# Direkt Profil : un système d'évaluation de textes d'élèves de français langue étrangère fondé sur les itinéraires d'acquisition

Jonas Granfeldt (1), Pierre Nugues (2), Emil Persson (1), Lisa Persson (2), Fabian Kostadinov (3), Malin Ågren (1), Suzanne Schlyter (1)

(1) Institut des langues romanes – Université de Lund
Box 201, S-221 00 Lund, Suède
{Jonas.Granfeldt, Malin.Agren, Suzanne.Schlyter}@rom.lu.se,
emil.person@telia.com
(2) Institut d'informatique – Université de Lund
Box 118, S-221 00 Lund, Suède
Pierre.Nugues@cs.lth.se, nossrespasil@hotmail.com
(3) Institut d'informatique – Université de Zurich
fabian.kostadinov@access.unizh.ch

**Mots-clés :** français langue étrangère, itinéraires d'acquisition, évaluation, annotation, analyse syntaxique partielle

**Keywords:** second language French, developmental sequences, evaluation, annotation, partial parsing

**Résumé :** *Direkt Profil* est un analyseur automatique de textes écrits en français comme langue étrangère. Son but est de produire une évaluation du stade de langue des élèves sous la forme d'un profil d'apprenant. *Direkt Profil* réalise une analyse des phrases fondée sur des *itinéraires d'acquisition*, i.e. des phénomènes morphosyntaxiques locaux liés à un développement dans l'apprentissage du français. L'article présente les corpus que nous traitons et d'une façon sommaire les *itinéraires d'acquisition*. Il décrit ensuite l'annotation que nous avons définie, le moteur d'analyse syntaxique et l'interface utilisateur. Nous concluons par les résultats obtenus jusqu'ici : sur le corpus de test, le système obtient un rappel de 83% et une précision de 83%

**Abstract:** *Direkt Profil* is an automatic analyzer of texts written in French as a second language. The objective is to produce an evaluation of the development stage of the students under the form of a learner profile. *Direkt Profil* carries out a sentence analysis based on developmental sequences, i.e. local morphosyntactic phenomena linked to a development in the learning of French. The paper presents the corpus that we use and briefly, the developmental sequences. Furthermore, it describes the annotation that we have defined, the parser, and the user interface. We conclude by the results obtained so far: on the test corpus the systems obtains a recall of 83% and a precision of 83%.

# 1 Introduction

Les systèmes d'évaluation des compétences linguistiques et d'apprentissage des langues assisté par ordinateur (ALAO) ont peu recours aux techniques de traitement automatique des langues (TAL). Les applications commerciales existantes produisent des exercices dont la correction dépend de techniques de reconnaissance de forme. Ces techniques limitent non seulement la qualité et la nature du feedback, mais elles restreignent aussi les types d'activités possibles. Nous présentons ici un système réalisant une analyse automatique de textes produits librement. Il est fondé sur l'étude de l'acquisition des langues étrangères à l'âge adulte. Pour analyser les phrases, nous avons créé un schéma d'annotation des textes, construit un analyseur syntaxique et développé un ensemble de règles.

L'interface du programme est conçue pour que les enseignants ou les chercheurs puissent copier des textes écrits par des apprenants et les soumettre à l'analyseur. Le programme identifie les structures caractéristiques du développement grammatical du français et affiche les résultats à l'utilisateur. Le premier objectif de Direkt Profil est d'être un outil dans la recherche des stades de développement du français écrit en identifiant des telles structures. À plus long terme, le but du système est de produire une évaluation des textes d'élèves en français sous la forme d'un profil d'apprenant.

# 2 Le corpus CEFLE de Lund

Pour le développement et l'évaluation de notre système, nous avons utilisé le corpus CEFLE de Lund (« Corpus Écrit de Français Langue Étrangère de Lund »). Ce corpus contient environ 100 000 mots (Ågren, 2005). Les textes qui le composent sont des récits de longueur et de niveaux variés. Nous l'avons rassemblé en demandant à 85 lycéens suédois et à 22 jeunes Français de raconter par écrit, entre autres, l'histoire évoquée par les images de la Figure 1. Le but du système étant d'analyser le français langue étrangère, nous avons utilisé les textes des Français comme groupe de contrôle.

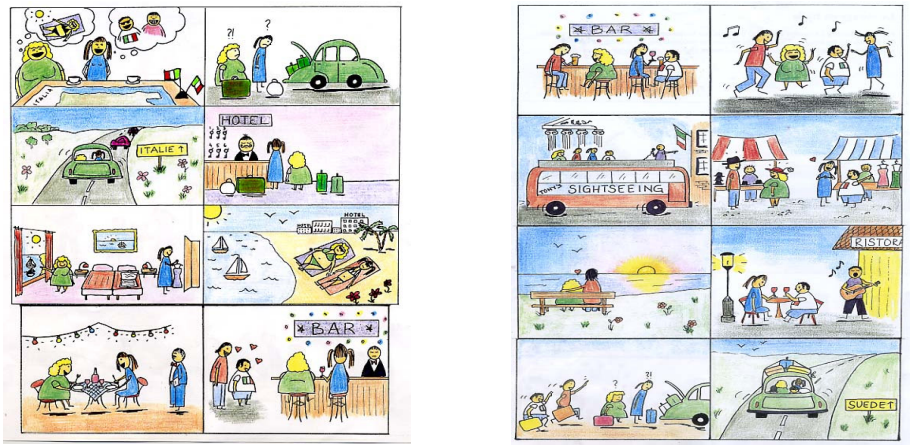


Figure 1 : *Voyage en Italie.*

Le récit qui suit est un exemple provenant d'une apprenante débutante :

> *Elles sont deux femmes. Elles sont a italie au une vacanse. Mais L'Auto est très petite. Elles va a Italie. Au l'hothel elles demande une chambre. Un homme a le clé. Le*

> *chambre est grande avec deux lies. Il fait chaud. C'est noir. Cette deux femmes est a une restaurang. Dans une bar cet deux hommes. Ils amour les femmes. Ils parlons dans la bar. Ils ont tres bien. Le homme et la femme participat a un sightseeing dans la Rome. Ils achetons une robe. La robe est verte. La femme et l'homme reste au un banqe. Ils c'est amour. La femme et l'homme est au une ristorante. es hommes va avec les femmes. L'auto est petite.*

Ce texte contient un certain nombre de constructions caractéristiques : parataxe, ordre des mots très simple, absence de pronoms objets, des formes verbales de base, fautes d'accord verbal et nominal, de genre, orthographe. Les travaux sur l'acquisition d'une langue étrangère ont montré que ces constructions (et d'autres) apparaissent avec une certaine systématicité selon le stade linguistique de l'apprenant. Ils permettent de décrire le développement de grammaires d'apprenants sous la forme d'itinéraires d'acquisition. Le corpus contient les textes bruts annotés avec leur stade de développement (Tableau 1).

## 3 *Direkt Profil et d'autres systèmes*

*Direkt Profil* est un analyseur de textes écrits en français comme langue étrangère. Il repose sur les constructions linguistiques des *itinéraires d'acquisition*. Nous avons établi une description systématique de ces itinéraires sous la forme d'une annotation et l'objectif du système est de les détecter automatiquement. L'analyseur parcourt le texte d'un apprenant en annotant et calculant les occurrences d'un phénomène particulier dans ses formes diverses. Le résultat est un profil de texte basé sur ces critères et, éventuellement, une indication du niveau du texte. L'interface présente les résultats à l'utilisateur en visualisant par des couleurs différentes les structures qu'il a détectées. Il est important de souligner que le système n'est pas un correcteur.

La plupart des outils informatiques relevant du domaine ont pour but d'aider à la rédaction. Ils identifient et parfois corrigent des fautes d'orthographe et des erreurs de grammaire. La lignée de programmes aboutissant à PLNLP (Jensen et al. 1993) et NLPWin (Heidorn 2000) est l'une des réalisations les plus notables. Le correcteur grammatical de PLNLP opère une analyse syntaxique complète. Il a été créé pour l'anglais et appliqué ensuite à d'autres langues dont le français. Il utilise des règles syntagmatiques binaires et prend en compte des relations de dépendance. PLNLP s'adresse essentiellement, mais non exclusivement, à des utilisateurs rédigeant dans leur langue maternelle.

D'autres systèmes relèvent de l'enseignement des langues assisté par ordinateur (ELAO) tels que *FreeText* (Granger et al., 2001) pour le français et *Granska* (Bigert et al. 2004) pour le suédois. *FreeText* se place dans une approche communicative à l'apprentissage des langues. Il utilise un analyseur syntaxique chomskyen pour le français. En cas d'échec, il opère à un relâchement de contraintes, par exemple sur les accords, pour diagnostiquer une erreur. *Granska*, à la différence de *FreeText*, réalise une analyse syntaxique partielle. Les auteurs justifient ce type d'analyse par une robustesse qu'ils jugent supérieure et qui permet d'accepter plus facilement des phrases incorrectes.

## 4 Une méthode d'analyse fondée sur les itinéraires d'acquisition du langage

Les systèmes actuels diffèrent en ce qui concerne le type d'analyse opérée : analyse complète ou partielle de la phrase. L'analyse complète et la correction d'erreurs sont difficilement

applicables aux textes d'apprenants de (très) bas niveau linguistique parce que le nombre de mots inconnus et de phrases incorrectes y sont très souvent élevés.

Dans notre corpus de test de 6 842 mots, la distribution des mots inconnus et de phrases incorrectes était la suivante. Au Stade 1, près de 100 % des phrases sont incorrectes (98,9 %) et 24,7 % des mots sont inconnus[1,2]. À ce stade, toute analyse complète des phrases nous semble très difficile. En revanche, dans le groupe de contrôle les chiffres correspondants sont de 32,7 % pour les phrases incorrectes et de 10,6% pour les mots inconnus. La Figure 2 montre aussi que la seule quantification des mesures de « mots inconnus » et « phrases incorrectes » est insuffisante pour définir le niveau linguistique des textes des apprenants. Les apprenants du Stade 3 produisent moins de phrases incorrectes que les apprenants du Stade 4 (70,5% vs. 80,2%). De plus, le pourcentage de mots inconnus chez le groupe de contrôle (natifs) est légèrement supérieur à celui des apprenants du Stade 4 (10,6% vs. 10,4%) cf. note 1 sur la définition du mot inconnu utilisé ici. Ainsi, le simple calcul des erreurs ne suffit pas pour distinguer les apprenants entre eux et les apprenants des natifs. La distinction des apprenants de différents niveaux linguistiques nécessite des analyses plus détaillées et des mesures plus fines. C'est exactement l'objet des *itinéraires d'acquisition* et de l'analyse de *Direkt Profil*.

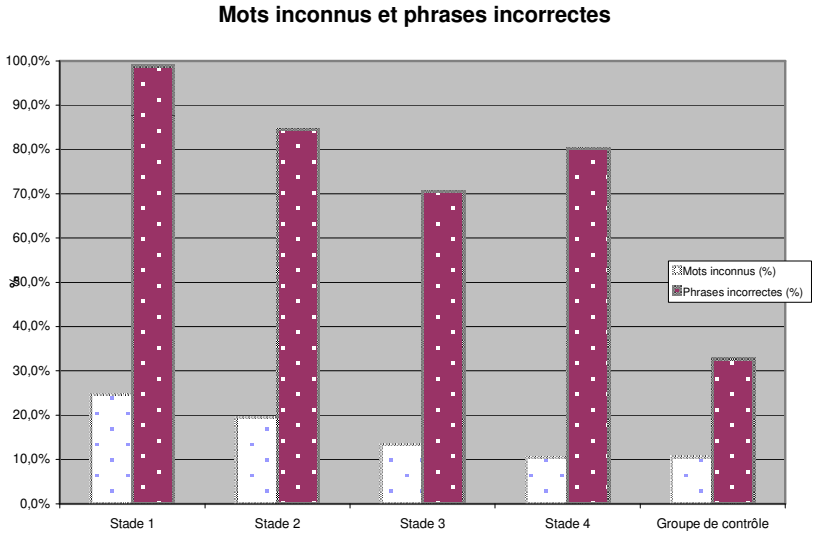


Figure 2. *Les mots inconnus et phrases incorrectes dans le corpus de test.*

[1] Est défini comme « mot inconnu » un mot qui dans sa forme actuelle n'apparaît pas dans le lexique employé par le système (en l'occurrence le lexique de l'ABU CNAM, voir ci-dessous).

[2] Est définie comme « phrase incorrecte » : toute phrase qui contient ou moins une erreur orthographique, syntaxique, morphologique ou sémantique. Pour le niveau de la sémantique, un critère d'interprétation était utilisé : une phrase était considerée sémantiquement incorrecte si elle n'avait aucun sens. Les phrases syntaxiquement etc correctes mais sémantiquement non appropriées dans le contexte (par exemple par rapport aux images, cf. section 2) étaient jugées correctes. De plus, aucune évaluation du choix de mots ou du registre des mots n'a été faite.

## 4.1 Les itinéraires d'acquisition

La différence la plus importante entre *Direkt Profil* et les autres systèmes actuels est la méthode sur laquelle se fonde l'analyse. *Direkt Profil* réalise une analyse des phénomènes locaux qui se sont avérés liés à un développement dans l'apprentissage du français. Ces phénomènes sont décrits sous forme d'*itinéraires d'acquisition*. Les itinéraires sont le résultat d'observations empiriques montrant que certaines constructions grammaticales sont acquises et peuvent être produites dans la langue parlée spontanée dans un ordre fixe. Cet ordre a aussi été nommé « ordre naturel », en anglais « natural order ».

Clahsen et al. (1983) ainsi que Pienemann & Johnston, (1987) ont établi des itinéraires d'acquisition pour allemand et l'anglais parlé. Pour le français parlé, Schlyter (2003) et Bartning et Schlyter (2004) proposent 6 stades de développement et des itinéraires d'acquisition couvrant plus de 20 phénomènes locaux. Ces phénomènes morphosyntaxiques sont décrits sous la forme de structures locales à l'intérieur du domaine verbal ou nominal. Le Tableau 1 en présente un sous-ensemble.

| **Ph** | **Stades** | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|---|
| **A.** | Énoncés contenant un verbe | 20-40% | 30-40% | 50% | 60% | 70% | 75% |
| **B.** | Formes conjuguées (Types de verbes en opposition) | Pas d'opposition | 10-20% | 50% | 75-100% | + | + |
| **C.** | Formes conjuguées (*je parle*) vs. (**je parlE*) (% occurrences) | 50% - 75% | 70-80% | 80-90% | 90-98% | + | + |
| **D.** | 1-2-3 pers singulier (*être/avoir*) | formule sans opposition: *j'ai/ c'est* | opposition (ex. *j'ai* vs. *il a*) | erreurs isolées (ex. **je va* * *je a)* | + | + | + |
| **E.** | 1re pers. pluriel (V-*ons)* (% correct) | - | 70-80% | 80-95% | erreurs dans des constr. complexes | + | + |
| **F.** | Sujet + *viennent*, *veulent*, *prennent* | - | - (ex. **ils prend*) | occs. isolées de V-*(n)ent* | 50% | problèmes encore | + |
| **G.** | Placement des pronoms objet | - | SVO | S(v)oV | SovV apparaît | SovV productif | + |
| **H.** | Genre Article-Nom (% correct) | 55-75% | 60-80% | 65-85% | 70-90% | 75-95% ? | 90-100% |

Tableau 1. *Itinéraires d'acquisition (exemples) et stades proposés.*
Légende : Ph = Phénomène ; occs. = occurrences ; - = pas d'occurrence / pas encore acquis ; + = acquis à 100% ; opp(osition) = deux formes différentes d'un verbe particulier dans le même enregistrement / texte ; formule = expression figées dans la langue de l'apprenant ; V-*ons*/V-*(n)ent* = Racine verbale + flexion.

L'axe horizontal indique le développement en fonction du temps d'un phénomène particulier, donc son itinéraire de développement. L'axe vertical indique l'ensemble de phénomènes grammaticaux étudiés regroupés de telle façon qu'ils présentent des caractéristiques pour l'établissement d'un stade acquisitionnel. Pour illustrer, comparons les phénomènes C (occurrences des formes verbales conjuguées) et G (pronoms objet).

Dès le Stade 1, les formes conjuguées et les formes non-conjuguées coexistent. On trouve aussi bien « je parle » (transcription de /je parl/ analysé comme une « forme conjuguée ») que « je parler » (transcription de /je parle/ analysé comme une forme « non-conjuguée »). L'estimation actuelle est qu'au Stade 1, il y a entre 50 et 75% de formes conjuguées (calculées sur les occurrences des verbes dans un contexte où ils sont normalement conjugués). Au Stade 4, le pourcentage de formes conjuguées a augmenté jusqu'à 90-98%. Pour ce phénomène morphologique, les itinéraires d'acquisition décrivent une morphologisation successive.

Le phénomène G décrit l'itinéraire d'acquisition des pronoms objets. Les premiers pronoms objets sont placés dans une position post-verbale selon le schéma Sujet-Verbe-Objet (SVO), par exemple **je vois le/la/lui* (pour *je le/la vois*). Au stade 3, les apprenants peuvent produire des énoncés selon le schéma SvoV (Sujet-verbe auxiliaire-pronom objet-Verbe), par exemple *Je veux le voir* (correct) mais aussi **j'ai le vu* (incorrect). Au stade 4, *je l'ai vu* apparaît. Pour ce phénomène syntaxique, les itinéraires d'acquisition décrivent un changement dans l'organisation linéaire des constituants concernés.

Ce tableau est soumis à une révision continuelle et les taux sont encore approximatifs. L'important est que les itinéraires d'acquisition nous fournissent un grand nombre d'hypothèses détaillées sous la forme de structures locales et qui ensemble couvrent une grande partie de la langue (domaine verbal et nominal).

# 5 Annotation

Le concept de groupe, nominal ou verbal, correct ou non, représente le support grammatical essentiel de notre annotation. La plupart des normes d'annotation syntaxique pour le français tiennent compte d'une manière ou d'une autre de tels groupes. Celle que propose Gendner et al. (2004) réconcilie un grand nombre de pratiques et forme une base consensuelle. Ces normes sont cependant insuffisantes pour rendre compte des constructions du Tableau 1.

Nous avons défini une annotation des textes propres au projet *Direkt Profil*. Elle repose sur l'inventaire des phénomènes linguistiques caractéristiques des itinéraires de développement et elle reprend les catégories décrites par Barting et Schlyter (2004) (Tableau 1). Nous avons représenté ces phénomènes par des arbres de décisions dont les nœuds terminaux correspondent à une catégorie d'analyse.

L'annotation de *Direkt Profil* utilise le format XML et annote les textes sur 4 niveaux. Seul le 3$^{e}$ niveau est réellement d'ordre syntaxique.

- Le premier niveau correspond à la segmentation du texte en mots.

- Le deuxième niveau annote les multimots et les expressions figées (par exemple *je m'appelle*). Ces expressions correspondent aux phrases « par cœur » qui ont une grande importance dans les premières années d'apprentissage du français.

- Le troisième niveau correspond à une annotation syntaxique partielle du texte, restreinte aux phénomènes à identifier. Ce niveau balise simultanément chacun des mots avec sa partie du discours et les groupes verbaux et nominaux auxquels ils appartiennent. Le groupe verbal incorpore les pronoms clitiques y compris les sujets. L'élément XML `span` marque les groupes et comporte un attribut pour indiquer leur type dans la grille. Les parties du discours utilisent un élément `tag` avec des attributs pour

indiquer le lemme, la partie du discours et les traits grammaticaux. Pour le groupe verbal, la phrase *Ils parlons dans la bar* extraite du texte d'introduction reçoit l'annotation `<span class="c03"><tag pos="pro:nom:pl:p3:mas"> Ils</tag> <tag pos="ver:impre:pl:p1"> parlons </tag></span> dans la bar`. La classe `c03` s'interprète comme « verbe lexical conjugué sans accord ».

- Le quatrième niveau dénombre les types de structures caractéristiques d'un stade d'acquisition. Il utilise un élément XML `counter`, `<counter id="counter.2" counter_name="passe_compose" rule_id="participe_4b" value="1"/>`.

# 6 Implantation

Le programme *Direkt Profil* est pour l'instant restreint à l'analyse des groupes verbaux et des pronoms clitiques sujets. Pour chacune des catégories du Tableau 1, le programme détecte les constructions correspondantes dans un texte et les dénombre.

L'analyseur utilise des règles écrites manuellement et s'appuie sur un lexique de formes fléchies. La variété des constructions contenues dans le corpus est grande et afin de ne pas multiplier le nombre de règles, nous avons choisi une stratégie d'analyse par renforcement de contraintes. Conceptuellement, l'analyseur recherche des classes de structures syntagmatiques dont tous les traits ont été ôtés. Il identifie les structures progressivement en faisant varier les valeurs des traits. La reconnaissance des limites des groupes se fait par un ensemble de mots vides et par des heuristiques à l'intérieur des règles. Elle suit ainsi une stratégie ancienne mais robuste utilisée notamment par Vergne (1999), *inter alia*, pour le français.

*Direkt Profil* applique en cascade trois ensembles de règles pour produire les quatre niveaux d'annotations. Le premier ensemble segmente le texte en mots. Un ensemble intermédiaire identifie les expressions figées. Le troisième ensemble annote simultanément les parties du discours et les groupes. Finalement, le moteur crée un groupe de résultats relié au Stade de l'apprenant. Il est à noter que le moteur n'annote pas tous les mots, ni tous les segments. Il ne considère que ceux qui sont pertinents pour la détermination du Stade. Le moteur applique les règles de gauche à droite puis de droite à gauche pour résoudre certains problèmes d'accord.

Les règles représentent des structures partielles de groupe et sont divisées en une partie condition et une partie action. La partie condition contient les paramètres de recherche. Il peut s'agir d'un lemme, d'une expression régulière ou d'une classe de flexion. Le moteur parcourt le texte et applique les règles à l'aide d'un arbre de décision. Il exécute la partie condition pour identifier les séquences de mots contigus. Chaque règle produit un résultat positif (« *match* ») ou négatif (« *no match* »). Ils sont exécutés suivant le résultat de la condition et ont pour effet d'annoter le texte, de compter le nombre d'occurrences du phénomène et d'enchaîner une autre règle. En parcourant les nœuds de l'arbre, le moteur mémorise les règles qu'il a parcourues sur son chemin ainsi que les résultats des parties condition de ces règles. Quand il arrive à un nœud terminal, le moteur applique les parties action de toutes les règles.

Le moteur recherche les mots dans un dictionnaire de formes fléchies. Il ne corrige pas les fautes d'orthographe à l'exception des accents et de certains radicaux. En effet, les apprenants construisent fréquemment des participes passés erronés à partir d'une généralisation abusive

de radicaux verbaux voisins. Un exemple est le mot **prendu* (*pris*) formé sur le radical de *prendre* et du suffixe de *rendu*. Nous avons utilisé le lexique disponible au site de l'Association des Bibliophiles Universels que nous avons corrigé, transposé en XML et que nous avons enrichi des radicaux des verbes.

# 7 Interface

*Direkt Profil* fusionne l'ensemble des niveaux d'annotations dans un objet résultat. Cet objet représente le texte d'origine, l'annotation et la trace de l'application des règles et des compteurs. L'objet résultat, qui peut être enregistré, est ensuite transformé par le programme pour être présenté à l'utilisateur. L'affichage utilise les spécifications XHTML 1.1 qui peuvent être lues par un navigateur internet. *Direkt Profil* fonctionne en mode client-serveur où le serveur réalise l'annotation d'un texte et le client, intégré à un navigateur, prend en charge l'affichage et l'interaction.

La Figure 3 est une copie d'écran de l'interface graphique de *Direkt Profil* qui montre l'analyse du texte présenté dans l'Introduction. L'interface indique à l'utilisateur par une couleur différente toutes les structures que l'analyseur a détectées. Le cadre inférieur donne le code des couleurs et le dénombrement de ces structures.

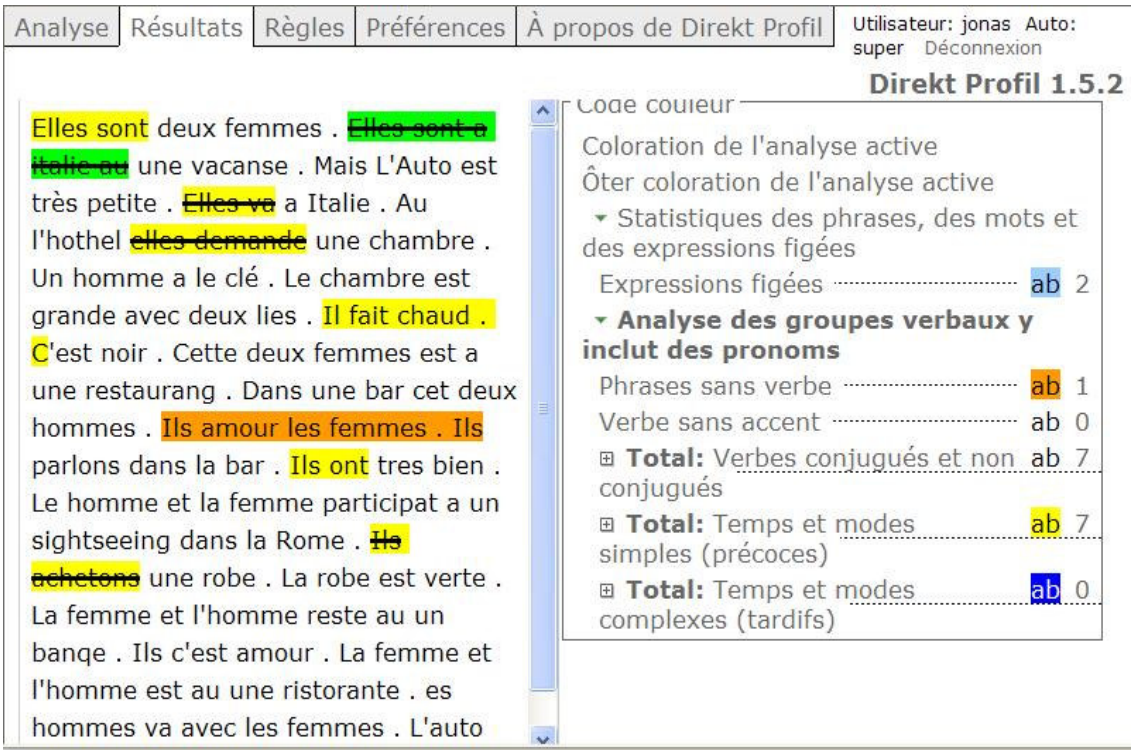


Figure 3. *L'interface graphique de Direkt Profil 1.5.*

# 8 Résultats et évaluation

Nous avons évalué *Direkt Profil* avec un sous-ensemble du corpus CEFLE de Lund. Nous avons choisi 20 textes au hasard répartis sur 4 Stades d'apprentissage. Nous avons aussi utilisé 5 textes provenant du groupe de contrôle. Nous n'avons pas testé dans cette version la correction des mots mal orthographiés : accent et radicaux. Le Tableau 2 présente quelques

statistiques sur la taille de textes et le Tableau 3 donne les résultats sous forme de rappel et précision.[3]

| | Stade 1 | Stade 2 | Stade 3 | Stade 4 | Contrôle | Total |
|---|---|---|---|---|---|---|
| Nombre de textes analysés | 5 | 5 | 5 | 5 | 5 | 25 |
| Nombre de mots | 740 | 1233 | 1571 | 1672 | 1626 | 6842 |
| Nombre de phrases | 85 | 155 | 166 | 126 | 107 | 639 |
| Longueur moyenne des textes (mots) | 148 | 247 | 314 | 334 | 325 | |
| Longueur moyenne des phrases | 8,7 | 7,9 | 9,5 | 13,3 | 15,2 | |

Tableau 2. *Corpus de test.*

| | Stade 1 | Stade 2 | Stade 3 | Stade 4 | Contrôle | Total |
|---|---|---|---|---|---|---|
| Nombre de structures de référence | 23 | 97 | 101 | 119 | 85 | 425 |
| Nombre de structures proposées | 27 | 98 | 100 | 112 | 92 | 429 |
| Nombre de structures détectées correctes | 15 | 81 | 89 | 96 | 73 | 354 |
| Nombre de structures non détectées | 5 | 16 | 12 | 20 | 11 | 64 |
| Nombre de structures surdétectées | 10 | 17 | 11 | 17 | 19 | 74 |
| Rappel | 65 | 84 | 88 | 81 | 86 | 83 |
| Précision | 56 | 83 | 89 | 86 | 79 | 83 |
| F-measure | 0,6 | 0,83 | 0,89 | 0,83 | 0,82 | 0,83 |

Tableau 3 : *Résumé des résultats de Direkt Profil 1.5.1*

À un niveau global, les résultats montrent que *Direkt Profil* parvient bien à détecter les phénomènes désirés. Il révèle aussi des différences intéressantes suivant les stades des textes. Le tableau montre que les textes du Stade 1 sont les plus difficiles à traiter (rappel de 65%). Ceci est dû en grande partie au nombre de *mots inconnus* à ce stade d'acquisition (cf. *infra* Figure 2). Le résultat en est une surdectection du phénomène « phrases sans verbes » à ce stade. Les résultats montrent aussi que *Direkt Profil* analyse mieux les textes des apprenants que les textes des étudiants français (Groupe de contrôle). Sans savoir exactement à quoi ce résultat est dû, nous pouvons constater qu'il suggère que la démarche adoptée qui vise à l'analyse des textes en français langue étrangère semble prometteuse.

# 9 Conclusion et travaux futurs

Nous avons présenté un système réalisant une analyse automatique de textes fondée sur les itinéraires d'acquisition dont le but est de produire un profil d'apprenant. Nous avons construit un analyseur syntaxique et développé un ensemble de règles pour annoter les textes. *Direkt Profil* est intégré dans une architecture client-serveur et dispose d'une interface permettant l'interaction avec l'utilisateur.

Les résultats montrent qu'il est possible de décrire sous forme de règles la vaste majorité des structures locales définies par les itinéraires d'acquisition. *Direkt Profil* peut ainsi les détecter

[3] Dans le corpus de test, nous n'avons pas d'exemples de textes d'apprenants des Stades 5 et 6 (cf. Tableau 1). Pour le moment, les textes du groupe de contrôle (des jeunes Français) peuvent servir d'exemples des textes de haut niveau linguistique.

et les analyser automatiquement. Nous pouvons ainsi vérifier la validité des critères d'acquisition établis pour l'oral sur notre corpus écrit.

*Direkt Profil* peut aussi avoir un rôle pédagogique car il sera possible d'analyser, de manière automatique et précise, le niveau grammatical d'un apprenant ou d'un élève dans la production écrite et libre. Le programme pourra être utilisé d'une part par des professeurs pour évaluer les textes de leurs élèves et d'autre part par les élèves eux-mêmes comme auto-évaluation et dans le perfectionnement de la langue.

Une version préliminaire de *Direkt Profil* est disponible en ligne à l'adresse http://www.rom.lu.se:8080/profil

## Références

BARTNING, I, S. SCHLYTER (2004) « Stades et itinéraires acquisitionnels des apprenants suédophones en français L2 ». *JFLS* (Journal of French Language Studies). À paraître.

BIGERT, J., KANN, V., KNUTSSON, O., SJÖBERGH J. (2004). Grammar checking for Swedish second language learners. Chapter in CALL in the Nordic Languages, Copenhagen Studies in Language, Copenhagen Business School.

CLAHSEN, H., MEISEL, J-M., PIENEMANN, M. (1983) *Deutsch als Fremdsprache. Der Spracherwerb ausländischer Arbeiter*. Tübingen: Narr

GENDNER, V., VILNAT, A., MONCEAUX, L., PAROUBEK, P., ROBBA I. (2004) Les annotations syntaxiques de référence PEAS.
http://www.limsi.fr/Recherche/CORVAL/easy/PEAS_reference_annotations_v1.6.html

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

HEIDORN G. E., Intelligent Writing Assistance, (2000) in Robert Dale, Hermann Moisl, et Harold Somers eds, *Handbook of Natural Language Processing*, Marcel Dekker.

JENSEN, K., HEIDORN G. E., RICHARDSON S. D., (1993) *Natural Language Processing: The PLNLP Approach*, Kluwer Academic Publishers.

PIENEMANN, M, JOHNSTON, M (1987) « Factors influencing the development of second language proficiency". In D. Nunan (ed.) *Applying second language acquisition research*, 45-141. Adelaide: National Curriculum Resource Centre.

SCHLYTER, S. (2003). « Stades de développement en français L2 ». Ms. Institut d'études romanes de Lund. Université de Lund. Disponible en ligne : http://www.rom.lu.se/durs/STADES_DE_DEVELOPPEMENT_EN_FRANCAIS_L2.pdf

VERGNE, J. (1999) *Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur. Analyse syntaxique automatique non combinatoire. Synthèse et Résultats.* Habilitation à Diriger des Recherches, 29 septembre 1999, Caen.

ÅGREN, M. (2005) « Le marquage morphologique du nombre dans la phrase nominale. Une étude sur l'acquisition du français L2 écrit » Ms. Institut d'études romanes de Lund. Université de Lund.