Faire face aux exigences de comparabilité dans l'évaluation des compétences en langue seconde

Isabelle Monnard, Chantal Tièche Christinat

L'enseignement traditionnel des langues ne produisant pas les résultats escomptés, diverses solutions sont volontiers proposées pour tenter d'y remédier. Cependant, il est très difficile de comparer la pertinence et le succès de ces diverses propositions, car on ne dispose pas d'instruments de mesure aisément transposables.
Deux obstacles majeurs peuvent être repérés. Premièrement, comment comparer les performances d'élèves recueillies dans des circonstances et par des procédures d'évaluation différentes, des situations de test et des situations d'expression spontanée, par exemple? En second lieu, comment établir de l'extérieur le niveau de difficulté des énoncés compris ou produits par les élèves lorsque la nature exacte des situations travaillées en classe n'est pas la même et de plus, souvent décrite de manière insuffisamment explicite?
Cet article présente une réflexion sur ces obstacles et la solution que nous proposons pour les surmonter. Nous avons développé un instrument permettant de définir le niveau de difficulté des énoncés quel que soit le contexte d'évaluation et les situations travaillées en classe. Un premier test de cet instrument, élaboré au moyen du paradigme de la généralisabilité figure dans la seconde partie.

Introduction

L'enseignement traditionnel des langues étrangères suscite plusieurs débats tant du point de vue de la pédagogie que de celui des curricula. Le choix d'une méthode au détriment d'une autre est nécessairement déterminé par les aspects économiques, sociaux et culturels, mais s'effectue également, dans un but d'optimalisation, en fonction des résultats obtenus au moyen d'évaluations comparatives des méthodes. L'enseignement traditionnel ne produisant pas les résultats escomptés, diverses solutions sont volontiers proposées pour tenter de remédier à ce constat. Les enseignements ont été modifiés dans leur approche de la langue, dans leur contenu et/ou dans leurs objectifs. Par ailleurs le temps consacré aux apprentissages a été modifié, et l'âge de l'enfant est pris partiellement en compte dans la détermination du mode d'enseignement, par immersion ou par ensei-

gnement traditionnel. La tendance actuelle dans l'enseignement de la langue seconde tend à développer la compétence communicative des élèves, qui repose à la fois sur le pôle réceptif et sur le pôle expressif. L'évaluation peut donc se réaliser dans quatre modalités: la compréhension orale et écrite d'une part, la production orale et écrite de l'autre. En classe, les évaluations sont construites en fonction des objectifs du programme poursuivi, des thèmes abordés, et, généralement, des types d'exercices proposés par les manuels utilisés. La compréhension écrite et orale de dialogues, de textes, de syntaxe, du lexique par exemple, est évaluée au moyen d'épreuves classiques: questionnaires ouverts ou à choix multiples, tests de closure, appariements, choix d'image en fonction d'une histoire, etc. Les critères de correction sont objectifs, une réponse étant obligatoirement juste ou fausse. En production, les élèves sont amenés à produire des textes, (par exemple une lettre, une description), à parler dans une situation donnée (par exemple créer un dialogue sur un thème précis à partir d'un message d'un répondeur téléphonique, s'entretenir en conversation libre, prendre position sur un sujet donné) ou à traduire des mots et des énoncés. Pour noter ces productions orales ou écrites, des grilles d'analyse sont en général utilisées. Les critères d'appréciation peuvent concerner la correction des énoncés, la quantité et la qualité des informations, la compréhensibilité, la prononciation, la fluidité, le choix des mots, la structure des énoncés.

En l'état, il est cependant très difficile de comparer la pertinence et le succès de ces diverses propositions, car on ne dispose pas d'instruments de mesure aisément transposables. Deux obstacles majeurs peuvent être repérés. En premier lieu apparaît la nécessité de s'interroger sur le bien-fondé des comparaisons mettant en jeu des performances d'élèves recueillies dans des circonstances et par des procédures d'évaluation différentes, par exemple des situations de tests et des situations d'expression spontanée. En second lieu, on doit se demander comment établir, de l'extérieur, le niveau de difficulté des énoncés compris ou produits par les élèves alors que la nature exacte des situations travaillées en classe n'est pas la même et qu'elle est le plus souvent décrite de manière insuffisamment explicite.

Forces et faiblesses des tests

Sans mettre en cause la valeur de ces épreuves dans le cadre des pratiques courantes d'évaluation en classe, deux questions se posent néanmoins aux chercheurs en éducation qui tentent de comparer les résultats de diverses situations d'enseignement et d'évaluations variées.

La première question évidente consiste à se demander si on peut utiliser les mêmes épreuves dans des situations d'enseignement différentes. À cette question, la réponse est généralement négative. En effet, les épreuves étant conçues en fonction des programmes, des manuels et des objectifs, il paraît évident d'un point de vue méthodologique qu'on ne peut utiliser les mêmes épreuves dans des contextes différents.

La seconde question, corollaire de la précédente, consiste à s'interroger sur

la pertinence des comparaisons de résultats obtenus au moyen d'épreuves différentes. Une réponse à cette préoccupation est discutée dans une étude réalisée par des étudiants du Séminaire des langues romanes de l'Université de Bâle (CDIP, 1993). Ce travail présente l'analyse d'une batterie de tests conçue pour permettre aux enseignants d'évaluer en classe les performances des élèves en allemand et en français langue 2 (L2) à la fin de la scolarité obligatoire. Cette batterie couvre les quatre domaines de la compétence communicative et présente un éventail très large de tests comprenant une grande partie des épreuves citées plus haut. De plus, elle est voulue comme une activité sensée, aussi proche de situations réelles que possible. Les analyses montrent que les degrés de difficulté diffèrent fortement d'une épreuve à l'autre et d'un item à l'autre. Par ailleurs, on observe des divergences dans les modes d'évaluation des enseignants, particulièrement en production orale. En effet, certains enseignants répartissent les élèves selon une courbe de Gauss, d'autres les rassemblent vers la note-seuil. L'évaluation dépend également des caractéristiques des enseignants. Plus particulièrement, les normes de correction, les convictions pédagogiques, les humeurs, l'environnement langagier, le choix du matériel didactique sont évoqués. Pour les auteurs du rapport, les critères d'évaluation (correction, liaison, compréhensible, fluidité, prononciation, stratégies propres), ne sont pas toujours pertinents et sont parfois flous, les niveaux de difficultés ne sont pas assez différenciés et semblent liés à une situation-test donnée ce qui leur confère une validité restreinte. Ce rapport de la CDIP montre de surcroît et de façon très nette que le problème majeur de l'évaluateur se situe au niveau de la production. Cependant cette difficulté n'est pas propre à l'évaluation seule, mais est partagée par la situation d'enseignement. En effet, la recherche en langue seconde au même titre que l'expérience commune ont clairement démontré qu'il est plus difficile de parler que de comprendre une langue seconde. Les types d'enseignement reflètent ce décalage, certains insistant davantage sur la compréhension alors que d'autres visent avant tout la capacité de production des élèves. De plus, dans certains travaux d'évaluation, les compétences sont mesurées en situation classique de test alors qu'elles sont évaluées dans d'autres dans le cadre de situations d'expression ou d'activités spontanées, qui malgré les précautions prises pour rendre les situations de tests «naturelles» et proches du vécu des élèves, conservent nonobstant un caractère artificiel. Chez de jeunes élèves, les tests peuvent inhiber leurs comportements langagiers et amener des résultats inférieurs aux apprentissages réalisés. Enfin, le caractère spécifique des énoncés à comprendre ou produire dans un test rend difficile toute généralisation. En résumé, on voit que les résultats des élèves varient en fonction des types d'épreuves, des critères choisis et de la manière d'évaluer des enseignants.

Malgré le peu d'études de ce type, le sérieux et la pertinence des analyses précédemment citées nous autorisent à affirmer, en réponse à notre seconde question que, de manière générale, les tests d'évaluation de la langue 2 ne sont pas comparables. En répondant négativement à nos deux questions initiales, il s'avère donc qu'on ne dispose pour l'heure, d'aucun outil permettant de comparer divers

modèles d'enseignement et des niveaux de compétence en langue 2.

Evaluation des compétences au moyen de grilles de niveaux

Pour parvenir à une meilleure appréciation de ce que les enfants savent faire en langue 2, la nécessité de prendre également en considération les productions spontanées semble incontournable aux yeux des évaluateurs qui ont développé des grilles d'analyse particulières pour faire face à ces situations de production. Les plus simples évaluent des compétences globales (capacité de comprendre des messages simples, de répondre à des questions, de décrire une image, de répéter des comptines, etc.) en rapport avec les objectifs généraux du programme. De telles grilles laissent une grande part à la subjectivité des enseignants et ne permettent pas d'évaluer les apprentissages de façon précise.

Certaines grilles, plus précises, se centrent sur un aspect de la production. C'est le cas par exemple de celle utilisée dans la recherche sur l'école maternelle bilingue en Vallée d'Aoste (Bourguignon, Py & Ragot, 1994). L'unité choisie est la complexité syntaxique: syntagme isolé, combinaison monosyntagmatique ou polysyntagmatique, énoncé complexe. Cet outil est certes intéressant, mais il sert avant tout à caractériser le langage des enfants. Les auteurs admettent d'ailleurs que le choix de la complexité syntaxique «n'est peut-être pas le meilleur descripteur de la compétence d'un jeune enfant dans l'apprentissage d'une langue seconde (les enfants de 5-6 ans ne produisent pas beaucoup d'énoncés complexes et corrects en langue maternelle)» (Bourguignon, Py & Ragot, 1994, p. 61). Des grilles beaucoup plus complexes ont été conçues. Ainsi la grille BB96 développée en Alsace (Association ABCM-Zweisprachigkeit, 1996) évalue de façon très détaillée la production orale: intonation, accent de mot et de phrase, articulation des différents phonèmes, respect des conventions phonotactiques, etc. Cette grille est utilisée pour une évaluation collective de la classe; elle est d'usage très complexe et nécessite des connaissances linguistiques poussées.

En se centrant sur les compétences langagières en situation de communication, les grilles d'évaluation pour l'oral (Schneider & North, 1997) permettent une autoévaluation comportant plusieurs repères et définissant divers niveaux. Ces repères portent sur des aspects linguistiques formels comme l'aspect syntaxique observé sous l'angle de la correction grammaticale et l'aspect lexical repéré par l'usage des expressions idiomatiques, sur des aspects pragmatiques comme la familiarité et la fréquence conversationnelle, sur des aspects prosodiques comme l'intonation et le débit et également sur des aspects socio-communicatifs comme l'aisance, la reformulation et les signaux de compréhension. Si la prise en compte de critères communicatifs et linguistiques nous paraît intrinsèquement nécessaire lors de l'évaluation des compétences langagières, l'absence de critères définis et stables pouvant être appliqués aux différents aspects précités rend difficile l'utilisation d'une telle grille.

Quels que soient le degré de complexité de ces grilles et la nature des obser-

vations qu'elles recueillent, comparer des résultats relevés au moyen de celles-ci et ceux provenant d'un test demeure difficile ; en particulier, on ne trouvera généralement dans une grille que les cas de «réussites», alors qu'un seul échec s'observe déjà clairement dans un test.

Dans le cadre d'une demande d'évaluation d'un enseignement bilingue, par immersion, nous avons été confrontées à ces questionnements et l'évocation de ces deux obstacles nous a incitées à élaborer une tentative de solution consistant en un instrument permettant de définir le niveau de difficulté des énoncés quels que soient le contexte d'évaluation et les situations travaillées en classe. Un premier test de cet instrument, élaboré au moyen du paradigme de la généralisabilité (Cardinet & Tourneur, 1985), figure dans la seconde partie. L'instrument que nous avons élaboré, et qui sera décrit ci-dessous, se veut un moyen d'esquisser une comparaison entre résultats recueillis au moyen de ces différents outils, mais ne remplace ni les uns ni les autres. Il permet par contre de situer sur une échelle commune, graduée en termes de niveaux de difficulté, toute production ou comportement langagier des élèves, indépendamment de la situation d'évaluation.

Etablissement du niveau de difficulté des productions dans une langue en cours d'apprentissage

A la différence des enseignements traditionnels d'une langue seconde, les enseignements par immersion ne font pas de la langue cible un objet d'étude, mais un moyen d'acquérir à travers elle des connaissances extralinguistiques, en même temps qu'une familiarisation de plus en plus grande avec l'outil. Pour l'évaluateur, l'objectif n'est plus de savoir quelle proportion de structures linguistiques et de mots de vocabulaire travaillés en classe a été acquise, mais de déterminer quel niveau de maîtrise de l'outil est atteint. Or une telle évaluation est loin d'être triviale; si, dans un enseignement traditionnel, il est relativement facile pour l'enseignant de moduler et d'incrémenter progressivement le niveau de difficulté des énoncés proposés ou demandés à l'élève, dans un enseignement par immersion, une telle progression est pratiquement impossible à assurer. Impossible ainsi, par exemple, de ne parler d'abord que par affirmation, sous prétexte qu'une telle structure est plus simple à comprendre et à produire. Et si des structures telles les négatives ou les interrogatives sont introduites rapidement et utilisées fréquemment, peut-on toujours statuer de leur plus grande complexité pour les élèves ou faut-il parler de complexité relative? Ainsi la complexité de la structure syntaxique comme seul et unique critère, ne suffit dès lors plus à établir le véritable niveau de difficulté des énoncés maîtrisés par les élèves. Pour pallier cet inconvénient, nous avons choisi de prendre en considération dans notre échelle quatre indices indépendants dont voici une brève présentation.

Indices de complexité d'un énoncé

1. La *caractéristique structurale* – ou type – de l'énoncé telle qu'elle a été définie par la grammaire générative (Chomsky, 1965). Tout énoncé, compris ou produit, se verra conférer un premier niveau de difficulté sur la base du critère de transformation de la phrase-noyau P introduit dans le modèle chomskien. Miller (1962) a suggéré que le passage d'une phrase à l'autre représente une opération plus complexe, selon le nombre de transformations qu'elle nécessite. Une phrase simple active affirmative déclarative (SAAD) telle que «le chat boit du lait», n'ayant subi aucune transformation syntaxique ni morphosyntaxique est a priori plus facile que la phrase interrogative «le chat boit-il du lait» ou «est-ce que le chat boit du lait» et la phrase négative «le chat ne boit pas de lait». Brown et Hanlon (1970) ont montré dans l'étude d'un corpus enfantin que l'ordre d'émergence des phrases respecte leur degré de complexité, alors que Maratsos et Kuckaj (1978) ont abouti à des résultats opposés. En 1980, quelques psycholinguistes tels Erreich proposent encore une théorie de l'acquisition de la syntaxe où l'enfant à partir des données observées dans le langage construit des hypothèses sous forme de règles transformationnelles. La psycholinguistique actuelle considère que la contextualisation des phrases et l'aspect pragmatique modifient les résultats obtenus durant les années 1960, et il est nécessaire de considérer la syntaxe comme étant non-autonome des conditions d'énonciation. Cet unique critère est donc largement insuffisant à décrire une situation de production. Par ailleurs l'acquisition des divers types de phrases n'est pas uniquement régie par leur statut grammatical, mais également par leur valeur fonctionnelle immédiate.

2. La longueur moyenne d'un énoncé, comptée en mots ou en morphèmes. Lors de l'apprentissage d'une langue première ou seconde, la longueur moyenne des énoncés maîtrisés tend en effet à s'accroître avec les compétences. Cette mesure permet de se centrer sur les significations et pour Brown (1973) cet indice exprime une meilleure prédictivité de la complexité linguistique des productions que l'âge du locuteur. Toutefois cet auteur souligne qu'au-delà de 4 morphèmes par énoncé, la longueur moyenne de l'énoncé produit par l'enfant est plus fonction de la qualité de l'interaction, que de ses connaissances verbales. Il apparaît clairement qu'à elle seule, la longueur d'un énoncé ne suffit pas non plus à en définir le niveau de difficulté. Une phrase comprenant plusieurs mots connus et dont la structure phonologique est simple n'est pas *a fortiori* plus difficile qu'une phrase courte comprenant des mots rares et/ou difficiles sur le plan articulatoire. Par ailleurs dans les formules de lisibilité des textes, la longueur moyenne des phrases constitue certes une variable importante qui offre l'avantage d'être un outil très simple, mais elle reste cependant très critiquable. En effet, comme l'ont montré Kintsch et Vipond (1979), un texte mieux structuré comprenant des connecteurs et dont les phrases sont plus longues est mieux compris dans son ensemble qu'un texte comprenant une succession de phrases courtes.

3. Le vocabulaire impliqué dans ces énoncés peut également influencer le niveau de difficulté de ceux-ci. On peut globalement estimer que les mots les plus faciles sont ceux que l'on rencontre dans la langue avec la plus grande *fréquence*,

alors que les mots rarement entendus puisque peu fréquents, sont plus difficiles à maîtriser. Toutefois, la seule connaissance lexicale des mots (test de vocabulaire par exemple) ne peut rendre compte de la compétence acquise par l'enfant. En effet, un mot évoqué seul ne relève que de la compétence mnésique de l'enfant. La capacité langagière relève quant à elle d'autres mécanismes tel le mécanisme d'inférence de sens (reconstruction du sens à l'aide des autres mots et du sens général plausible).

4. Pour une langue seconde, et qui plus est, après une période de familiarisation relativement brève où la variété et le nombre de mots perçus sont peu étendus, ce ne sont pas nécessairement les mots les plus fréquents dans la langue en question qui ont été le plus souvent prononcés et qui donc devraient être les plus faciles. Certains mots rares dans la langue peuvent fort bien avoir été en classe d'un emploi relativement fréquent, alors que d'autres, pourtant fréquents dans la langue, peuvent n'avoir jamais été prononcés devant l'enfant. En conséquence, nous avons choisi de tenir compte également du degré de *familiarité* de l'énoncé ou des mots le composant, suivant pour cela les indications fournies par les enseignants. Prises isolément, la familiarité des mots et la fréquence d'usage de ceux-ci dans les différentes situations travaillées en classe ne donnent cependant pas non plus un aperçu exact de la compétence en L2 de l'enfant. Si elles permettent de voir ce que l'enfant a retenu de ce qui a été fait en classe et indiquent éventuellement la compétence de reproduction dans une situation donnée, elles ne renseignent pas sur ce que l'enfant a réellement construit en L2.

Calcul du niveau de difficulté des énoncés

Compte tenu des remarques qui précèdent, nous avons décidé de combiner les indices de structure, de longueur, de fréquence et de familiarité des énoncés pour calculer un niveau unique de difficulté. Ce niveau de difficulté des énoncés est fixé par combinaison des quatre indices, répartis en trois niveaux en fonction des critères, selon le schéma présenté dans le tableau 1.

Si l'on additionne alors les niveaux de difficulté atteints par un énoncé sur les quatre indices ci-dessus, le résultat peut aller de 4 (niveau 1 sur les 4 indices) à 12 (niveau 3 sur les 4 indices). Une note de difficulté est attribuée en fonction du total atteint: de 1 pour un total de 4 à 3 pour un total de 12. On peut alors calculer le niveau de difficulté de n'importe quel énoncé. Pour illustrer le fonctionnement de cette grille, nous présentons quelques exemples de cotation d'énoncés (tableau 2).

La phrase «Ich kaufe Orangen» est d'une longueur de niveau 1 (3 mots), de type 1 (déclarative simple), de fréquence 1 (mots fréquents) et de familiarité 1 (énoncé souvent utilisé en classe), soit un total de 4 et donc une note de difficulté égale à 1. «Willst du mir Trauben kaufen» est d'une longueur de niveau 2 (5 mots), de type 3 (interrogative ne commençant pas par un mot interrogatif), de fréquence 2 (un mot peu fréquent) et de familiarité 3 (si l'énoncé n'a jamais utilisé en classe), soit un total de 10 qui correspond à une note 7.

Tableau 1: Indices, niveaux et caractéristiques utilisés pour établir le niveau de

difficulté d'un énoncé en langue 2

Indice	Niveau	Critère
Longueur de l'énoncé	**1**	énoncés courts (3 mots au plus)
	2	énoncés de longueur moyenne (4 ou 5 mots)
	3	énoncés longs (6 mots et plus)
Type de la phrase	**1**	phrases déclaratives ou impératives simples
	2	phrases déclaratives complexes, interrogatives et négatives simples
	3	autres types de phrases
Fréquence des mots utilisés	**1**	tous les mots sont généralement d'usage fréquent
	2	un ou deux mots de l'énoncé sont d'usage moins fréquent
	3	certains mots sont relativement rares dans la langue
Familiarité	**1**	l'énoncé a été souvent utilisé en classe ou tous les mots qu'il comporte sont bien connus des élèves
	2	l'énoncé a été utilisé en classe, mais pas très fréquemment ou certains mots qu'il comporte ne sont que moyennement connus des enfants
	3	l'énoncé n'a pratiquement jamais été utilisé en classe ou les mots qu'ils comportent sont peu connus des enfants

Tableau 2: Illustration de cotation de phrases et établissement du niveau de difficulté

	Longueur	Type	Fréquence	Familiarité	Total	**Note**
Ich kaufe Orangen.	1	1	1	1	4	**1**
Was kostet eine Orange ?	2	2	1	1	6	**3**
Willst du Orangen kaufen ?	2	3	1	1	7	**4**
Willst du Orangen oder Bananen kaufen ?	3	3	1	2	9	**6**
Willst du mir Trauben kaufen ?	2	3	2	3	10	**7**
Ich will heute keine Traube essen.	3	3	3	3	12	**9**

On peut alors, si on le souhaite, déterminer pour chaque énoncé un niveau de difficulté global en divisant en trois segments égaux cette différence; on obtient alors la classification présentée dans le tableau 3.

Tableau 3: Calcul du niveau global de difficulté d'un énoncé en langue 2

Niveau	Critères d'évaluation
I	tout énoncé dont la somme des niveaux atteints sur les quatre indices varie entre 4 et 6 (note 1 à 3)
II	tout énoncé dont la somme des niveaux atteints sur les quatre indices varie entre 7 et 9 (note 4 à 6)
III	tout énoncé dont la somme des niveaux atteints varie entre 10 et 12 (note 7 à 9).

Dans le cas des exemples cités, «Ich kaufe Orangen» est donc d'une difficulté globale de niveau I et «Willst du mir Trauben kaufen?» de niveau III.

Etude statistique de l'instrument

Pour évaluer les qualités de l'instrument développé, il était nécessaire, d'une part, de vérifier la précision des critères pour chacun des quatre indices et, d'autre part, de voir si les niveaux de difficulté calculés correspondent bien à des difficultés rencontrées par les enfants.

Fidélité interjuges

Dans le cadre de l'évaluation d'expériences d'enseignement bilingue précoce (Gurtner, Monnard & Tièche Christinat, 1996), des tâches de compréhension, de production et de répétition en langue 2 ont été soumises à des enfants de classe enfantine (5-6 ans). L'épreuve comprenait au total 30 énoncés dont le niveau de difficulté a été apprécié indépendamment au moyen de notre grille d'analyse par trois juges formés à son utilisation. La familiarité des mots et des énoncés a été évaluée à partir des listes fournies par l'enseignante.

Une étude de généralisabilité sur les valeurs obtenues a alors été menée pour contrôler la fidélité interjuges et déterminer dans quelle mesure on peut faire confiance au jugement d'une personne unique utilisant la grille. Basé sur l'analyse de variance, le modèle de la généralisabilité permet notamment de vérifier la fiabilité d'instruments d'évaluation en testant dans quelle mesure on peut généraliser résultat à l'ensemble des observations possibles à partir des valeurs observées (Cardinet & Tourneur, 1985).

Le plan d'observation comporte trois facettes: les énoncés (E), les juges (J) et les indices (I). Nous avons considéré que l'univers des phrases possibles et des juges peut être infini. Par contre, l'univers des indices ne peut être élargi et nous avons donc retenu 4 niveaux correspondant aux 4 indices décrits. Le plan de mesure est entièrement croisé: chacun des 30 énoncés est évalué par chacun des juges sur chacun des 4 indices. L'étude visant à vérifier la fidélité interjuges, le plan de mesure comprend la facette E sur la face de différenciation et les facettes J et I sur la face d'instrumentation (E/JI).

Les résultats obtenus montrent que les critères choisis pour définir chacun des indices sont suffisamment précis puisque le coefficient de généralisabilité relatif atteint .95. La variabilité due aux interactions juges/énoncés et juges/indices est faible. On peut en conclure que des juges indépendants donnent bien les mêmes notes aux énoncés en fonction des critères. Toutefois, en situation normale d'utilisation de la grille, une seule personne serait amenée à utiliser la grille pour évaluer le niveau de difficulté d'énoncés. Pour calculer la fidélité de l'outil dans cette situation, le coefficient obtenu doit être corrigé en demandant dans le plan d'optimisation le coefficient pour un seul juge. Comme le coefficient de généralisabilité atteint dans ce cas .86, il apparaît qu'on peut avoir confiance dans le jugement d'un seul juge.

Parmi les quatre indices utilisés dans la grille, deux laissent une place à l'interprétation: la fréquence et la familiarité des mots. Ces deux indices ont donc été traités séparément. Le plan de mesure ne comportait alors plus que deux facettes: les juges et les indices. Si le critère de familiarité peut être jugé comme satisfai-

sant (coefficient de généralisabilité = .86), ce n'est pas le cas pour la fréquence des mots. Le coefficient obtenu pour ce critère est égal à .63. L'interaction entre les juges et les énoncés représente une source de variation importante (61%). Il semble donc que cet indice ne soit pas assez précis et qu'il laisse une trop grande place à la subjectivité. Toutefois, dans l'évaluation globale du niveau de difficulté, ces inexactitudes sont atténuées par la fiabilité des autres indices.

Pour augmenter encore la précision de la grille, lors d'une seconde expérience similaire à la première (Gurtner, Monnard & Tièche Christinat, 1996), nous avons coté les énoncés en utilisant des listes de référence (vocabulaire de base, français-allemand fondamental) pour évaluer la fréquence des mots.

La fidélité interjuges calculée sur les 4 indices est à nouveau très satisfaisante; le coefficient de généralisabilité atteint .96. Pour un juge unique, le coefficient est égal à .90. La fidélité sur le critère fréquence a bien été améliorée par l'utilisation d'une liste de référence, elle atteint ici .85, tandis que la variation due à l'interaction entre les juges et les phrases a fortement diminué (32%). La fidélité des mesures de familiarité par contre est légèrement inférieure lors de cette deuxième cotation. Atteignant .80, elle reste toutefois satisfaisante. Il faut préciser que, pour cette seconde expérience, les données fournies par l'enseignante concernant les mots et énoncés utilisés en classe étaient très lacunaires. Il semble donc que même sans connaître précisément le contexte scolaire, il soit possible d'évaluer assez fidèlement la familiarité d'un énoncé.

Cohérence interne des indices

Concernant la définition d'une difficulté des énoncés, on peut se demander dans quelle mesure les quatre indices sont cohérents entre eux et dans quelle mesure la somme des quatre indices est corrélée avec le niveau global de difficulté qu'elle définit. Le coefficient alpha de Cronbach a été calculé pour mesurer cette cohérence interne. Pour éliminer l'influence des juges, la moyenne des trois juges a été prise en compte pour chacun des 60 items des deux expériences décrites. On constate que le dispositif présente une bonne cohérence puisque l'*alpha* atteint .80.

Valeurs observées selon les trois niveaux de difficulté

Dans le cadre de l'évaluation d'une expérience d'enseignement bilingue (Gurtner, Monnard & Tièche Christinat, 1996), plusieurs tests ont été soumis à des élèves d'école enfantine: traduction français-allemand et allemand-français, répétition d'énoncés en allemand. Au total, 2 séries de 35 items ont été proposés à 14 et 17 enfants respectivement. Les résultats obtenus confirment les étalonnages établis (tableau 4).

Tableau 4: Pourcentages de réussite aux items en fonction du niveau de difficulté des indices et du niveau global de difficulté

Indices	Niveau I	Niveau II	Niveau III
Longueur	55	53	27
Type de phrase	57	38	32
Fréquence	49	47	18
Familiarité	51	57	25
Niveau global de difficulté	58	49	24

Le niveau de difficulté de chaque indice défini selon les critères établis est plus ou moins lié aux difficultés rencontrées par les enfants, le niveau global de difficulté calculé correspond généralement mieux aux résultats observés (tableau 5).

Tableau 5: Pourcentages de réussite aux différents tests en fonction du niveau global de difficulté

	Niveau global de difficulté		
	Niveau I	Niveau II	Niveau III
Allemand-français	68	41	34
Allemand-allemand	79	55	18
Français-allemand	25	11	3

Une analyse de variance montre que la difficulté des énoncés des trois niveaux est significativement différente ($F = 8.96$, $p<.001$). Quel que soit le type de test (traduction ou répétition), les items de niveau I sont ceux qui sont le mieux réussis par les élèves; à l'inverse, les items de niveau III obtiennent les plus faibles pourcentages de réussite. Aucun enfant ne produit/comprend d'énoncé de niveau II (respectivement III) s'il n'a pas réussi la majorité des énoncés de niveau I (respectivement II). Même s'ils doivent encore être considérés avec prudence, ces résultats ouvrent donc des perspectives intéressantes pour les études de comparabilité.

Nous avons réalisé une analyse de régression pour voir dans quelle mesure l'utilisation de la grille permet de prédire la difficulté d'un énoncé particulier. La régression des critères longueur, fréquence, type de phrase, familiarité et niveau global de difficulté sur le nombre de réussites est significative (valeur de $F = 2.88$, $p<.05$); le niveau global de difficulté a un poids au moins cinq fois plus grand que chacune des autres variables. Dans une régression pas à pas, seule cette variable reste dans l'équation et explique à elle seule 17% du nombre de réussites ($F = 14.3$, $p<.001$). C'est aussi cette variable qui est la plus fortement corrélée avec le nombre de réussites.

Discussion

L'instrument que nous avons élaboré cherche à dépasser les contraintes et les difficultés intrinsèques à la situation de comparaison des systèmes d'enseignement de L2 du point de vue de leur efficacité. Conçu pour résoudre les difficultés liées à la comparaison des évaluations, il ne prétend en aucun cas remplacer les instruments d'évaluation utilisés dans les classes. Il pourrait par contre être utilisé pour discuter le niveau de difficulté des énoncés que comportent les épreuves habituelles et de ce fait constitue un instrument pouvant juger de la complexité des instruments établis. Il a pour objectif de se situer en dessus des pratiques pédagogiques d'enseignement et d'évaluation tout en étant un dénominateur commun annulant les effets indésirables des différences propres aux méthodes d'enseignement, aux activités en classe, voire même aux situations d'évaluations retenues d'une expérience à l'autre. Annulant dès lors l'effet des différences, il facilite des comparaisons inter et intragroupe. Cet instrument travaille à la manière d'une grille d'analyse, mais débouche sur une estimation unique du niveau de difficulté de chaque énoncé par composition de deux critères structuraux (longueur et type grammatical des énoncés) et deux critères d'usage (simplicité du lexique et familiarité des situations évoquées). Cette combinaison de critères, plutôt que le choix d'un seul donne à l'ensemble, comme nous l'avons vu, une bonne fiabilité générale pour l'ensemble de l'épreuve.

L'instrument comporte quelques faiblesses qu'il faudra tenter d'éliminer. Ainsi dans le cas de la compréhension d'un énoncé, un seul mot peut parfois permettre de comprendre le sens général sans qu'il y ait obligatoirement recours à l'ensemble et à la combinaison des indices qui définissent la difficulté de l'énoncé.

Il est apparu au cours de l'analyse de régression que certains énoncés avaient été cotés trop faciles au vu des résultats obtenus. Il s'avère en effet que la variable psychologique du stress ou de l'inquiétude n'est pas neutralisée vu le nombre trop restreint d'élèves testés. Cette variable peut influencer les résultats en rendant difficile ce qui est posé comme facile. En effet, la situation de test était effectuée dans des conditions de certification quasi examinales, ce qui pouvait créer chez les jeunes enfants une inquiétude, et dans une situation qui tout en cherchant à créer un contexte conversationnel agréable est demeurée néanmoins une situation nouvelle qui les a rendus moins performants qu'en situation naturelle. La fiabilité de cet outil sera d'autant meilleure que l'application sera menée en situation naturelle ou pour un échantillon de population suffisamment large.

Il nous paraît également important de discuter l'omission volontaire dans cet instrument, d'un certain nombre de variables qui a été pris parfois en compte dans d'autres tests d'évaluation de L2 (Bourguignon, Py & Ragot, 1994 ; Bregy, Brohy & Fuchs, 1996 ; Association ABCM, 1996). La situation examinale introduit les enfants dans une situation totalement nouvelle, tant par le lieu que par l'examinateur. Nous nous trouvons dans une situation de contrat différent de l'enseignant et la finesse de l'analyse sera d'autant plus grande que nous pourrons neutraliser au mieux cette variable en appliquant par exemple un protocole

précis de passation. D'autre part, le phénomène de décontextualisation de nos énoncés n'a pas été pris en compte, alors que nous pensons que la reconnaissance d'un contexte, ou l'analogie contextuelle facilite la tâche. Dès lors, notre instrument pourrait se limiter à la première phase d'apprentissage d'une langue étrangère au terme de laquelle on attend que l'élève puisse produire des phrases isolées. D'autres phénomènes se situant dans le cadre de l'analyse pragmatique du langage, telle la qualité de l'information, la familiarité de l'énonciateur et les présupposés qui en découlent ne sont pas pris en compte dans cet instrument. L'instanciation de ces variables dans les items expérimentaux modifierait peut-être les performances des élèves testés, mais pas nécessairement la fiabilité générale de l'instrument. On pourrait rajouter un indice d'insertion textuelle et insérer les items dans une suite dialogique ou dans une suite narrative par exemple. De ce fait, le niveau de difficulté serait fixé par combinaison non plus de quatre mais de cinq indices. Toutefois, une telle opération rendrait le maniement de l'instrument plus lourd pour un gain en fiabilité vraisemblablement très faible et la cohérence des indices pris en compte en serait amoindrie. Il nous paraît qu'un instrument qui mesure de manière cohérente un premier niveau de compétence langagière est préférable à un instrument moins homogène dans lequel un critère diffère conceptuellement des autres.

Bien que notre préoccupation première ait été la comparabilité dans l'évaluation d'une langue seconde, nous tenons à spécifier que cet instrument n'a pas pour fonction d'analyser en profondeur les acquisitions linguistiques des apprenants ; en effet les indices linguistiques que nous avons retenus sont grossiers et peuvent ou doivent être affinés selon l'usage auquel on le destine. Cet outil répond à l'exigence de comparabilité des épreuves d'évaluation et ce, aussi bien dans le contexte restreint d'une classe que dans des contextes aussi vastes que les *surveys* nationaux ou internationaux. A l'usage des chercheurs en pédagogie, il offre en plus un canevas de réflexion permettant la construction d'instruments fiables d'évaluation multicritériée ainsi que l'interprétation des résultats. Conçu avant tout pour la recherche, un tel instrument pourrait même bien dépasser ce cadre et servir à l'enseignement. En l'état actuel, il offre en effet l'avantage d'être d'un usage simple et rapide et ne nécessite aucune connaissance linguistique préalable. Il pourrait donc être utilisé par les enseignants eux-mêmes pour réguler le niveau de difficulté des expressions qu'ils emploient dans leur enseignement ou pour permettre aux élèves de constater les progrès qu'ils effectuent ou les éventuelles limitations qu'ils tendraient à donner eux-mêmes à leurs expressions en langue 2.

Références

Association ABCM-Zweisprachigkeit (1996). *BB96*. Communication personnelle, Deuxièmes

rencontres intersites à propos de l'apprentissage bilingue, Aoste, 27-29 mars 1996.
Bourguignon, C., Py, B. & Ragot, A.-M. (1994). *Recherche sur l'école maternelle bilingue en Vallée d'Aoste. Aspects psycholinguistiques.* Aoste: IRRSAE.
Bregy, A.-L., Brohy, C. & Fuchs, G. (1996). *Evaluation de l'expérience d'apprentissage bilingue de Sierre 1994/95* . Neuchâtel: IRDP.
Brown, R. (1973). *A first language*. London: G. Allen & Unwin.
Brown, R. & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In R. Hayes (Ed.), *Cognition and the development of language. (pp. 155 - 207)* New York: Wiley and Sons.
Cardinet, J. & Tourneur, Y. (1985). Assurer la mesure. Berne: Peter Lang.
CDIP [Conférence suisse des directeurs cantonaux de l'instruction publique] (1993). *Analyse des tests utilisés pour évaluer les performances des élèves en allemand et en français L2 à la fin de la scolarité obligatoire*. Berne: CDIP.
Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, Mass.: MIT Press.
Erreich, A., Valian, V. & Winzemer, J. (1980). Aspects of the theory of language acquisition. *Journal of Child Language, 7,* 157-179.
Gurtner, J.-L., Monnard, I. & Tièche Christinat, C. (1996). *Enseignement d'une langue seconde à l'école enfantine. Évaluation scientifique des expériences fribourgeoises de Villars-sur-Glâne et de Morat 1994-1995.* Neuchâtel: IRDP.
Kintsch, W. & Vipond, D. (1979). Reading comprehension and readibility in educational practice and psychological theory. In L. G. Nilsson (Ed.), *Perspectives on Memory Research* (pp 329-365). Hillsdal: Erlbaum.
Maratsos, M. P. & Kuczaj, S. A. (1978). Against the transformalist account: a simpler analysis of auxiliary overmarkings. *Journal of Child Language, 5,* 81-99.
Miller, G. (1962). Some psychological studies of grammar. *American Psychologist, 17,* 748 - 762.
Schneider, G. & North, B. (1997). Portfolio européen des langues pour jeunes et adultes. In *Conseil de l'Europe (Ed.), Portfolio européen des langues. Propositions d'élaboration* (pp. 75-88). Strasbourg: Conseil de l'Europe.

Die Anforderungen der Vergleichbarkeit in der Evaluation der Kompetenzen in der Zweitsprache bewältigen

Zusammenfassung

Da der traditionelle Sprachunterricht nicht das erwartete Ergebnis erbracht hat, werden gerne verschiedene Lösungen als Abhilfe vorgeschlagen. Es ist jedoch sehr schwierig, den Wert und den Erfolg dieser verschiedenen Vorschläge zu vergleichen, da man nicht über leicht übertragbare Messinstrumente verfügt.

Dabei sind insbesondere zwei Hindernisse zu beachten: Erstens stellt sich die Frage, wie es möglich ist, Leistungen von Schülern zu vergleichen, die in unterschiedlichen Situationen und nach unterschiedlichen Bewertungsverfahren ermittelt wurden, wie zum Beispiel Testsituationen oder spontane Ausdrucksfähigkeit. Zweitens ergibt sich das Problem, aussenstehend den Schwierigkeitsgrad der von den Schülern verstandenen oder erzeugten Aussagen festzulegen, wenn die Art und Weise der in den Klassen eingearbeiteten Situationen nicht die gleiche, und zudem oft ungenügend präzise geschildert ist?

Dieser Artikel stellt eine Überlegung bezüglich dieser Hindernisse dar und die Lösung, die zu deren Überwindung beitragen könnten. Wir haben ein Mess-instrument entwickelt, um den Schwierigkeitsgrad der Äusserungen zu definieren, unabhängig vom Kontext der Evaluation und der in der Klasse erarbeiteten Situationen. Ein erster Test dieses Instrumentes, ausgearbeitet mittels des Paradigmas der Generalisierbarkeit, ist Gegenstand des zweiten Teils.

Mastering the Requirements of Equivalency in the Evaluation of Competency in a Second Language

Summary

As traditional language teaching does not produce the expected results, various solutions are readily proposed as a remedy. However, due to a lack of transposable measuring instruments, comparison of these various proposals' results is very difficult.

Two major impediments can be pointed out. First, how should pupils performances collected in different circumstances and by different types of evaluations be compared? Second, how can one objectively establish the difficulty level of statements understood and produced by pupils when classroom the situation differs, and moreover, are often described in an insufficiently explicit way?

This paper evaluates those difficulties and the solutions proposed to overcome them. A specific instrument was also developed that allows a definition of the difficulty level of statements, in accordance with the evaluations context and the classroom situation. A first test of this instrument, elaborated by means of the generalisation paradigm, appears in the second part.

Come far fronte alle esigenze di comparabilità nella valutazione delle competenze nell'apprendimento delle lingue seconde

Riassunto

Poichè l'insegnamento tradizionale delle lingue non porta ai risultati attesi, diverse soluzioni vengono proposte quale rimedio. Tuttavia è molto difficile comparare la pertinenza ed il successo di queste proposte in quanto non sono disponibili strumenti di misura facilmente trasponibili a questi problemie.

Due ostacoli possono essere individuati. In primo luogo, come paragonare prestazioni di allievi raccolte in circostanze e con procedure di valutazione differenti, come per esempio tramite test e sulla base di espressione spontanea ? In secondo luogo, come stabilire il livello di difficoltà delle espressioni ascoltate o prodotte dagli allievi quando la natura esatta delle situazioni studiate in classe non è uguale ed inoltre spesso descritta in modo insufficientemente preciso ?

Questo articolo presenta alcune riflessionie su queste difficoltà e sulle soluzioni che proponiamo per superarle. Abbiamo sviluppato uno strumento che permette di definire il livello di difficoltà delle espressioni indipendentamente dal contesto di valutazione ed dalla situazione di classe. Nella seconda parte dell'articolo si propone un primo test di questo strumento-, elaborato sulla base del paradigma della generalizzabilità.