Evaluation automatique de productions lexicales : une analyse à 4 niveaux

Olivier Kraif

▶ **To cite this version:**

Olivier Kraif. Evaluation automatique de productions lexicales : une analyse à 4 niveaux. UNTELE 2005, 2005, Compiègne, France. hal-00790341

HAL Id: hal-00790341
https://hal.science/hal-00790341

Submitted on 19 Feb 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Evaluation automatique de productions lexicales : une analyse à 4 niveaux

Olivier Kraif (Olivier.Kraif@u-grenoble3.fr)
Université Stendhal - Grenoble 3

Résumé : De nombreuses activités pour l'apprentissage des langues incluent des tests destinés à l'autoévaluation de l'apprenant. Dans bien des cas, ces tests s'appuient sur des questions fermées du type QCM. Lorsque les questions sont semi-ouvertes, comme les quiz ou les tests de closure, les réponses font souvent l'objet d'un traitement binaire : ces tests sont traités comme des questions fermées, ignorant la variabilité des réponses de l'apprenant.
Pour pallier ce manque, nous proposons de recourir à des techniques simples et maîtrisées issues du traitement automatique des langues : normalisation, étiquetage morphosyntaxique, lemmatisation et consultation de lexique sémantiques (type Wordnet). Afin de minimiser l'effet de certaines ambiguïtés (morphosyntaxiques et sémantiques) nous proposons une heuristique de moindre différence, où la réponse de l'apprenant joue le rôle d'indice permettant d'orienter, par similarité, l'analyse linguistique. Nous avons implanté ces principes dans un prototype effectuant une évaluation automatique à quatre niveaux, afin de distinguer différents types de différences : différences superficielles, différences orthographiques, différences morphosyntaxiques, différences lexicosémantiques.
Ce prototype entre actuellement en phase de test : par la mise en ligne d'exercices lacunaires, nous prévoyons de recueillir un corpus de réponses d'apprenants, qui nous permettra d'évaluer la validité de notre approche et d'affiner les feed-back de notre système.

1 Introduction

Quand on examine la plupart des produits pédagogiques développés en Apprentissage des langues assisté par ordinateur (ALAO) force est de constater que le traitement automatique des réponses et productions de l'apprenant est le plus souvent réduit au strict minimum : quand l'apprenant répond à des questions fermées, de type *questionnaire à choix multiple*, l'évaluation est triviale et facilement automatisable ; à l'opposé, quand il s'agit de productions libres, on laisse le soin à un enseignant d'effectuer la correction lui-même. Entre ces deux extrêmes, les activités de type *Question à réponses ouvertes courtes* (QROC), telles que des exercices lacunaires, suivent en général l'un ou l'autre des deux paradigmes : soit on considère qu'il existe une (ou quelques) bonne(s) réponse(s) prédéfinie(s), tout le reste étant considéré comme faux, et l'on peut procéder à une analyse automatique ; soit l'évaluation est faite manuellement, par correction de l'enseignant ou auto-correction en donnant *la* réponse jugée correcte.

S'il est toujours possible d'affiner ce genre d'évaluation en essayant de *prévoir* plusieurs bonnes réponses possibles, cette "prédiction" comporte deux écueils. D'une part, elle complique la tâche du concepteur de l'activité qui doit anticiper des possibilités parfois nombreuses. D'autre part, rien ne garantit que la "prédiction" soit couronnée de succès. Pour une évaluation plus fine, il serait intéressant de prévoir également les *erreurs*, afin d'être en mesure de fournir un diagnostique en guise de feed-back, plus intéressant qu'une simple évaluation. Mais l'étendue des possibilités à intégrer *a priori* deviendrait rédhibitoire. Habituellement, on ne considère qu'un seul type de réponse fausse, sans hiérarchisation ni distinction quant au degré de l'erreur.

En d'autres termes, pour les CROQ, l'évaluation automatique se cantonne à une logique du tout ou rien. On sait cependant que toutes sortes de divergences peuvent intervenir quand on compare la réponse donnée avec la (ou les) réponses attendu(es) :

- Variations graphiques (majuscules, espacements)
- Fautes d'orthographe
- Fautes de grammaire
- Problèmes de registre
- Choix d'expressions synonymes
- Faux-sens, contre-sens,
- ....

Ces variations sont difficilement *prédictibles* (p. ex. en créant une liste prédéfinie) mais elles sont *calculables* : c'est là tout l'objet des techniques du *Traitement automatique de la langue* (TAL). L'idée d'intégrer ces techniques dans des applications de l'ALAO n'est pas neuve (Chanier & Selva, 2000, Brun et al. 2002, Antoniadis et al. 2004, Kraif, 2003), mais on note pourtant qu'elles sont sous-employées en ce qui concerne l'évaluation. Ce sous-emploi est imputable aux deux raisons suivantes :

- Les analyses automatiques produites par le TAL sont souvent peu fiables. Tout utilisateur de logiciels "grand public" tels que le correcteur de grammaire sous Word ou Correcteur 101 didactique a pu faire le constat que de nombreuses erreurs restent non détectées, tandis que des constructions correctes sont parfois signalées comme erronées.
- Les techniques du TAL sont difficiles à mettre en oeuvre : outre un haut degré de technicité dans ses traitements (analyse lexicale, morphologique, syntaxique, sémantique, etc.), elles requièrent des ressources coûteuses (dictionnaires électroniques, grammaires formelles, réseaux sémantiques, corpus électroniques, etc.) pour leur implantation.

Nous ne contestons ni l'une ni l'autre de ces deux assertions : nous pensons cependant qu'il existe, au niveau de l'état de l'art, un corpus de techniques élémentaires, simples dans leur mise en oeuvre, qui permettent d'effectuer un véritable saut qualitatif dans l'évaluation des productions courtes. Afin d'illustrer ce point de vue, et de proposer des solutions concrètes au problème de l'évaluation, nous présentons dans cet article une méthode basée sur une analyse à 4 niveaux, permettant à la fois la prise en compte de plusieurs réponses correctes et le diagnostique de différents types d'erreur.

2 Principe de l'analyse et heuristique

La méthode d'analyse que nous décrivons s'intéresse à un type de traitement limité, qui se révèle adapté dans le cadre des QROC : il s'agit simplement de comparer deux lexèmes, que nous noterons respectivement RD (réponse donnée) et RA (réponse attendue).

L'analyse linguistique de deux lexèmes peut se heurter à différents types d'ambiguïtés :

- Ambiguïtés morphosyntaxiques dues à l'homographie.
- Ambiguïtés sémantiques dues à la polysémie.

Par exemple, supposons qu'on ait RA=*voiture* et RD=*véhicule*, dans le contexte "Je n'ai plus de ....". La réponse de l'apprenant peut être jugée correcte sur les plans orthographiques, morphologiques et sémantiques. Pour reconnaître cette similarité, l'analyse linguistique doit cependant éliminer des interprétations possibles : p. ex. *véhicule* comme forme verbale, à la première personne, indicatif présent, du verbe *véhiculer*. L'ambiguïté touche aussi RA : si on retient, pour *voiture*, l'acception précise des chemins de fer, c'est-à-dire */wagon pour le transport des voyageurs/*, RA et RD apparaîtrons plus éloignés sémantiquement que si on retient /*automobile*/. Bref, pour décider de l'identité relative, ou de la différence, de RA et RD, il faudra faire des choix entre des analyses concurrentes.

On sait que la résolution des ambiguïtés peut nécessiter des traitements complexes, tels que l'application de règles morphosyntaxiques ou de modèles probabilistes (arbres de décision, chaînes de Markov,....). Pour traiter de façon économique une partie de ces ambiguïtés, nous proposons d'utiliser l'heuristique suivante :

"Toute ressemblance entre RD et RA sera présumée non-fortuite, résultant du choix de l'apprenant. L'analyse doit donc être guidée par les ressemblances."

Cette heuristique de moindre différence se base en quelque sorte sur une présomption d'exactitude (relative) de la réponse. L'analyse pourra donc s'appuyer en priorité sur les ressemblances les plus significatives, c'est-à-dire les moins susceptibles d'être dues au hasard :

1. Identité des graphies après normalisation
RA= *circuit* vs RD=*Circuit*

2. Ressemblance graphique
RA=*circuit* vs RD=*circui*

3. Ressemblance morphosyntaxique
RA=*circuit* (nom) vs RD=*voyage* (nom) et pas *voyage* (verbe)

4. Ressemblance sémantique
RA=*circuit* vs RD=*tour* (/tourisme/) et non *tour* (/bâtiment/)

La comparaison des traits morphosyntaxiques prend appui sur le(s) lemme(s) (forme(s) canonique(s)) potentiel(s), ainsi que les traits correspondant aux différentes analyses morphosyntaxiques. On peut se baser :

- Sur un dictionnaire de forme fléchies, comportant des entrées comme ci-dessous :

Forme	Lemme	Analyses
porte	*porte*	Adj:InvGen+SG
porte	*porte*	Nom:Fem+SG
porte	*porter*	Ver:IPre+SG+P1:IPre+SG+P3: SPre+SG+P1:SPre+SG+P3: ImPre+SG+P2

- Sur les sorties d'un étiqueteur/lemmatiseur (avec élimination des ambiguïtés).

Pour évaluer la portée de l'heuristique de moindre différence, nous n'avons utilisé, dans la version actuelle, que les entrées d'un dictionnaire de formes fléchies (dictionnaire libre disponible à http://abu.cnam.fr/DICO/).

3 Description des 4 niveaux

Niveau 1: normalisation graphique

Le premier niveau concerne les différences superficielles, purement graphiques : présence de blancs inutiles, différences dans la casse (majuscules/minuscules), variantes de caractère (*oeuvre* vs *œuvre*, *Etre* vs *Être*), variantes orthographiques (*événement* vs *évènement*, *plate-forme* vs *plateforme*, *paraît* vs *parait*, *aiguës* vs *aigües*, etc.). Pour

reconnaître l'identité de deux chaînes malgré ces différences, il faut utiliser une fonction de *normalisation graphique*, classique en TAL, simple à réaliser dans un langage tel que Perl. La normalisation peut aussi aller jusqu'à la suppression de mots outils (p. ex. articles, pronoms), facultatifs dans certaines réponses :

Par exemple, pour la question :
Question.: Quel est le fruit du pin ?
(RA : La pomme de pin)

La réponse suivante pourra donner une évaluation favorable :
RD : Pomme de Pin -> "OK"

Si une différence subsiste malgré la normalisation, on passe au niveau 2.

Niveau 2 : différence orthographique

Plusieurs cas de figure sont envisagés :

2.1 - Si la forme est connue du dictionnaire, graphiquement ressemblante, mais avec un lemme différent : on a sans doute une confusion orthographique (par homophonie ?). Par exemple :
RD : pomme de pain -> "Attention à la confusion entre pain et pin"

2.2 - Si la forme est inconnue et ressemblante : on a sans doute une faute d'orthographe. Par exemple :
RD : pome de pain -> "Vérifiez l'orthographe"

2.3 - Si un des lemmes potentiels est identique à un des lemmes de RA, examiner le niveau 3.
RD : pommes de pin (-> niveau 3)

Notons qu'ici l'heuristique de moindre différence permet de retenir le lemme *pomme* (nom) et non *pommer* (verbe).

2.4 - Si la forme est inconnue et dissemblable, proposer des formes ressemblantes, puis examiner le niveau 4 (sémantique).
RD : cone -> cône (-> niveau 4)

2.5 - Si la forme est connue mais dissemblable, examiner le niveau 4 (sémantique).
RD : pigne (-> niveau 4)

Niveau 3 - Différence morphosyntaxique

Ce niveau traite les cas où l'on a trouvé un lemme identique mais avec une analyse morphosyntaxique différente. Par exemple :
RA= pomme (Fem, SG) vs RD= pommes (Fem, PL) -> "vérifier le nombre"

Là encore, un dictionnaire de formes fléchies peut suffire. Il faut néanmoins utiliser l'heuristique pour la comparaison des traits. Par exemple :

RA : si j'avais su Lemme: avoir	Ver:IImp+SG+P1	ou	IImp+SG+P2
RD : si j'aurais su Lemme: avoir	Ver:CPre+SG+P1	ou	CPre+SG+P2

Ici, on observe des différences pour le temps/mode ainsi que pour la personne :
P1 ≠ P2, CPre ≠ IImp

Grâce à l'heuristique de moindre différence, on ne compare que les analyses les plus proches (entre pointillés), ce qui permet de ne retenir qu'une seule différence, entre le conditionnel présent et l'imparfait : CPre ≠ IImp

On pourra donc renvoyer un feed-back du type"Vérifiez le temps et le mode pour le verbe *avoir*". Voici un autre exemple comportant des ambiguïtés :

RA : les auxiliaires
auxiliaire, Adj:InvG+PL, Nom:InvG+PL, Nom:Mas+PL
auxiliaires, Nom:Mas+Pl
RD : les auxiliaire
auxiliaire, Adj:InvG+SG, Nom:InvG+SG, Nom:Mas+SG

En appliquant l'heuristique, on ne retient que la différence de nombre, et non celle de catégorie ou de genre.

Niveau 4 - différence lexicosémantique

Ce niveau traite les cas où tous les lemmes potentiels de RD sont différents de ceux de RA. Il peut également se combiner avec le niveau 3 pour mettre en évidence des différences morphosyntaxiques. On compare deux à deux tous les lemmes potentiels de même catégorie. S'il existe une relation sémantique étroite entre les deux lemmes, de type synonymie, hyperonymie, hyponymie ou antonymie, on considère que cette relation n'est pas le fruit du hasard. On peut alors élaborer un feed-back approprié en fonction du type de similarité :

RD : *fruit* (hyperonyme) -> "Soyez plus précis."
RD : *cône* (synonyme) -> "Donnez une expression synonyme."
RD : *pignon* (meronyme) -> "Le *pignon* est une partie de ce qu'on cherche."

A nouveau, on est confronté à de nombreuses ambiguïtés sémantiques, car du fait de la polysémie les lemmes sont susceptibles de porter différents sens. Mais l'heuristique de moindre différence permet de contourner cet écueil, car on ne retient que les comparaisons entre éléments similaires : si l'apprenant répond *cône* ou *pignon* pour *pomme de pin*, il est peu probable qu'il se référait aux acceptions géométrique ou mécanique de ces termes (ce qui rendrait sa réponse invalide). Si l'on dispose d'une ressource sémantique de type EuroWordNet, on pourra se baser sur le "chemin" le plus court entre deux unités pour sélectionner le sens présumé.

RD=poire --hyper--> *fruit* --hypo--> *RA=pomme de pin*

Certes, un terme appartenant à un autre champ sémantique serait évalué positivement. Par exemple:

RD=engrenage --hypo--> *pignon* --holo--> *RA=pomme de pin*

Mais il est peu probable qu'à la question "*Quel est le fruit du pin ?*" un apprenant réponde *engrenage*.

4 Mise en oeuvre et perspectives

L'implantation des niveaux 1 à 3 font partie des techniques de base du TAL : elles sont fiables et font appel à des ressources peu coûteuses : normalisation, tokenisation, dictionnaire de formes fléchies, lemmatiseur et étiqueteur morphosyntaxique. Le niveau 4 dépend de la richesse de la ressource lexicale employée (type Wordnet).

La version actuelle de notre prototype est basée sur le dictionnaire librement distribué par l'ABU (*Association des Bibliophiles Universels*) et le réseau français d'EuroWordNet. Nous l'avons testée, avec succès, sur un échantillon extrait d'un corpus de productions d'apprenant rassemblé par (Echinard, 2004), mais il s'agissait de formes erronées tirées de productions libres (par ex. RA=*sortait* vs RD=sorta*i*s, RA=*bijoux* vs RD=*bijous*, RA=*cadeaux* vs RD=*cadeaus*, etc.). Ces fautes d'accord ou d'orthographe ne rendent pas compte de toutes les divergences susceptibles d'apparaître dans un QROC, par exemple un exercice lacunaire. Pour effectuer une évaluation véritable de notre méthode, nous prévoyons de mettre en ligne des exercices lacunaires et d'enregistrer un corpus spécifique de productions d'apprenants, dans le cadre du projet Mirto (Antoniadis & Ponton, 2004). Ce corpus nous permettra entre autre de :

- Tester l'architecture à 4 niveaux.
- Mesurer l'efficacité de l'heuristique de moindre différence (avec dictionnaire ou étiqueteur).
- Affiner les feed-back pour chaque cas détecté.

Dans l'avenir, deux améliorations pourront être apportées :

- Prise en compte des unités polylexicales.

Pour le moment nous ne comparons que des formes simples. La comparaison d'expressions polylexicales (mot composé, expression verbale avec auxiliaire, expression idiomatique, etc.) pourra être effectuée simplement sur la base d'une comparaison forme par forme.

- Prise en compte d'erreurs dues à une confusion de paradigme.

Quand un apprenant écrit *bijous* au lieu de *bijoux*, ou *prendez* au lieu de *prenez*, il ne s'agit évidemment pas d'une simple faute d'orthographe. Ces erreurs fréquentes résultent de la mauvaise application d'une règle morphologique par ailleurs correcte : par exemple, le paradigme de *rendre* appliqué au verbe *prendre*. Or ces erreurs peuvent être automatiquement diagnostiquées. Par exemple, on peut générer un dictionnaire de formes erronées contenant des formes artificiellement fléchies en leur appliquant des paradigmes majoritaires (*cheval* -> **chevals*) ou voisins (*rendez* -> **prendez*). Pour ce faire, il faut préalablement coder les modèles morphologiques les plus fréquents dans la langue.

5 Conclusion

L'application des techniques du traitement automatique de la langue à l'ALAO n'en est qu'à son début. Concernant l'évaluation des réponses d'apprenant, le manque de fiabilité des techniques les plus avancées du TAL, telles que l'analyse syntaxique ou la compréhension automatique (extraction d'une représentation sémantique), apparaît parfois comme rédhibitoire. Des projets ambitieux, comme le projet FreeText (L'haire & Vandeventer, 2003), ouvrent des perspectives intéressantes, mais les applications tardent à venir du fait de la lourdeur de mise en oeuvre des technologies concernées.

Pourtant, dans la perspective plus modeste de l'évaluation de réponses courtes, les solutions existent, et n'impliquent pas d'investissement technique important. Avec la description d'un analyseur de réponse permettant de distinguer différents types d'erreurs et d'accepter différents types de bonnes réponses, nous espérons avoir défriché quelques pistes, que la communauté des utilisateurs non-spécialistes du TAL (enseignants et concepteurs de produits pédagogiques) pourra aisément emprunter.

La prochaine étape de ce travail consistera à mener une évaluation précise de la méthode à partir d'un corpus de productions d'apprenant, recueillies dans le contexte spécifique des activités visées (de type QROC), afin de valider notre approche et de l'adapter aux cas de figure rencontrées.

6 Références

Antoniadis, G., Ponton, C. (2004) MIRTO : Un système au service de l'enseignement des langues. *UNTELE'2004*, 6-20 mars 2004, Compiègne.

Antoniadis, G., Echinard, S., Kraif, O., Lebarbé, T., 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, Geneva, August 28th, 2004.

Brun, C., Parmentier, T., Sandor, A., Segond, F. (2002) Tal et e-formation en langues, in F. Segond (dir.) *Multilinguisme et traitement de l'information*, Hermès, Paris, p. 223-250.

Chanier, T., Selva, T. (2000) Génération automatique d'activités Lexicales dans le système ALEXIA, *Sciences et Techniques Educatives (STE*), vol 7, 2. Hermès, Paris, p 385-412.

Echinard, S. (2004) *Vers la création d'un système d'analyse des réponses d'apprenants*, Mémoire de DEA, Sous la dir. Georges Antoniadis et Claude Ponton, Université Stendhal Grenoble 3.

Kraif, O. (2003) Propositions pour l'intégration d'outils TAL aux dispositifs informatisés d'apprentissage des langues, Intercompréhension en langues romanes, LIDIL, N° 28, C. Degache (sous la dir. de), Grenoble.

L'haire, S., Vandeventer Faltin, A. (2003) Diagnostic d'erreurs dans le projet FreeText, *ALSIC*, Vol. 6, numéro 2, décembre 2003, pp. 21 - 37