Évaluation des stades de développement en français langue étrangère*

Jonas GRANFELDT[1], Pierre NUGUES[2]
[1] Centre de langues et de littérature, Université de Lund, S-221 00 Lund
[2] Institut d'informatique, Institut de Technologie de Lund, S-221 00 Lund
Jonas.Granfeldt@rom.lu.se, Pierre.Nugues@cs.lth.se

Résumé. Cet article décrit un système pour définir et évaluer les stades de développement en français langue étrangère. L'évaluation de tels stades correspond à l'identification de la fréquence de certains phénomènes lexicaux et grammaticaux dans la production des apprenants et comment ces fréquences changent en fonction du temps. Les problèmes à résoudre dans cette démarche sont triples : identifier les attributs les plus révélateurs, décider des points de séparation entre les stades et évaluer le degré d'efficacité des attributs et de la classification dans son ensemble. Le système traite ces trois problèmes. Il se compose d'un analyseur morphosyntaxique, appelé Direkt Profil, auquel nous avons relié un module d'apprentissage automatique. Dans cet article, nous décrivons les idées qui ont conduit au développement du système et son intérêt. Nous présentons ensuite le corpus que nous avons utilisé pour développer notre analyseur morphosyntaxique. Enfin, nous présentons les résultats sensiblement améliorés des classificateurs comparé aux travaux précédents (Granfeldt *et al.*, 2006). Nous présentons également une méthode de sélection de paramètres afin d'identifier les attributs grammaticaux les plus appropriés.

Abstract. This paper describes a system to define and evaluate stages of development in second language French. The task of identifying such stages can be formulated as identifying the frequency of some lexical and grammatical features in the learners' production and how they vary over time. The problems in this procedure are threefold : identify the relevant features, decide on cutoff points for the stages, and evaluate the degree of efficiency of the attributes and of the overall classification. The system addresses these three problems. It consists of a morphosyntactic analyzer called Direkt Profil and a machine-learning module connected to it. We first describe the usefulness and rationale behind the development of the system. We then present the corpus we used to develop our morphosyntactic analyzer called Direkt Profil. Finally, we present new and substantially improved results on training machine-learning classifiers compared to previous experiments (Granfeldt *et al.*, 2006). We also introduce a method of attribute selection in order to identify the most relevant grammatical features.

Mots-clés : analyseur morphosyntaxique, apprentissage automatique, acquisition des langues.

Keywords: morphosyntactic parser, machine learning, language acquisition.

* Une première version de cet article a été présentée à la 16e conférence nordique de traitement automatique des langues, NODALIDA 2007, Tartu, Estonie, sous le titre *Evaluating Stages of Development in Second Language French : A Machine-Learning Approach*.

1 Introduction

L'un des points essentiels des recherches sur l'acquisition des langues étrangères est l'identification et l'analyse des stades de développement que l'on traverse en apprenant une deuxième langue ou une langue étrangère. La notion de stade de développement peut s'appliquer aux données de tous les niveaux linguistiques, mais elle est particulièrement intéressante pour rendre compte de l'acquisition de la morphologie et de la syntaxe. Dans ce cadre, on peut considérer la grammaire interne de l'apprenant comme un système propre qui se développerait et subirait des restructurations au cours du temps.

La modélisation du développement de la grammaire de l'apprenant et de ses propriétés à des moments différents revient, pour l'essentiel, à identifier les phénomènes grammaticaux pertinents, à définir des points de séparation entre les stades et à les évaluer de manière systématique. Dans cet article, nous décrivons et nous évaluons un système qui a entièrement automatisé ce processus. Avant de le présenter, nous décrivons de façon simplifiée comment on identifie en général les stades de développement dans le domaine de l'acquisition des langues.

2 Contexte

2.1 Méthode actuelle pour identifier les stades de développement

La première étape pour identifier les stades de développement est de déterminer et d'extraire des phénomènes grammaticaux dans la production, orale ou écrite, d'une population représentative d'apprenants. Le point crucial dans le choix de ces phénomènes est qu'ils aient une validité interne, dont les réalisations peuvent traduire un changement qualitatif de la grammaire. Une deuxième étape est de comprendre et modéliser leur développement.

Certains phénomènes linguistiques montrent un développement linéaire simple et les pourcentages d'usages corrects ont une augmentation stable avec le temps. D'autres phénomènes ont un développement non linéaire, parfois en forme de « U », où les pourcentages d'usages corrects au début de l'acquisition sont élevés mais diminuent dans une deuxième phase pour ensuite regagner un niveau élevé de rectitude dans une troisième phase. Une explication qui a souvent été proposée pour ce type d'évolution est que la première phase contient un certain nombre d'expressions fixes apprises de façon holistique par l'apprenant. Ces structures, peut-être apprises par cœur, représenteraient alors des structures linguistiques non-analysées dans la grammaire interne des apprenants. Ensuite, une fois que les séquences de développement sont connues, il faut décider des points de séparation dans les données où l'apprenant a atteint un nouveau stade de développement. La plupart du temps, il est préférable de prendre en compte plusieurs phénomènes grammaticaux pour en même temps réaliser un « profilage grammatical ».

2.2 Problèmes de la méthode actuelle

L'analyse morphosyntaxique détaillée de textes ou d'énoncés produits par les apprenants est une partie centrale dans la méthode décrite dans le paragraphe précédent. La plupart des analystes travaillant sur l'acquisition d'une première et d'une deuxième langue ont maintenant accès à de grands corpus de productions orales et écrites. Dans notre cas, ce sont des textes écrits mais on

pourrait tout aussi bien l'appliquer à des transcriptions de productions orales. Pour des langues répandues, comme l'anglais et le français, on dispose également d'outils tels que des analyseurs morphologiques et des étiqueteurs de parties du discours (MacWhinney, 2000). Ces outils peuvent réduire de façon considérable le temps de l'analyse morphosyntaxique qui autrement serait très fastidieuse.

Cependant même avec ces outils, beaucoup d'analyses manuelles restent à faire. D'abord il n'existe actuellement aucun outil automatisé fiable pour l'analyse syntaxique de textes d'apprenants malgré quelques tentatives récentes pour l'anglais (Sagae *et al.*, 2005). Pour le français, une partie des structures linguistiques utilisées dans le profilage grammatical peuvent être détectées en utilisant des outils comme CHILDES. Mais pour d'autres structures plus complexes telles que l'accord entre les constituants, c'est impossible. Un autre problème est qu'avec les outils actuels, on ne peut effectuer que des requêtes simples sur un phénomène individuel alors que dans le profilage grammatical, on doit en analyser un grand nombre en même temps.

Un troisième problème concerne le côté artificiel des stades. Le résultat de l'analyse morphosyntaxique est présenté typiquement sous forme de fréquences de certains phénomènes. Pour un phénomène linguistique particulier, par exemple l'accord sujet-verbe à la troisième personne du singulier au présent, on identifie les différentes réalisations de cette structure et on les compte. Les données compilées pour tous les apprenants et tous les phénomènes et structures faisant partie du profil grammatical sont ensuite inspectées afin d'identifier intuitivement des stades de développement. Il y a actuellement de multiples façons de traiter cette étape, mais aucune n'a reçu d'évaluation systématique. Une raison possible est que personne n'ait relié les deux premières étapes, l'analyse morphosyntaxique et l'analyse des fréquences, à un traitement statistique. Si un traitement entièrement automatisé du processus était disponible, toutes ses étapes auraient pu être évaluées plus complètement.

Dans le reste de l'article, nous présentons notre système qui vise à surmonter les problèmes mentionnés précédemment. Nous commençons par un bref résumé des travaux précédents sur le développement morphosyntaxique du français langue étrangère. Nous décrivons ensuite le corpus que nous employons ainsi que de façon brève notre analyseur, Direkt Profil. Dans les derniers paragraphes, nous discutons de notre démarche fondée sur l'apprentissage automatique pour définir et évaluer les stades de développement et pour sélectionner les attributs. Nous présentons enfin nos résultats actuels.

3 Développement morphosyntaxique du français deuxième langue

Une partie des recherches sur le développement morphosyntaxique de français langue étrangère a pour objectif d'atteindre une description détaillée de la façon dont les apprenants développent leur grammaire en fonction du temps. L'étude de Bartning & Schlyter (2004) en est un exemple pour le français parlé, où les auteurs ont identifié environ 25 constructions morphosyntaxiques différentes et ont proposé une définition de leur développement dans le temps pour des Suédois adultes. Pris ensemble, ces phénomènes délimitent six stades sous la forme de profils grammaticaux qui s'étendent des débutants aux apprenants très avancés. Des exemples de constructions sont donnés dans le tableau 1. Au fur et à mesure que l'apprenant automatise la mise en œuvre de la langue cible, les structures produites deviennent plus fréquentes, plus complexes et plus correctes. Les itinéraires d'acquisition décrivent ce processus en termes linguistiques.

Stades	1	2	3	4	5	6
% Formes conjuguées de verbes lexicaux en contextes obligatoires	50-75	70-80	80-90	90-98	100	100
% Accord 1re personne pluriel S-V (*nous V-ons*)	–	70-80	80-95	100	100	100
% Accord 3e pers pluriel avec verbes irréguliers lexicaux comme *viennent*, *veulent*, *prennent*	–	–	qq cas	≈ 50	qq erreurs	100
Placement des pronoms objets	–	SVO	S(v)oV	SovV app.	SovV prod	acquis (*y* et *en*)
% Accord genre grammatical	55-75	60-80	65-85	70-90	75-95	90-100

TAB. 1 – Itinéraire de développement d'après Bartning & Schlyter (2004). Légende : – = pas d'occurrences ; app = apparaît ; prod = productif niveau avancé.

4 Un corpus écrit de français langue étrangère

Pour développer notre analyseur (voir § 5) et expérimenter notre approche d'apprentissage automatique des stades, nous avons utilisé le Corpus Écrit de Français Langue Étrangère de Lund (Ågren, 2005) – CEFLE. CEFLE se compose de textes en français provenant de 85 étudiants suédois à différents niveaux de développement. Il contient approximativement 400 textes et 100 000 mots. Il comporte également des textes d'un groupe de contrôle de 22 jeunes Français du même âge. CEFLE a été compilé lors de l'année scolaire 2003/2004 pendant laquelle chaque étudiant a écrit quatre ou cinq textes à deux mois d'intervalle.

Pour notre étude, nous avons utilisé un sous-ensemble de 317 textes du corpus dont les caractéristiques sont données dans le tableau 2. En employant les critères décrits dans Bartning & Schlyter (2004), un membre de l'équipe a au préalable annoté un texte de chaque étudiant et a estimé le stade de développement qu'il reflétait. Pour les expérimentations que nous décrivons dans les paragraphes qui suivent, nous avons attribué de façon systématique la même classification aux trois ou quatre autres textes du même étudiant dans le corpus. Nous avons ainsi propagé le stade annoté à la main à tous les textes du même étudiant. Nous avons supposé que généralement l'étudiant ne monterait pas d'un stade pendant la courte période où a eu lieu la collecte des textes.

CEFLE			**Sous-ensemble de CEFLE** (moyenne)			
Tâche	Type d'élicitation	Mots	Stade	Textes	Taille texte	Taille phrases
Homme	Images	17 260	Stade 1	23	78	6,9
Souvenir	Récit personnel	14 365	Stade 2	98	161	8,4
Italie	Images	30 840	Stade 3	97	212	9,8
Moi	Récit personnel	30 355	Stade 4	58	320	11,6
Total		92 820	Contrôle	41	308	15,2

TAB. 2 – La description générale du corpus CEFLE et du sous-ensemble utilisé dans les expériences rapportées dans cet article.

5 Direkt Profil

Direkt Profil (Granfeldt *et al.*, 2005, 2006) est un analyseur morphosyntaxique conçu pour du français langue étrangère. Le but initial était de mettre en œuvre une analyse automatique des phénomènes et des constructions grammaticaux contenus dans le tableau 1. Dans sa version actuelle, le système ne détecte pas certains des phénomènes indiqués par Bartning & Schlyter (2004), mais en contrepartie il en détecte un grand nombre d'autres. Le système a fait l'objet d'une présentation détaillée dans des articles précédents et nous nous contenterons d'en donner une brève description.

Le concept de groupe, nominal ou verbal, correct ou non, représente le support grammatical essentiel de notre analyse. Nous avons défini une annotation des textes, propre au projet, fondé sur ces groupes. Elle prend en compte les phénomènes linguistiques caractéristiques des itinéraires de développement d'après les catégories décrites par Bartning & Schlyter (2004). La version actuelle de Direkt Profil, V. 2.1, détecte trois types de groupes syntaxiques : nominaux non-récursifs, verbaux et prépositionnels.

L'analyseur utilise des règles écrites manuellement et s'appuie sur un lexique de formes fléchies. De façon conceptuelle, l'analyseur recherche des classes de structures syntagmatiques sans considérer leurs traits grammaticaux. Il identifie ensuite les structures progressivement en tentant d'affecter des valeurs à ces traits. La reconnaissance des limites des groupes se fait par un ensemble de mots vides et par des heuristiques à l'intérieur des règles. Direkt Profil applique en cascade trois ensembles de règles pour produire 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.

La version actuelle de Direkt Profil est accessible en ligne à l'adresse www.rom.lu.se :8080/profil. La performance de la version 1.5.2 pour la détection des segments a été évaluée dans Granfeldt *et al.* (2005). Les résultats ont donné une moyenne harmonique F globale de précision et de rappel de 0,83.

6 Une méthode d'apprentissage automatique pour évaluer les stades de développement

À l'heure actuelle, il existe des quantités de méthodes pour définir les stades de développement, mais à notre connaissance aucune façon systématique pour les évaluer. Dans leur article, Bartning & Schlyter (2004) avaient défini six stades de développement. Le corpus CEFLE en utilise cinq attribués par un annotateur humain. Un problème essentiel dans cette dernière étape est que l'analyse montre une augmentation progressive des fréquences avec l'acquisition qui suggère plutôt un développement en continu que selon des stades discrets. D'une certaine manière, il faut donc accepter que n'importe quelle définition soit en partie arbitraire.

Dans notre système, l'analyse des fréquences des constructions grammaticales est obtenue automatiquement et correspond à la sortie de Direkt Profil. Elle forme le support qui permet d'établir

les stades de développement. Dans le paragraphe suivant, nous évaluons la probabilité de l'existence de ces cinq stades différents en utilisant des techniques d'apprentissage automatique.

6.1 Première expérience : classification utilisant tous les attributs

Comme conditions expérimentales, nous avons employé chacun des textes des 85 étudiants qui a reçu manuellement un stade de développement. Nous avons ensuite réutilisé la même classification pour les trois ou quatre autres textes du même étudiant dans le corpus, ce qui donne comme résultat 276 textes d'apprenant classifiés. Les 41 textes supplémentaires viennent du groupe de contrôle des natifs, ce qui aboutit à un total de 317 textes classifiés. La phase d'apprentissage induit automatiquement des classifieurs à partir des vecteurs de 142 attributs que nous extrayons des textes au moyen de l'analyseur.

Nous avons employé trois algorithmes d'apprentissage automatique : ID3/C4.5 (Quinlan, 1986), les machines à vecteurs de support (SVM) (Boser *et al.*, 1992) et les arbres de modèles logistiques (LMT) (Landwehr *et al.*, 2003). Dans un premier temps, nous avons regroupé les cinq stades dans trois stades plus généraux, où les stades 1 et 2 ainsi que les stades 3 et 4 ont été fusionnés et nous avons entraîné les algorithmes sur ces trois stades. Nous avons ensuite réalisé une deuxième évaluation avec les cinq stades d'origine. Nous avons réalisé toutes nos expériences avec l'ensemble d'algorithmes d'apprentissage automatique disponible dans Weka[1] (Witten & Frank, 2005) et nous les avons évalués en appliquant 10 fois une validation croisée sur le corpus d'apprentissage. Les tableaux 3 et 4 présentent les résultats pour les 317 textes pour 3 et 5 classes respectivement.

	C4.5			SVM			LMT		
Stades	P	R	F	P	R	F	P	R	F
1–2	0,66	0,70	0,68	0,70	0,71	0,71	0,76	0,75	**0,75**
3–4	0,70	0,68	0,69	0,71	0,72	0,71	0,76	0,79	**0,77**
Contrôle	0,71	0,66	0,68	0,70	0,63	0,67	0,89	0,83	**0,86**

TAB. 3 – Résultats de la classification des textes en trois stades pour les trois classifieurs. Chaque classifieur a employé 142 attributs et a été entraîné sur 317 textes du corpus CEFLE. P : Précision. R : Rappel, F : Moyenne harmonique de la précision et du rappel.

	C4.5			SVM			LMT		
Stades	P	R	F	P	R	F	P	R	F
1	0,37	0,42	0,39	0,54	0,58	**0,56**	0,44	0,33	0,38
2	0,50	0,52	0,51	0,60	0,60	**0,60**	0,59	0,61	**0,60**
3	0,42	0,46	0,44	0,45	0,46	0,45	0,51	0,54	**0,53**
4	0,48	0,38	0,42	0,52	0,50	0,51	0,64	0,66	**0,65**
Contrôle	0,71	0,66	0,68	0,70	0,63	0,67	0,89	0,83	**0,86**

TAB. 4 – Résultats de la classification des textes en cinq stades pour les trois classifieurs. Chaque classifieur a employé 142 attributs et a été entraîné sur 317 textes du corpus CEFLE.

[1] Accessible à partir de ce site : http ://www.cs.waikato.ac.nz/ml/weka/.

Ces résultats peuvent être comparés à ceux que nous avons obtenus avec une version précédente de Direkt Profil (1.5.4) employant un nombre plus petit de 33 attributs et un corpus d'entraînement moins important de 80 textes. Ces résultats (Granfeldt *et al.*, 2006) ont montré que le meilleur classifieur, SVM, obtenait une moyenne harmonique de précision et de rappel de près de 70% pour la classification en trois stades, et une moyenne de 43% de précision et 36% de rappel pour une classification en cinq stades. Les résultats actuels avec plus de 100 attributs supplémentaires et un corpus d'entraînement qui est quatre fois plus grand montrent une amélioration de presque 10%. Le meilleur algorithme, cette fois LMT, obtient une moyenne de précision et de rappel de 79% pour la classification en trois stades (Tableau 3). Pour la classification en cinq stades, l'amélioration est encore plus importante (Tableau 4). LMT obtient 62% de précision et 59% de rappel. En comparant la performance des deux meilleurs algorithmes, SVM et LMT, nous observons que LMT est supérieur à SVM sur les stades intermédiaire et avancé – 3, 4, et le groupe de contrôle des natifs – mais pas sur les deux premiers stades de développement. Nous n'avons aucune explication pour ce fait.

En conclusion de cette première expérience, nous pouvons affirmer que le plus grand nombre d'attributs et le corpus d'entraînement plus important ont eu comme résultat une meilleure performance globale pour les trois classifieurs. Mais l'amélioration n'a pas été aussi grande qu'espérée. Une hypothèse possible est que nous avons introduit un certain nombre d'attributs non pertinents parmi les quelques 100 nouveaux. Pour cette raison, nous avons appliqué une procédure de sélection afin d'identifier les meilleurs attributs. Les résultats de cette deuxième expérience sont présentés dans le paragraphe suivant.

6.2 Deuxième expérience : classification utilisant une sélection d'attributs

Pour évaluer les 142 attributs, nous avons mesuré le gain d'information (Quinlan, 1986) pour chaque attribut par rapport à la classe. Ce critère est à la base de l'algorithme ID3 et fait partie de la boîte à outils Weka. Nous avons employé la méthode de recherche de rang qui classe les différents attributs en fonction de leur évaluation. Les tableaux 5 et 6 présentent les résultats pour respectivement les 10 et 20 meilleurs attributs selon cette méthode. Dans un deuxième temps, nous avons réalisé deux nouvelles classifications en utilisant les mêmes algorithmes que dans la première expérience et le même choix de 317 textes du corpus, mais cette fois avec un nombre d'attributs réduit aux meilleurs d'entre eux.

Lors de la première classification, nous avons évalué la performance des classifieurs en utilisant les 10 meilleurs attributs. Les résultats produits sont mitigés (voir le tableau 7 pour la classification en cinq stades). En moyenne, la réduction radicale du nombre d'attributs de 142 à 10 ne semble pas beaucoup affecter les résultats. Les moyennes des précision et rappel pour LMT sont respectivement de 66% et 58%. Ceci suggère qu'il y a beaucoup de bruit dans les 132 attributs restants. En revanche, les résultats pour le premier stade de développement se détériorent de façon importante. L'algorithme SVM n'identifie plus un seul texte au stade 1. Ceci suggère que le reste des 132 attributs contient des informations très importantes pour identifier ce stade. Dans notre deuxième évaluation, nous avons incorporé les 10 attributs suivants (attributs 11–20, soit au total 20 attributs). Les résultats pour la classification en cinq stades sont présentés dans le tableau 8.

Les résultats globaux de cette deuxième classification sont meilleurs et le vecteur de 20 attributs permet à chacun des trois classifieurs d'identifier des textes au stade 1. Cependant la moyenne pour LMT est en légère baisse par rapport à la classification avec 10 attributs. On note également

Mérite	Rang	Attribut
0,405	1,4	Pourcentage de séquences déterminant-nom avec accord (nombre et genre)
0,354	2,2	Pourcentage de mots inconnus
0,33	3,2	Pourcentage de GNs avec accord en genre
0,313	3,9	Pourcentage de prépositions (sur toutes les parties de discours)
0,311	4,3	Longueur moyenne des phrases
0,208	6,2	Pourcentage de séquences nom-adjectif avec accord (nombre et genre)
0,198	7,4	Pourcentage d'accord sujet-verbe avec des verbes modaux + INF
0,187	8,3	Pourcentage d'accord sujet-verbe avec verbes au passé composé
0,177	9,3	Pourcentage d'accord sujet-verbe avec être/avoir au 3e personne pluriel
0,176	9,8	Pourcentage d'accord sujet-verbe avec verbes modaux et sujets pronominaux

TAB. 5 – Les 10 meilleurs attributs. Attributs 1–10.

Mérite	Rang	Attribut
0,168	11,4	Pourcentage de verbes au présent (sur tous les temps)
0,165	11,8	Pourcentage de verbes au passé composé (sur tous les temps)
0,15	14	Pourcentage d'accord sujet-verbe avec aux. de mode (sur tous les sujets)
0,142	15,7	Pourcentage d'accord sujet-verbe avec aux. de mode au sg.
0,14	16,2	Pourcentage d'accord sujet-verbe avec aux. de mode au présent et sujet pronominal 3e personne
0,136	16,7	Pourcentage verbes lexicaux conjugués dans des contextes conjugués
0,133	17,3	Pourcentage d'accord sujet-verbe avec verbes lexicaux conjugués
0,131	18,1	Pourcentage d'accord sujet-verbe avec sujet pronominal sg et aux. de mode
0,125	19,3	Pourcentage d'accord sujet-verbe avec verbes lexicaux à la 3e personne du pluriel
0,116	21,4	Pourcentage d'accord sujet-verbe avec sujet pronominal et être/avoir

TAB. 6 – Les 10 attributs suivants. Attributs 11–20.

	C4.5			SVM			LMT		
Stades	P	R	*F*	P	R	*F*	P	R	*F*
1	0,46	0,46	**0,46**	0,00	0,00	0,00	0,78	0,29	0,42
2	0,50	0,49	0,49	0,53	0,72	0,61	0,57	0,70	**0,63**
3	0,43	0,42	0,43	0,50	0,43	0,46	0,55	0,49	**0,52**
4	0,50	0,57	0,53	0,62	0,71	**0,66**	0,63	0,64	0,63
Contrôle	0,84	0,76	0,79	0,94	0,76	**0,84**	0,78	0,78	0,78

TAB. 7 – Résultats de la classification des textes en cinq stades pour les trois classificateurs. Chaque classificateur a employé les 10 meilleurs attributs évalués avec le critère du gain d'information de Weka et a été entraîné sur 317 textes du corpus.

	C4.5			SVM			LMT		
Stades	P	R	*F*	P	R	*F*	P	R	*F*
1	0,56	0,38	0,45	0,60	0,38	**0,46**	0,53	0,38	0,44
2	0,51	0,53	0,52	0,61	0,62	**0,62**	0,61	0,61	0,61
3	0,49	0,47	0,48	0,54	0,57	0,56	0,56	0,59	**0,57**
4	0,45	0,55	0,50	0,61	0,69	**0,65**	0,61	0,62	0,62
Contrôle	0,78	0,68	0,73	0,83	0,73	0,78	0,86	0,88	**0,87**

TAB. 8 – Résultats de la classification des textes en cinq stades pour les trois classificateurs. Chaque classificateur a employé les 20 meilleurs attributs évalués avec le critère du gain d'information de Weka et a été entraîné sur 317 textes du corpus.

une différence dans les chiffres de précision et de rappel pour le stade 1, cette fois-ci par rapport à la première expérience utilisant les 142 attributs (voir les tableaux 3 et 4). Alors que ces chiffres étaient relativement proches, ils sont très différents dans les deux expériences suivantes avec un rappel considérablement inférieur à la précision. Ceci signifie que la qualité du rappel dépend d'un ensemble d'attributs beaucoup plus grand pour le stade de développement le plus bas que pour les autres stades. Puisque la précision et le rappel pour les stades plus avancés sont proches dans toutes expériences menées, ceci pourrait signifier que le stade 1 est le plus hétérogène.

7 Conclusion

Dans cet article, nous avons présenté et évalué un système pour identifier des stades de développement en français langue étrangère. Le système se compose d'un analyseur morphosyntaxique et d'un module d'apprentissage automatique. Dans un premier temps, l'analyse morphosyntaxique nous permet de représenter chaque texte par un vecteur de 142 attributs. Grâce aux second module, nous avons ensuite entraîné trois classifieurs différents pour évaluer l'hypothèse qu'on pouvait partitionner les textes du corpus en cinq stades de développement. Cette démarche a permis la classification automatique d'un ensemble de 317 textes du corpus CEFLE selon le stade de développement qu'ils reflétaient.

Les résultats d'une première expérience de classification employant un vecteur contenant l'ensemble des 142 attributs ont montré une amélioration importante de plus de 10% comparés à nos résultats précédents. Pour une classification simplifiée à trois stades, la moyenne de précision et de rappel pour le système est maintenant de 79%. Dans le but d'identifier les meilleurs attributs pour la classification, nous avons introduit un critère de sélection fondée sur le gain d'information. À notre surprise, les résultats ont montré que la performance globale n'était pas affectée de façon sensible par la réduction radicale du nombre d'attributs (de 142 à 10 et 20 respectivement). Cependant les résultats pour le stade de développement le plus bas sont dégradés de façon très importante. Une interprétation possible est que les textes du stade 1 sont tellement hétérogènes qu'on doive remettre en cause l'unicité de ce niveau et qu'il serait préférable de le diviser en plusieurs sous-classes.

Remerciements

La recherche présentée dans cet article bénéficie d'un financement du Conseil suédois pour la science, contrat numéro 2004-1674, et de bourses de la fondation Elisabeth Rausing pour la recherche dans les sciences humaines et de la fondation Erik Philip-Sörenssen pour la recherche.

Références

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

BARTNING I. & SCHLYTER S. (2004). Stades et itinéraires acquisitionnels des apprenants suédophones en français L2. *Journal of French Language Studies*, **14**(3), 281–299.

BOSER B., GUYON I. & VAPNIK V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, p. 144–152, Pittsburgh : ACM.

GRANFELDT J., NUGUES P., PERSSON E., PERSSON L., KOSTADINOV F., ÅGREN M. & SCHLYTER S. (2005). Direkt Profil : un système d'évaluation de textes d'élèves de français langue étrangère fondé sur les itinéraires d'acquisition. In *Actes de la conférence Traitement Automatique des Langues Naturelles, TALN & RECITAL 2005*, volume Tome 1 – Conférences principales, p. 113–122, Dourdan, France.

GRANFELDT J., NUGUES P., ÅGREN M., THULIN J., PERSSON E. & SCHLYTER S. (2006). CEFLE and Direkt Profil : A new computer learner corpus in French L2 and a system for grammatical profiling. In *Proceedings of the 5th International Conference on Language Ressources and Evaluation*, p. 565–570, Genoa, Italy.

LANDWEHR N., HALL M. & FRANK E. (2003). Logistic model trees. In N. LAVRAC, D. GAMBERGER, L. TODOROVSKI & H. BLOCKEEL, Eds., *Proceedings of the 14th European Conference on Machine Learning (ECML)*, volume 2837 of *Lecture Notes in Computer Science*, p. 241–252. Springer.

MACWHINNEY B. (2000). *The CHILDES project : Tools for analyzing talk*. Mahwah, New Jersey : Lawrence Erlbaum.

QUINLAN J. R. (1986). Induction of decision trees. *Machine Learning*, **1**(1), 81–106.

SAGAE K., LAVIE A. & MACWHINNEY B. (2005). Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics 2005*, p. 197–2004, Ann Arbor, USA.

WITTEN I. H. & FRANK E. (2005). *Data Mining. Practical Machine Learning Tools and Techniques*. Amsterdam : Elsevier.