has not arrived*. These failures in precision, while forgivable for a tool intended for lexicographers, are unacceptable for a tool intended directly for learners as they browse the Web. We choose to retain the higher recall created by our 5-word window and focus on increasing precision by adding a second stage that sorts these results into patterns, as described above. While our current sorting method does make it much easier to filter out false positives by hand, this is not improvement enough for a browser-based version for learners. The sorting strategy that we mentioned early of exploiting POS patterns, though requiring refinements, does hold promise. For example, it would enable the chunker to discard false positives like *Tell him that unfortunately the time has not arrived* since, as we have seen, it can discover that cases like this with a 3-word gap between *tell* and *the time* exhibit no regularities in the POS sequences appearing in that gap. However, the implementation of this strategy must be much more fully articulated and tested. Recently, we have added entropy measures to detect how stable the POS patterns are that occur within the strings. Preliminary results suggest high entropy as a promising indicator of false positives, allowing for automatic filtering for increased precision of our results.

References

BRANTS T. (2000). "TnT-A statistical part-of-speech tagger". In *Processings of ANLP-2000*. Seattle, Washington.

CHURCH K, GALE W., HANKS P., HINDLE D. (1991). "Using statistics in lexical analysis". In U. Zernik (ed.), *Lexical Acquisition: Exploiting Online Resources to Build a Lexicon*. Lawrence Erlbaum Associates, Hillsdale: 115-164.

SINCLAIR J. (ed.) (1987). *Looking UP: An Account of the COBUILD Project in Lexical Computing*. Collins, London.

SMADJA F. (1993). "Retriving Collocations from Text: Xtract". In C*omputational Linguistics* 19: 143-177.

WIBLE D., KUO C.-H., TSAO N.-L. (2004a). "Contextualizing Language Learning in the Digital Wild: Tools and a Framework". In *Proceedings of IEEE International Conference on Advanced Learning Technologies (ICALT)*. Joensuu.

WIBLE D., KUO C.-H., TSAO N.-L. (2004b). "Improving the Extraction of Collocations with High Frequency Words". In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*. Lisbon.

WEINERT R. (1995). "The Role of Formulaic Language in Second Language Acquisition: A Review". In *Applied Linguistics* 16: 180-205.

WRAY A. (2002). *Formulaic Language in the Lexicon*. Cambridge University Press, Cambridge.

Appendix 2: Kaleidoscope presentation

Traitement automatique des langues et Apprentissage des langues assisté par ordinateur: perspectives d'intégration

Integrating Natural Language Processing and Computer-Assisted Language Learning

Kaleidoscope http://www.noe-kaleidoscope.org

- a Network of Excellence in Technololgy Enhanced learning (TEL)
- Aim: coordinate research in TEL across Europe
- 23 countries, 76 research units, 800 researchers
- Academic + Private research
- Supported by the European Community,
  under the Information Society Technologies priority
  of the 6th Framework Programme

Kaleidoscope actions

- Backbone Activities
  - Academia-Industry Digital Alliance
  - Advanced Training Activities
  - Shared Virtual Laboratory
  - Virtual Doctoral School
  - Users Group

- Special Interest Groups
- European Research Teams
- Jointly-Executed Integrated Research Projects (JEIRP)

JEIRP: Integrated Digital Language Learning

- Aim: design a model of NLP that integrates Natural Language Processing, Computer Corpus Research and Language Didactics
- Partners:
  - Université catholique de Louvain (coordinator)
    - Centre for English Corpus Linguistics (S. Granger)
    - Centre de traitement automatique du langage (C. Fairon)
  - Université Stendhal
    - Laboratoire de linguistique et de didactique des langues étrangères et maternelle (G. Antoniadis)
    - Centre universitaire d'études françaises (C. Cavalla)
  - York University
    - Department of French as a Second Language (A. Avolonto)
  - Tamkang University
    - Ubiquitous Web-based Intelligent Language Learning (D. Wible)

Three case studies

- NLP-enhanced ESP terminology (medical English)
  - Integrate an intelligent glossary into a learning management system (Moodle) in order to facilitate reading comprehension and vocabulary acquisition
  - Louvain (& Tamkang)
- Learner-corpus-enhanced error detection and feedback
  - Use error-tagged learner corpora of L2 French to detect errors and provide intelligent feedback
  - Grenoble (& Louvain); testbed: CUEF & York
- NLP-enhanced collocation detection and didactic exploitation
  - Use NLP tools to detect collocations and integrate them into didactic materialsTamkang (& Louvain)

Two dissemination actions

- Integrated Digital Language Learning webpage

URL: http://www.idill.org

- Integrated Digital Language Learning doctoral module

Idill webpage

- URL: http://www.idill.org
- Focus on integration of NLP into digital language learning
- Contents: select bibliography, links to relevant websites, forthcoming events, repository of papers and theses
- Audience: students/researchers in computer science, linguistics and didactics

Idill virtual doctoral module

- Virtual module on Digital Language Learning to be integrated into doctoral programmes
- Three sections
  - Introduction to NLP
  - NLP and CALL
  - Learner corpora and (I)CALL
- Two languages
  - English
  - French
- Deadline: December 2006

Interested in contributing to the Idill webpage and/or doctoral module?

- Contact Julia Medori
  - medori@tedm.ucl.ac.be
- or any of the “Idillic” team members!
  - granger@lige.ucl.ac.be
  - cedrick.fairon@uclouvain.be
  - Georges.Antoniadis@u-grenoble3.fr