ASp
la revue du GERAS

23-26 | 1999
Varia

Le « *washback* » et les tests intégrés

À propos de l'utilité pédagogique de l'évaluation linguistique

Ross Charnock

Édition électronique

URL : <http://journals.openedition.org/asp/2614>

DOI : 10.4000/asp.2614

ISBN : 978-2-8218-0380-0

ISSN : 2108-6354

Éditeur

Groupe d'étude et de recherche en anglais de spécialité

Édition imprimée

Date de publication : 1 décembre 1999

Pagination : 389-416

ISSN : 1246-8185

Référence électronique

Ross Charnock, « Le « *washback* » et les tests intégrés », *ASp* [En ligne], 23-26 | 1999, mis en ligne le 07 janvier 2012, consulté le 01 mai 2019. URL : <http://journals.openedition.org/asp/2614> ; DOI : 10.4000/asp.2614

Ce document a été généré automatiquement le 1 mai 2019.

Tous droits réservés

Le « washback » et les tests intégrés

À propos de l'utilité pédagogique de l'évaluation linguistique

Ross Charnock

- 1 Les nombreuses nouvelles approches pédagogiques régulièrement proposées pour l'enseignement des langues aux spécialistes des autres disciplines sont le signe de la vitalité incontestable de la discipline, même si elles mettent en même temps en évidence un certain manque de satisfaction par rapport au statu quo. Il est regrettable que l'évaluation soit moins souvent prise en compte, au point de constituer actuellement le point faible de la discipline. Bien qu'importante, la question n'est pas encore prise au sérieux ni au niveau de l'administration, ni sur le plan pédagogique. En règle générale, en langues vivantes comme dans les autres disciplines universitaires, les systèmes d'évaluation sont tenus pour acquis, la notation étant vue comme une corvée obligatoire à accomplir en plus du service annuel proprement dit. L'indifférence des universitaires face à la question de l'évaluation les laisse très en retard par rapport aux professionnels, en Formation continue, ou en Français langue étrangère (FLE). Des enquêtes à ce sujet n'ont pas abouti, les formulaires à remplir étant peu adaptés aux pratiques habituelles. Et les théories présentées dans l'importante littérature de recherche paraissent tellement éloignées des réalités du terrain qu'il semble impossible d'en tenir compte¹.
- 2 Pourtant, l'évaluation est un élément fondamental du système pédagogique, qui doit être pleinement intégré au cours. Malheureusement, il est difficile de trouver un accord entre enseignants de langues vivantes à propos de la définition des compétences à évaluer.

« Washback »

- 3 Tout pédagogue reconnaît l'importance d'objectifs clairs et réalisables, nécessaires si on veut éviter de présenter la langue étrangère comme un gouffre sans fond. Pour être d'une plus grande utilité, ces objectifs doivent correspondre à des critères d'évaluation tout aussi précis. En effet, ces critères sont fondamentaux pour le diagnostic, et pour la motivation des étudiants. En effet, tout étudiant rationnel cherche à maximiser sa note, et fait donc particulièrement attention aux aspects de son travail qui vont être contrôlés.

- 4 En outre, chacun cherche par une sorte d'auto-évaluation perpétuelle à connaître ses points forts et ses faiblesses, afin de pouvoir se comparer aux autres. L'étudiant de langue utilise ainsi le test linguistique comme un instrument de mesure, tout comme l'athlète professionnel se sert du chronomètre. Il s'agit dans les deux cas d'un outil indispensable à l'entraînement.
- 5 Étant donné son rôle important, le système de notation adopté aura forcément un effet important sur les stratégies d'apprentissage. Puisque le travail des apprenants est étroitement focalisé sur les critères adoptés, l'enseignant aussi est amené à s'y adapter, pour mieux répondre aux besoins des étudiants.
- 6 Ces « effets en retour » ont toujours existé², mais pendant longtemps ils ont eu mauvaise presse. En effet, si les critères d'évaluation ne correspondent pas aux objectifs pédagogiques valables, les candidats seront amenés à perdre leur temps en faisant du bachotage, plutôt qu'en pratiquant la langue de manière utile. Il s'agit alors du *washback* négatif.
- 7 On constate qu'en langue, le système d'évaluation est souvent en contradiction avec les méthodes didactiques. On déconseille par exemple le travail de type mot à mot, alors que les étudiants savent qu'ils vont être notés sur des tests de vocabulaire, préparés à l'aide de listes de mots pris hors contexte. De la même façon, pendant les années 1970, de nombreuses institutions ont adopté les nouvelles méthodes communicatives, tout en gardant les systèmes d'évaluation traditionnels. La contradiction entre les objectifs affichés et les critères réels est alors devenue flagrante. Morrow (1979) souligne l'importance pédagogique de la question, en préconisant un système d'évaluation qui serait en accord avec la nouvelle définition communicative des compétences linguistiques. Un tel système, malgré les faiblesses prévisibles au niveau de la fiabilité, présenterait l'avantage d'inciter les étudiants à prendre plus au sérieux les objectifs proposés.
- 8 Malheureusement, puisque les capacités communicatives sont difficilement susceptibles de définition précise, les objectifs, et donc les critères d'évaluation, sont restés vagues. Il est par conséquent problématique de concevoir des tests objectifs pour mesurer cette compétence.
- 9 Les exercices le plus souvent utilisés chez les spécialistes des langues vivantes (thème, version, commentaire, dissertation) ne correspondent nullement aux besoins des spécialistes des autres disciplines. La réussite à ce genre d'exercice dépend autant des compétences extralinguistiques que des capacités en langue proprement dites. De plus, les tests élémentaires utilisés dans le secondaire ne permettent ni de tenir compte des connaissances dans la discipline principale, ni de mesurer les véritables compétences communicatives.

L'Idéal

- 10 Avant de rappeler les inconvénients des tests couramment utilisés, il convient de préciser le cahier de charges du test idéal, en tenant compte des contraintes et des spécificités du secteur LANSAD. On verra qu'il est plus facile de définir les critères du test idéal que de les mettre en pratique, en partie parce que ces critères sont parfois contradictoires entre eux.

- 11 Tout d'abord, étant donné les flux massifs de candidats auxquels il faut actuellement faire face, le test doit être facile à créer et à noter, et en même temps simple à organiser. Il faut aussi que la validité soit acceptable, c'est-à-dire que la note doit correspondre de manière directe au véritable niveau de compétence. La notion de validité peut sembler aller de soi lorsqu'il s'agit d'évaluer le comportement en situation réelle, par exemple lors des tests de conduite ; mais c'est une question plus délicate lorsqu'il s'agit d'estimer les compétences à partir des connaissances théoriques. Les choses sont particulièrement difficiles en langue, puisqu'il n'y a unanimité ni sur la définition des compétences à évaluer, ni sur les critères d'évaluation.
- 12 De surcroît, il ne suffit pas que la validité soit acceptable ; il faut aussi que cette validité soit claire et évidente pour tous les acteurs concernés ; pour les candidats, qui doivent avoir l'occasion de s'habituer au système d'évaluation adopté, et pour l'administration, qui n'aura pas confiance si le format de test ne correspond pas aux habitudes institutionnelles. Rappelons qu'il s'agit ici d'une perception plutôt que d'un fait objectif, car la validité apparente (*face validity*) ne correspond que rarement à la validité réelle ; en effet, il est tout à fait possible de prendre au sérieux un test dont les résultats ne veulent rien dire, et à l'inverse de se méfier d'un test dont les résultats sont systématiquement confirmés par des études théoriques. La validité apparente reste néanmoins importante pour des raisons psychologiques ; si les candidats ne prennent pas le test au sérieux, ils ne feront pas d'effort.
- 13 Il faut aussi réduire autant que possible l'élément de chance. Si deux candidats de compétence équivalente obtiennent à peu près la même note, ou si un même candidat obtient à peu près la même note à chaque tentative, on pourra considérer que le taux de fiabilité (*reliability*) est élevé. Si, à l'inverse, les étudiants ont l'impression d'avoir (ou de ne pas avoir) de chance avec les questions posées, ou bien si le correcteur peut juger les copies de manière plus ou moins sévère selon les jours, alors les notes risquent de varier de manière inacceptable, et le taux de fiabilité sera alors trop bas. Pour améliorer la fiabilité de la notation entre correcteurs (*inter-rater reliability*), il faut que leurs notes soient sensiblement équivalentes, ce qui n'est pas facile avec les tests subjectifs, qui font appel par définition au jugement personnel. Si le taux de fiabilité n'est pas acceptable, on risque de voir des étudiants collés aux examens de juin, mais reçus en septembre avec des notes parfois élevées. Pourtant, malgré les efforts consentis dans certains cas, il est peu probable que la majorité des candidats ait pu faire de tels progrès pendant les vacances. Les différences constatées entre les notes de juin et de septembre correspondent la plupart du temps à la marge d'erreur inhérente à la notation.
- 14 On voit tout de suite une contradiction entre les critères de validité et de fiabilité. En effet, il est toujours possible d'améliorer la fiabilité en ne posant que des questions factuelles, dont les réponses sont vraies ou fausses. Mais alors, puisque ce type de question ne permet pas d'évaluer les compétences communicatives en situation réelle, le taux de validité ne sera plus acceptable. D'autre part, si, pour améliorer la validité, on donne l'occasion au candidat de s'exprimer librement, on ne pourra plus garantir la fiabilité de la notation.
- 15 L'idéal serait de trouver le moyen de mesurer ce qui est important, plutôt que de prendre pour important ce qui est directement mesurable. Les effets en retour seront alors favorables, et on pourra conseiller l'utilisation du test non seulement pour le travail en classe, mais aussi pour le travail en autonomie.

- 16 Au-delà de ces considérations générales, il existe aussi certaines contraintes particulières pour l'enseignement aux étudiants des autres disciplines. Il est souvent utile, dans la mesure où les connaissances dans la discipline principale sont susceptibles de faciliter l'acquisition de la langue, d'adapter le contenu du cours pour tenir compte des intérêts des étudiants. Il faut alors pouvoir en tenir compte dans les examens, en évitant toutefois de donner trop d'importance aux connaissances extralinguistiques dans la note de langue.
- 17 D'autre part, étant donné les disparités importantes en ce qui concerne les niveaux de départ, le test idéal doit être utile à tous les niveaux. En d'autres termes, il devra être abordable pour les faibles, sans paraître enfantin aux yeux des forts, qui risqueraient autrement d'obtenir des résultats improbablement élevés, sans avoir fait des progrès. Il serait également utile de pouvoir faire des distinctions fines à tous les niveaux, et même de détecter les progrès de semaine en semaine. On conviendra en effet que, bien plus que l'ambiance en cours ou la personnalité du professeur, c'est surtout la conscience du progrès qui motive les étudiants, et qui les incite à travailler de manière utile.

Insuffisance du système actuel

- 18 La situation que l'on constate aujourd'hui, à l'université, ne correspond nullement à l'idéal qui vient d'être décrit. Les procédures adoptées sont tellement variées et variables que cela ne peut même pas être qualifié de système. Certes, la plupart du temps, les correcteurs peuvent s'arranger pour avoir par exemple le même pourcentage de reçus, mais il n'y a souvent pas d'accord sur les critères. La situation chez les enseignants de langue aux spécialistes des autres disciplines est particulièrement difficile, pour plusieurs raisons.
- 19 D'abord, pour des raisons qui relèvent de la tradition ancienne, l'institution exige une note unique sur vingt, qui doit tenir compte de la compétence linguistique, de l'effort fourni, et éventuellement des connaissances extralinguistiques. Il n'est jamais question de référentiels (*rating scales*), pourtant indispensables pour la construction d'un barème adéquat et pour la définition des critères objectifs. La plupart des étudiants sont donc notés par leur propre professeur, à partir des questions traitées en cours. Une telle note de langue ne fournit aucune information utile, lisible depuis l'extérieur, sur la compétence linguistique du candidat.
- 20 D'autre part, si le cours porte non seulement sur la langue mais aussi sur un contenu extralinguistique, il faudra nécessairement en tenir compte dans la notation. Pourtant, il n'y a aucun consensus concernant la part de la spécialité à prévoir dans la note (unique) de langue.
- 21 Enfin, il n'est pas question de supprimer la note de « participation ». Celle-ci est utilisée en principe pour encourager l'expression orale, qu'on n'est pas en mesure de tester de manière systématique, en partie à cause du manque de moyens. Mais elle est également utile pour éviter les injustices. Étant donné les niveaux de départ variés, les étudiants les plus faibles n'ont aucune chance de rattraper les plus forts. À la différence des enseignants de la discipline principale, la plupart des collègues linguistes considèrent qu'il faut essayer d'assurer un enseignement utile pour la majorité des inscrits, plutôt que de préparer une sorte de concours éliminatoire en fin d'année. Ils répugnent à l'idée d'abandonner ceux parmi les faibles qui font consciencieusement tout ce qu'on peut

raisonnablement leur demander de faire, pendant toute l'année. La solution évidente (mais trop rarement adoptée) serait d'attribuer deux notes, l'une pour le travail, et l'autre en rapport avec le niveau absolu. Cela permettrait par ailleurs la constitution de groupes de niveau, comme dans les Grandes Écoles. Malheureusement il s'agit d'une notion qui va à l'encontre des impératifs administratifs dans la plupart des universités.

- 22 Dans la pratique, cette note de participation, qualifiée de « note de gueule » dans de nombreuses Facultés, est souvent utilisée pour récompenser la bonne volonté, plutôt que la compétence linguistique, et pour corriger les notes invraisemblables attribuées à la suite de tests peu fiables. Suivant une pratique assez répandue, inspirée du secondaire, de nombreux collègues préfèrent garder cette note en réserve pour contrôler le comportement des étudiants qui ne viennent en cours que par obligation.
- 23 Si les étudiants parlent souvent de « tenter leur chance » à l'université, c'est qu'il leur semble que la notation est justement une affaire de chance. L'amélioration du système ne semble pas facile, étant donné qu'il faudrait d'abord changer les habitudes acquises depuis longtemps. En effet, chaque enseignant considère que son système personnel est tout à fait adéquat, même si cela relève de l'expérience et du flair et ne peut donc être précisé (Isani 1996 : 358). Des critères précis pourraient pourtant faciliter la notation, et seraient aussi utiles pour les candidats à l'examen (sans parler des enseignants responsables) ; mais vu le nombre de candidats à évaluer, une notation plus détaillée semble hors de question. Il faut aussi admettre que, plus les enseignants expérimentent de nouvelles méthodes pédagogiques, plus ils auront besoin d'un test commun, pour évaluer à la fois le niveau des étudiants et l'efficacité des différentes approches. Le problème est d'autant plus urgent pour les Facultés qui disposent déjà de Centres de ressources, et qui sont en mesure de proposer aux étudiants des cursus individualisés. Pour des raisons pédagogiques évidentes, lorsqu'un candidat a suivi un cursus qu'il a fixé lui-même, en suivant les conseils de l'enseignant responsable, il faut l'évaluer par rapport à ce cursus. Pourtant, il est nécessaire de connaître en même temps son niveau « absolu », par rapport aux critères plus généraux. Il est difficile de réconcilier ces deux impératifs.

Les systèmes traditionnels

- 24 Parmi les procédures d'évaluation courantes, on peut distinguer deux approches principales, qui correspondent à des méthodes pédagogiques opposées. Il s'agit d'un côté des méthodes fondées sur la forme et sur la structure de la langue, et de l'autre des méthodes directes, établies à partir de la compréhension et la communication. Pour tester la forme, on utilise le plus souvent des questions ponctuelles, susceptibles de notation objective, alors que pour évaluer la compétence communicative, on utilise autant que possible le discours authentique, ce qui impose une évaluation subjective. Comme les méthodes d'enseignement correspondantes, les deux approches ont des avantages et des inconvénients.

Les questions ponctuelles

- 25 Les questions sur les points séparés (*discrete point items*) sont le plus souvent utilisées pour évaluer les connaissances de base en grammaire ou en vocabulaire. Tant qu'il s'agit de composer une petite vingtaine de questions, ce genre de test est aussi facile à construire

qu'à corriger. Mais pour que les résultats soient significatifs, il faut un nombre plus important de questions, conçues d'une manière plus élaborée.

- 26 L'expérience montre le manque de corrélation entre les résultats obtenus à ce genre de test et la compétence communicative. Puisqu'il ne semble pas possible de prévoir les capacités linguistiques en situation réelle à partir des connaissances en grammaire scolaire, il faut en conclure que la validité réelle est quasi inexistante. Pourtant, ce genre de test correspond à une tradition institutionnelle, et a acquis par ce biais une certaine validité apparente. Le maintien des méthodes scolaires d'évaluation peut même avoir un effet stabilisateur pour certains étudiants, bien que ce soit plutôt décourageant pour les plus avancés.
- 27 Puisqu'il n'y a en principe qu'une seule bonne réponse, ces exercices sont corrigés de manière objective, la fiabilité de la notation étant par conséquent élevée. Mais s'il n'y a pas plus de questions, cela laissera toujours trop de place à la chance. Un candidat faible risque de passer au-dessus de la moyenne, en tombant sur des questions qui lui conviennent, ou vice versa. Dans la pratique, les candidats à ce type de test sont évalués non pas en rapport avec des critères objectifs (*criterion referenced*), mais les uns par rapport aux autres (*norm referenced*).
- 28 Ces exercices ne sont utilisables qu'avec les candidats les plus faibles, pour le diagnostic des faiblesses élémentaires. Puisque les plus forts obtiennent des notes trop élevées, ce format de test ne permet pas de faire des distinctions fines.
- 29 Puisque du point de vue pédagogique, ces exercices n'aident pas à faire des progrès, il faut admettre que les effets en retour sont mauvais (*negative washback*). Ces tests traditionnels vont à l'encontre des méthodes d'enseignement modernes, en principe communicatives. Sauf pour les quelques vrais débutants qui se sont « infiltrés » à la Faculté, le problème des étudiants spécialistes des autres disciplines n'est pas l'apprentissage des items lexicaux et des règles élémentaires de grammaire (si cela suffisait, ils seraient tous « doués en langue »). La fonction du professeur est plutôt d'aider les étudiants à utiliser les connaissances qu'ils ont déjà, pour s'exprimer en situation authentique. Or, par définition, les questions ponctuelles sont prises hors contexte, et ne sont pas associées à un sens ou à un contenu ; elles ne permettent pas d'utiliser la langue à des fins communicatives.
- 30 Pire, les étudiants ne manquent pas de relever l'incohérence fondamentale qui consiste à décourager le travail du type mot à mot pendant le cours, alors qu'à l'examen il faut justement sortir des mots séparés, ou appliquer des règles scolaires. S'ils savent qu'un tel examen est prévu, ils persisteront, malgré tous les conseils prodigués en classe, à perdre leur temps en travaillant de manière inutile.
- 31 Le problème n'est pas le même avec le système de questions à choix multiples, qui constituent l'exemple prototypique du test sur questions ponctuelles. Les QCMs sont faciles à corriger (surtout si on dispose de moyens technologiques modernes), mais elles sont souvent difficiles à organiser, à cause des effectifs importants. Ils sont surtout extrêmement difficiles à composer. L'obligation de proposer plusieurs mauvais choix, quand même plausibles, impose des contraintes importantes. De plus, la validation des questions implique des calculs statistiques impressionnants, qui dépassent de loin les compétences des enseignants de langue.
- 32 Lorsqu'on utilise les QCMs pour contrôler le vocabulaire et la grammaire, on se heurte aux mêmes problèmes de validité que pour les tests classiques. Mais des recherches

menées aux États-Unis (surtout chez l'ETS à Princeton) montrent qu'il est possible d'utiliser des QCMs produits par des professionnels et testés au préalable à grande échelle, afin d'évaluer de manière indirecte toutes les formes de connaissances linguistiques susceptibles de définition précise, y compris la compréhension de l'implicite, la reconnaissance des sous-entendus, des présupposés et des références culturelles. Des études statistiques permettent même d'utiliser les résultats obtenus aux tests objectifs pour prévoir les capacités en expression libre. Malheureusement, il est peu probable que les QCMs maison soient dignes de confiance. Ils sont pourtant acceptés au niveau de l'administration, et apparemment pris au sérieux par les candidats.

- 33 L'objectif principal de ces batteries de tests est de confirmer l'existence d'un niveau élevé en langue ; par conséquent, pour une majorité des étudiants, les résultats aux tests TOEFL, par exemple, ne vont généralement pas au-delà de ce qu'on pourrait attribuer au hasard. Les tests qu'on peut trouver sur le marché ne permettent pas de faire des distinctions fines à tous les niveaux. Il n'existe d'ailleurs pas de QCMs spécifiquement adaptées aux différentes disciplines.
- 34 D'autre part, les QCMs sont conçues précisément pour évaluer le niveau linguistique des candidats indépendamment de la préparation. Par conséquent (et par principe), elles n'ont pas d'effets en retour, ni bon ni mauvais. En format test, les résultats ne fournissent pas d'informations utiles permettant de déceler les faiblesses éventuelles. Les candidats travailleront sur ce type d'exercice non pas pour progresser, mais uniquement pour se familiariser avec le format.
- 35 De manière générale, les tests sur questions ponctuelles proposés à l'université ont un taux de validité très contesté. Les problèmes semblent difficiles à résoudre, puisqu'inhérents au système. Il est inutile par exemple d'essayer d'introduire les éléments d'un contexte dans un test qui porte sur la structure de la langue ; cela fera immédiatement ressortir de nouvelles fautes imprévues, et aura pour effet de compliquer la notation sans pour autant améliorer la validité. Il en va de même pour les questions de compréhension, éventuellement purement factuelles, qui imposeraient l'adoption de critères supplémentaires. À cause de la contradiction entre le système d'évaluation et les objectifs de l'enseignement, les effets en retour semblent donc défavorables. Il est préférable de travailler autant que possible avec le discours authentique.

Tests d'expression libre

- 36 Parmi les différents types de tests linguistiques, les dissertations et les compositions figurent parmi les plus traditionnels. À proprement parler, ce ne sont pas des tests, notés de manière quantitative, mais des épreuves, évaluées de manière subjective ; elles sont pourtant bien acceptées comme système d'évaluation, à la fois par l'institution et par les candidats. Le sujet est facile à préparer, même si la correction des copies prend du temps. Des épreuves de ce type permettent, et imposent même, l'intégration du contenu du cours. Et puisque l'exercice correspond parfaitement aux objectifs affichés, la validité du test semble également acceptable. L'exercice permet en effet de juger l'expression écrite de manière directe.
- 37 Le problème principal est celui de la fiabilité, la notation étant nécessairement subjective. Les expériences indépendantes ont montré en effet que les notes attribuées aux copies sensiblement équivalentes varient bien plus que l'on ne voudrait le croire³. Il est

notamment difficile de comparer des copies longues, avec beaucoup de fautes, et des copies plus courtes, qui en comportent forcément moins. Il semble d'ailleurs difficile de se mettre d'accord sur la gravité relative des différentes fautes possibles, dont certaines pourraient être qualifiées de rédhitoires. S'il faut aussi évaluer la cohérence et l'organisation de l'écrit, le jugement porte alors explicitement sur un critère non linguistique, sans qu'il soit possible de préciser à l'avance l'importance de ce critère. En outre, il faut obligatoirement tenir compte des idées, qui conditionnent les difficultés d'expression. Pourtant, malgré l'importance de ce critère, il n'est pas possible de préciser son rôle dans la notation. L'évaluation est donc forcément impressionniste, ce qui a pour effet de baisser le taux de fiabilité. C'était pour des raisons pédagogiques, en rapport avec le *washback*, que le Test TOEFL comprenait pendant longtemps une question de composition, sans que celle-ci soit notée (Spolsky 1995 : 322). Faute d'accord sur les critères, le taux de fiabilité n'a jamais été considéré comme acceptable.

- 38 Puisque les copies ne sont pas facilement comparables entre elles, il semble délicat d'évaluer les candidats par rapport à une norme. L'utilisation des référentiels permet en revanche de les classer par niveau ; pourtant, ce genre de classement ne permet pas d'établir des distinctions fines, et ne justifie pas l'attribution d'une note précise sur vingt.
- 39 Il existe des techniques pour « objectiviser » l'évaluation de l'écrit, malgré la subjectivité inhérente aux épreuves de ce type. La formation des examinateurs permet ainsi d'améliorer la fiabilité, par une meilleure harmonisation de la notation. On peut également proposer une méthode de notation analytique, en notant séparément les différents aspects de la production, selon des critères plus ou moins complexes. Cependant, il s'agit alors de remplacer une note holistique, ou impressionniste, par une multiplicité de jugements subjectifs ; la note finale dépendra en fin de compte des critères et des coefficients adoptés, qui restent toujours contestables⁴.
- 40 Afin d'éviter ces difficultés, on pourrait demander des résumés et des commentaires à la place des rédactions « libres », afin de rendre les copies comparables en imposant une certaine structure, et pour simplifier l'exercice en fournissant les éléments du vocabulaire. Mais cela risque de soulever de nouveaux problèmes de notation. En effet, il faudra tenir compte en même temps de la compréhension et de l'expression. Il n'est pas question d'interdire l'emploi des expressions tirées du texte, mais il faut éviter de donner des points pour ce qui semble être copié de manière trop directe. La composition guidée est certes plus facile à corriger, mais elle a le défaut de paraître enfantine aux yeux des plus avancés.
- 41 Les problèmes sont encore pires pour les exposés à l'oral. Il faut tenir compte de la correction grammaticale, de l'intérêt du contenu, et de la prononciation, le tout en temps réel. De surcroît, on est en droit de demander aux candidats de faire preuve d'intelligence — même s'il faut reconnaître que ni cette qualité, ni l'aptitude à la prise de parole en public ne peuvent être considérées comme des critères linguistiques, proprement dits. Malgré ces difficultés, l'interrogation orale est souvent utilisée pour porter un jugement sommaire et définitif à l'occasion des différents concours. En revanche, aux examens universitaires, on reconnaît volontiers le manque de précision dans la notation en accordant quelques points de plus, au titre du rattrapage, à ceux qui en auraient besoin. Pour améliorer la fiabilité, il est parfois nécessaire de limiter ces entretiens à des échanges structurés élémentaires.
- 42 En France, on a récemment relevé le défi d'évaluer l'expression libre à la fois à l'écrit et à l'oral, afin d'évaluer les capacités linguistiques des adultes professionnels, candidats au

« Diplôme de Compétence en Langue » (DCL). Suivant un « scénario » unique, censé simuler des situations réelles en rapport avec la vie professionnelle, le candidat doit d'abord recueillir des informations dans un temps limité, à partir de documents écrits et audio ; il doit ensuite recueillir d'autres informations en interrogeant un examinateur interlocuteur, afin d'établir un rapport écrit. Il doit enfin présenter son travail à l'oral. Pour être en mesure d'évaluer les candidats à ces épreuves, les examinateurs suivent un stage de deux jours ; ils doivent être en mesure de remplir des grilles analytiques très détaillées, portant soit sur des critères pragmatiques (la communication interactive), soit sur des critères purement linguistiques (analyse grammaticale classique). Pour les épreuves à l'oral, le candidat se trouve en présence de deux examinateurs. Cette complexité d'organisation et de notation semble effectivement nécessaire pour assurer la fiabilité de l'épreuve, en ce qui concerne le classement des candidats sur cinq « degrés de compétence » (y compris le degré « zéro »). Bien qu'il ne soit pas possible de noter de manière plus précise, il semble utile, pour des raisons pédagogiques, de communiquer aux candidats le détail des jugements des examinateurs selon les différents critères, tel qu'il apparaît sur les grilles analytiques.

- 43 Puisque ce qu'on demande aux candidats correspond directement à leurs objectifs, la validité de ce certificat semble clairement acceptable. Cependant, l'utilisation de cette procédure en l'état semble problématique dans les universités. Il s'agit d'un exercice très difficile, à peine abordable pour la plupart des étudiants spécialistes d'autres disciplines. Dans ces conditions, le *washback* ne sera pas forcément favorable, surtout pour les candidats qui ne risquent jamais de dépasser le degré « un ». D'autre part, étant donné l'importance des connaissances professionnelles dans la notation, il faut reconnaître qu'il s'agit d'une épreuve (de langue) qui ne porte pas uniquement sur la langue. Enfin, les universités n'ont pas tous les moyens financiers de mobiliser deux enseignants pour chaque candidat, ou de les former au métier d'examineur. Il ne faut pas oublier non plus que la conception des scénarios est très difficile, et doit être considérée comme une activité à plein temps.
- 44 Au niveau universitaire, il est donc question d'une version simplifiée du DCL, dans la forme d'une « certification de la Compétence en Langues pour les Études Supérieures Spécialisées » (cCLESS). La simplification des tâches risque pourtant d'entraîner une perte de fiabilité en soulevant de nouveaux problèmes de notation. En revanche, il sera probablement possible de simplifier les grilles, pour éviter par exemple de séparer artificiellement les critères linguistiques des critères pragmatiques, ou encore de juger séparément, de manière irréaliste, de l'acceptabilité des sujets et des prédicats (GN/GV) produits par le candidat.
- 45 En résumé, les rédactions, comme les exposés, constituent des exercices utiles, avec des effets en retour favorables. Mais il faut un niveau élevé pour réussir à ce genre d'épreuve. En effet, les compétences requises correspondent aux objectifs finaux de l'enseignement, et ne sont pas forcément abordables pour les étudiants de niveau intermédiaire, qui risquent alors d'être découragés. Pour que la fiabilité soit acceptable, il faut trouver le moyen « d'objectiviser » la notation ; et pour que ces exercices aient une utilité pédagogique quelconque pour les étudiants LEP (*Limited English Proficiency* ou « *linguistically challenged* » cf. *Focus on Evaluation and measurement*, 1992), il faut trouver le moyen de les simplifier.
- 46 Il existe donc d'un côté des exercices fondés sur des questions ponctuelles, susceptibles de notation objective, dont la fiabilité est élevée, mais qui n'encouragent pas l'expression

dans des situations authentiques ; leur validité n'est donc pas acceptable et leurs effets en retour sont négatifs. De l'autre côté, il existe des exercices sur le discours libre, dont la validité est plus élevée, et qui permettent de juger des capacités communicatives de manière plus directe ; ceux-ci ont des effets en retour plutôt favorables, mais ils ne sont pas susceptibles de notation objective et sont donc peu fiables. On constate alors une contradiction (*trade-off*) entre les deux impératifs. Au lieu d'accepter de faire un choix, ou un compromis apparemment inévitable, entre deux systèmes, chacun inacceptable, il semble préférable d'essayer de réconcilier les différents critères, à l'aide du discours continu.

Tests sur le discours continu

- 47 Ni les méthodes fondées sur les points séparés, ni les méthodes plus directes, établies à partir du discours libre, ne permettent d'obtenir les résultats escomptés. Les premières attirent l'attention surtout sur la forme, et ne donnent pas l'occasion à l'élève d'utiliser ses connaissances, ou de s'exprimer de manière satisfaisante en situation réelle. Les secondes en revanche se concentrent sur le sens intentionnel, ce qui est loin de garantir l'acquisition de la structure linguistique de base. À cause de leurs insuffisances respectives, ces deux types d'approches ont longtemps alterné, par une sorte d'effet de pendule. Cette expérience, vérifiée sur plusieurs siècles⁵, laisse penser que le vrai problème ne se situe ni au niveau de la phrase (la signification sémantique), ni au niveau de l'énoncé (le sens intentionnel), mais plutôt à l'interface entre les deux⁶. Il faudrait alors rechercher une méthode qui permettrait de travailler justement sur le rapport entre ces deux formes de connaissance linguistique.
- 48 Il en va de même pour l'évaluation. Il ne suffit pas d'alterner les tests sur la forme et les tests fondés sur le contenu ; il faut plutôt rechercher des formats qui permettent d'intégrer les deux approches, en cumulant les avantages des deux côtés, tout en évitant les inconvénients.
- 49 À la fin des années 70, un certain nombre de « tests intégrés » ont été proposés ; leur objectif était de confirmer l'existence d'une « compétence unitaire » en langue (Oller 1979). Grâce à la « compétence unitaire » postulée, les chercheurs pensaient être en mesure de prévoir les compétences séparées à partir d'une évaluation unique ; les tests supplémentaires seraient ainsi rendus superflus. Certes, cette hypothèse n'est plus d'actualité, au moins dans sa forme forte⁷, mais les corrélations observables entre les différentes compétences n'ont pas pour autant été remises en question. Ces tests font surtout appel aux attentes linguistiques et aux inférences pragmatiques ; ils dépendent donc de la mise en relation de la forme et le contenu, au niveau de la signification et du sens. Ce sont donc des exercices utiles à partir d'un niveau élémentaire, jusqu'aux niveaux les plus avancés. Le format du test permet également une notation objective, ce qui permet d'obtenir une bonne fiabilité et un taux élevé de validité, avec des effets en retour favorables.
- 50 À la différence des « scénarios » du DCL, ces tests, fondés non pas sur l'expression libre mais sur le discours continu, sont faciles à préparer et à corriger. La notation est objective, ce qui garantit un taux élevé de fiabilité, alors qu'en même temps la continuité du discours permet l'interprétation contextuelle, nécessaire pour l'acquisition de la langue. Seule la validité apparente est susceptible de poser problème, puisqu'il s'agit d'un type de test qui ne correspond ni aux habitudes de travail des étudiants, ni aux pratiques

administratives. Il faut donc que les étudiants aient l'occasion de travailler en classe avec des exercices du même type, et que les objectifs administratifs soient clairement expliqués. La validité réelle se voit corroborée par le fait que les notes obtenues aux différents tests intégrés restent sensiblement les mêmes. Mais les différents formats ne fonctionnent pas tous de manière aussi efficace ; la procédure « cloze » en particulier présente des faiblesses dont il faudrait tenir compte.

Le Cloze test

- 51 Le cloze test, où il s'agit de compléter des textes à trous, constitue l'exemple prototypique du test intégré. Il s'agit d'une procédure utilisée comme exercice depuis de nombreuses années, et comme test depuis plus de quarante ans (Spolsky 1995 : 160)⁸. Il est à remarquer que les cloze tests permettent d'intégrer sans problème le contenu du cours à l'intérieur de l'exercice linguistique. Prenons un exemple journalistique, sur un sujet économique et commercial :

High taxes and welfare ___ are the real obstacles to job ___ in France. This is what the ___ should be thinking about, rather than dubious ___ for sharing existing work. High benefits for ___ out of work mean there is often ___ benefit from taking a job. The minimum ___ and hefty labour taxes have ___ up the cost to employers of ___ on new workers. This is exacerbated by ___ restrictions on hiring and firing. ?/10 (Corrigé en annexe 1)

- 52 Il s'agit d'un test de vocabulaire qui a le mérite de placer les mots en contexte. Cela exige en même temps la compréhension du texte et la connaissance de la grammaire. La validité de l'exercice est vérifiée par le fait que les résultats obtenus sont en étroite corrélation avec les capacités réelles en compréhension et en expression. Puisqu'il s'agit d'un exercice courant, dont les candidats ont l'habitude, la validité apparente semble également acceptable, à la fois du point de vue des étudiants et de l'administration.
- 53 Le cloze test est facile à préparer et aussi à organiser. Selon une convention généralement acceptée, il suffit d'enlever un mot tous les sept mots environ, dans un texte donné, puis de comparer les réponses avec l'original. Malheureusement, le fait que les sujets parlants de langue maternelle (*native speakers*) soient souvent incapables de trouver le mot manquant montre qu'il ne suffit pas de connaître la langue ; il faut aussi trouver la bonne approche psychologique et faire preuve d'imagination. La fiabilité présente aussi quelques difficultés. Le nombre de trous étant limité, l'exercice comporte souvent trop peu de questions, ce qui laisse une place importante à la chance. D'autre part, même lorsque le candidat trouve le mot juste, il peut commettre une faute d'orthographe. Il faut alors trouver le moyen de le pénaliser par rapport aux candidats qui n'ont pas fait de faute, tout en l'avantageant par rapport à ceux qui n'ont pas trouvé le mot. Enfin, on voit souvent des cas où plusieurs réponses sont admissibles, dont certaines pourraient être prises pour des améliorations du texte original. La notation exige donc un jugement subjectif de la part du correcteur.
- 54 Il s'agit là de problèmes que l'on peut éviter, soit en donnant la première lettre du mot qui manque, afin de réduire le choix de mots possible, soit en présentant les questions en forme de QCM, afin d'éviter les fautes d'orthographe. Cependant, dans les deux cas, cela

revient à enlever autant d'éléments utiles pour l'évaluation. D'autre part, si la correction est plus commode, l'exercice devient alors plus difficile à préparer.

- 55 Étant donné que certains textes conviennent mieux aux étudiants faibles, et d'autres aux plus avancés, cet exercice ne peut pas établir des distinctions fines à tous les niveaux. Il est d'ailleurs difficile sinon impossible de calculer à l'avance le niveau de difficulté d'un exercice donné. Comme on l'a souvent fait remarquer, il suffit pour changer le niveau de difficulté de choisir un autre point de départ pour les trous. Par exemple, avec le même texte de base que précédemment :

High taxes and welfare benefits are the real ___ to job creation in France. This is ___ the government should be thinking about, ___ than dubious schemes for sharing existing _____. High benefits for those out of work ___ there is often little benefit from ___ a job. The minimum wage and hefty ___ taxes have driven up the cost to ___ of taking on new workers. This is ___ by complex restrictions on hiring and _____. ?/10 (Corrigé en annexe 1)

- 56 Par conséquent, il ne semble pas possible d'utiliser la procédure cloze comme instrument de mesure permettant de préciser le niveau absolu du candidat ; il s'agit plutôt d'un exercice qui permet de comparer les candidats entre eux.
- 57 Pour ce qui concerne le *washback*, le fait que le test fonctionne comme exercice constitue un avantage. Il peut servir par exemple, moyennant une préparation particulière, pour la vérification des connaissances de structures grammaticales spécifiques. Il s'agit d'un bon exercice de compréhension écrite, même si le rôle de l'étudiant est trop passif pour que cela puisse être considéré comme une préparation utile pour l'expression.
- 58 En format test, la procédure cloze permet alors de vérifier les connaissances linguistiques de manière assez efficace ; elle peut par ailleurs servir d'exercice linguistique. Mais, à cause de ses défauts de fiabilité, elle ne permet pas une évaluation quantifiée, en rapport avec des critères objectifs.

Le C-test

- 59 Il suffit d'avoir jeté un œil distrait sur la littérature pertinente des quinze dernières années pour savoir que le C-test semble avoir permis de résoudre la plupart des problèmes associés à la procédure cloze. Il s'agit toujours d'un test intégré, qui exploite le principe de redondance réduite, mais de manière plus efficace. Dans ce format, au lieu de laisser un trou tous les sept ou huit mots, on enlève la (seconde) moitié du mot tous les deux mots. Si le nombre de lettres est impair, on enlève une lettre de plus, les mots à une seule lettre étant fournis « gratuitement ». Ainsi, si une seule lettre est donnée, le mot à trouver doit en comporter deux ou trois. Et si l'indice comprend deux lettres, alors la réponse doit en comporter soit quatre, soit cinq. Ainsi, « t__ » pourrait correspondre à « to », « two », « too », « tea », « the », etc., alors que « th__ » correspondra à « this », « they », « that », « there », « then », « those », etc. D'habitude, on donne, toujours « gratuitement », la première phrase du paragraphe afin de fournir les indications contextuelles utiles. Pour s'en convaincre, voyons d'abord un exemple en français, avant de revoir le texte en langue étrangère :

Les nombreuses nouvelles approches pédagogiques régulièrement prop_____ pour l'enseig_____ des lan_____ aux spécia_____ des aut_____ disciplines so_____ le si_____ de l_____ vitalité incont_____ de l_____ discipline, mê_____ si el_____

montrent e__ même te__ un cer__ man__ de satisf__ par rap__
 au st__ quo. I__ est regre__ que l'évalu__ soit mo__ souvent pr__
 en com__, au po__ de const__ actuellement l__ point fai__ de l__
 discipline. ?/30 (Corrigé en début d'article)

High taxes and welfare benefits are the real obstacles to job creation in France. Th__ is wh__ the gover__ should b__ thinking ab__, rather th__ dubious sch__ for sha__ existing wo__. High bene__ for th__ out o__ work me__ there i__ often lit__ benefit fr__ taking a j__. The min__ wage a__ hefty lab__ taxes ha__ driven u__ the co__ to empl__ of tak__ on n__ workers. Th__ is exace__ by com__ restrictions o__ hiring a__ firing. ?/31 (Corrigé en annexe 1)

- 60 Puisque, pour chaque question, il n'y a qu'une seule réponse possible, la notation est purement objective. Même dans les rares cas où il faut choisir entre deux mots grammaticalement acceptables (par exemple lorsqu'il faudrait choisir entre « ten » ou « two »), le contexte impose un choix clair et incontestable. Il n'y a donc plus de problème de fiabilité entre correcteurs.
- 61 Ce système « sans douleur » permet évidemment de tenir compte de la même façon de la discipline principale. Il permet également de multiplier le nombre de questions posées par rapport à un texte donné, ce qui réduit d'autant l'élément de chance. L'exemple ci-dessus comporte 31 items dans un paragraphe relativement court. Quatre ou cinq exercices de ce type tiennent facilement sur une page (recto verso), et permettent le plus souvent de 120 à 150 questions. Comme dans la procédure cloze, les étudiants ayant apparemment trouvé le bon mot risquent toujours de faire des fautes d'orthographe. Mais, en règle générale, accorder des demi-points pour les fautes qu'on estime moins graves ne fait que compliquer inutilement la notation, sans changer le classement général. L'expérience montre en effet qu'il suffit de compter tout simplement les fautes de manière mécanique, en exigeant la perfection, si on veut obtenir des résultats clairs et nets. Il s'agit bien d'une note générale, conformément aux exigences institutionnelles ; mais l'analyse des copies permet d'obtenir des informations plus détaillées, pédagogiquement utiles.
- 62 À première vue, on pourrait prendre ce type d'exercice pour facile et élémentaire. Hughes (1989 : 71) s'inquiète d'y voir l'aspect d'un « puzzle », et se demande s'il ne suffirait pas d'adopter une stratégie particulière pour améliorer sa note. Il s'agit là d'une fausse impression ; en effet, la recherche plus récente (voir bibliographie) montre que la validité du test est en réalité très élevée. Il est à noter que les « natifs » obtiennent systématiquement, en employant uniquement leurs connaissances linguistiques, et presque en temps réel, la note de 20/20, ce que les étudiants étrangers ne peuvent presque jamais faire⁹.
- 63 Loin d'être simpliste, le test se révèle utile à tous les niveaux. Le nombre élevé de questions permet d'obtenir un échantillon représentatif des problèmes linguistiques, de difficulté variable. En ce qui concerne le vocabulaire, la longueur du mot ne semble pas poser de problème particulier, car l'indice devient automatiquement plus long. En revanche, la morphologie est souvent source de difficultés, surtout lorsque le mot comporte une apostrophe. Les prépositions sont relativement faciles, probablement parce qu'elles sont à la fois courtes et invariables. L'exercice ne prend que peu de temps ; dix ou quinze minutes par paragraphe suffisent amplement, sans que la durée de l'épreuve soit ressentie comme une contrainte. Les candidats les plus faibles obtiennent toujours quelques points, mais même les plus forts ne trouvent presque jamais toutes les réponses.

Lors d'une expérience récente, avec des étudiants de première année (en Sci Eco) la fourchette était de 18 à 80 fausses réponses sur 146 questions, avec une moyenne de 49 ; ce qui permet d'élaborer un barème en plaçant la moyenne autour de 35 ou 40 fautes¹⁰.

- 64 Weir (1990 : 49) signale des problèmes concernant la validité apparente ; il faut effectivement admettre que les étudiants n'ont pas l'habitude de ce genre de test. Comme bon nombre de musiciens, pas toujours d'accord avec le métronome qui constitue pourtant leur outil de travail, certains étudiants de langues vivantes n'apprécient guère les contraintes imposées par ce format de test, contraintes qui sont pourtant nécessaires pour la précision de la notation. Il s'agit de ceux qui caressent toujours l'espoir de faire la dissertation de leur vie le jour de l'examen, et d'obtenir une note plus élevée que d'habitude. Mis à part le fait que ces étudiants sont souvent déçus, leur objectif ne correspond pas vraiment à celui de l'administration.
- 65 Il existe aussi quelques candidats (un ou deux par groupe TD) qui ont réussi à développer une certaine facilité linguistique (*fluency*), souvent au détriment de la justesse grammaticale, et qui espèrent cacher leurs faiblesses linguistiques, soit avec la sympathie et l'aisance de la présentation, soit avec l'originalité des idées ; ils peuvent se sentir frustrés face au C-test, qui ne mesure que les compétences strictement linguistiques. On pourrait certes envisager de récompenser d'une manière ou d'une autre la facilité linguistique (il faut encourager toutes les stratégies d'apprentissage valables), mais cela n'oblige pas à renoncer à l'emploi d'un exercice aussi fiable que le C-test, qui présente l'avantage d'inciter les candidats à travailler sur leurs faiblesses plutôt que sur leurs points forts.
- 66 De la même façon, un petit nombre de candidats n'arrivent pas à s'exprimer de manière convenable en classe, mais réussissent au-delà des attentes avec le C-test. Pour ces étudiants aussi, il serait envisageable d'obtenir un échantillon de l'anglais écrit, en ajoutant une question de rédaction, afin de confirmer les résultats obtenus. Cela permettrait éventuellement de moduler la note objective vers le haut ou vers le bas ; l'expérience montre que ce genre d'ajustement est rarement nécessaire.
- 67 Comme avec le cloze-test classique, certains exercices sont plus difficiles que d'autres, sans que le niveau soit toujours prévisible. On constate sans surprise que plus le texte est prévisible, plus l'exercice est facile. C'est pourquoi, pour les spécialistes des autres disciplines, les textes en rapport avec la spécialité sont souvent plus faciles que ceux qui traitent des situations quotidiennes. En revanche, les textes littéraires et poétiques peuvent se révéler problématiques, même pour les anglophones. Il n'est pas inutile de donner ici quelques exemples, de difficulté variable, tirés de romans célèbres.

Under certain circumstances there are few hours in life more agreeable than the hour dedicated to the cere_____ known a__ afternoon t__. There a__ circumstances i__ which, whe_____ you par_____ of t__ or n__ —some peo_____ of cou_____ never d__— the situ_____ is i__ itself delig_____. Those th__ I ha__ in mi__ in begi_____ to unf_____ this sim_____ history off_____ an admi_____ setting t__ an inno_____ pastime. ?/25

"Yes, of course, if it's fine to-morrow", said Mrs Ramsey. "But you'll have to be up wi__ the la____", she ad____. To h__ son th__ words conv_____ an extrao_____ joy, a__ if i__ were set_____ the exped_____ were bo__ to ta__ place, a__ the won_____ to wh__ he h__ looked for____, for ye__ and ye__ it see____, was, af__ a nig__ dark_____ and a da__ sail, wit_____ touch. ?/19

Stately, plump Buck Mulligan came from the stairhead, bea____ a bo____ of
lat____ on wh____ a mir____ and a ra____ lay cro____. À yel____ dressing-
gown, ungi____, was sust____ gently beh____ him b____ the mi____ morning
a____. He he____ the bo____ aloft a____ intoned: -Intr____ ad alt____ Dei. ?/27
(Corrigés en Annexe 2)

- 68 Ces extraits font clairement apparaître l'importance de la connaissance des idiomes et des expressions figées.
- 69 On constate aussi que, pour un même texte, si on fait démarrer les trous sur un autre mot, les résultats ne sont pas forcément les mêmes. Mais les différences restent minimes par rapport au cloze test. En effet, il y a eu trois points d'écart en moyenne lorsque deux groupes TD ont essayé le même texte, l'un dans la présentation donnée précédemment, l'autre avec celle qui suit, les écarts entre les plus forts et les plus faibles étant de 5 à 20 fautes pour le premier groupe, et de 8 à 23 fautes pour l'autre.

High taxes and welfare benefits are the real obstacles to job creation in
France. This i____ what t____ government sho____ be thin____ about, rat____
than dub____ schemes f____ sharing exis____ work. Hi____ benefits f____
those o____ of wo____ mean th____ is of____ little ben____ from tak____ a job.
T____ minimum wa____ and he____ labour ta____ have dri____ up t____ cost t____
employers o____ taking o____ new wor____. This i____ exacerbated b____ complex
restri____ on hir____ and fir____. ?/31 (Corrigé en annexe 1)

- 70 Les différentes présentations peuvent alors changer le niveau de difficulté, sans que les écarts soient statistiquement significatifs. Certes, à la fin du paragraphe en question, il y aura incontestablement plus de fautes avec *hiring* et *firing* qu'avec *on* et *and*, mais les difficultés ont tendance à se compenser (ainsi *exacerbated* sera probablement plus difficile à trouver que *by*). Sur quatre paragraphes de même longueur les différences semblent s'égaliser. Il est alors possible de dire si les étudiants d'un groupe sont collectivement (et objectivement) plus ou moins forts que ceux d'un autre.
- 71 Étant donné les niveaux de difficulté parfois imprévisibles, il faut noter par rapport à une norme. Mais les mêmes exercices peuvent être utilisés plusieurs années de suite, et à grande échelle, ce qui permet les comparaisons nécessaires pour se faire une idée précise du niveau absolu. Pour s'en convaincre, il suffit de constater l'utilisation du C-test pour une étude (comparative et multilingue) menée à grande échelle, sur les niveaux de compétence linguistique des étudiants des divers pays européens (Coleman 1996).
- 72 C'est le washback qui constitue l'avantage principal de ce test, qui fonctionne non seulement comme test de compréhension mais aussi comme test indirect des capacités d'expression écrite. En effet, il n'est pas possible de deviner les réponses, et il ne suffit pas de cocher des cases. Au contraire, il faut contribuer de manière relativement active à l'élaboration (la reconstruction) du texte. De plus, l'emploi du discours continu permet de vérifier le raisonnement contextuel, l'emploi de la cohésion, l'implicature, l'anaphore, etc.
- 73 Grâce justement à la continuité du discours, cet exercice fait travailler sur les attentes, et sur les inférences pragmatiques, nécessaires pour l'utilisation des connaissances linguistiques en contexte réel.
- 74 Les résultats au C-test font apparaître des distinctions assez fines pour encourager et pour motiver les étudiants ; ceux-ci peuvent éventuellement utiliser ces exercices pour la révision des textes déjà vus en cours. Il est d'ailleurs possible, avec des étudiants

spécialistes d'autres disciplines, d'utiliser les mêmes textes plusieurs fois, à des intervalles réguliers, afin de mesurer leurs progrès (Connelly 1997).

- 75 En revanche, puisque le C-test présente justement un échantillon représentatif des problèmes linguistiques, il ne semble pas possible de vérifier la connaissance de certains points particuliers de grammaire, comme l'article zéro, ou la présence ou l'absence de l'auxiliaire dans la formation du passé composé. Pour tenir compte de ce type de problème, il faut utiliser un autre type de test intégré.

La lecture des épreuves

- 76 La lecture (et la correction) des épreuves a été proposée plus récemment que le C-test, comme test linguistique ; ce format n'a pas encore été étudié de manière détaillée. Il s'agit de proposer aux candidats un texte comportant une (et une seule) faute par ligne. Les fautes peuvent être de types différents. On peut non seulement y introduire des fautes d'orthographe ou de morphologie, mais aussi rajouter un mot au texte original, ou enlever un mot grammaticalement nécessaire. Il est également possible d'inverser l'ordre des mots ou de remplacer un mot par un autre. L'exemple suivant contient des fautes de tous ces types :

- 1) ...Important taxes and welfare benefits are the real
- 2) ...obstacles to jobs creation in France. This is what
- 3) ...the government should be thinking about, rather that
- 4) ...dubious schemes for share existing work. High
- 5) ...benefits for those out of work mean there often is
- 6) ... a little benefit from taking a job. The minimum wage
- 7) ...and hefty labour taxes driven up the cost to
- 8) ...employers of taking on new worker. This is
- 9) ... exacerbated by complexe restrictions on hiring and firing.

- 77 Selon les règles couramment utilisées, les candidats sont prévenus qu'un certain nombre de lignes (en général un maximum de 20 %) peuvent ne pas comporter de faute (ils ne savent pas lesquelles). Leur tâche consiste à repérer les fautes, les souligner dans le texte (éventuellement en utilisant des signes typographiques conventionnels), et inscrire la correction adéquate dans l'emplacement prévu à gauche (éventuellement utilisé pour qualifier la ligne de « correct »). On s'aperçoit très vite que la plupart des candidats ont des difficultés pour repérer les fautes à corriger, et proposent souvent de modifier des mots ou des structures qui ne comportent aucune erreur. Comme avec les autres tests intégrés, celui-ci permet d'évaluer les compétences linguistiques de manière assez détaillée, tout en traitant d'un sujet en rapport avec la discipline principale.
- 78 Il faut certes admettre que le fait de travailler avec des textes qui comportent des erreurs n'est pas sans danger ; en effet, cela va à l'encontre des traditions pédagogiques selon lesquelles les modèles proposés aux élèves doivent toujours être impeccables. Néanmoins, cette objection (cognitivist) ne suffit pas pour faire renoncer à l'exercice. En effet, les imperfections ne peuvent jamais être éliminées, ni chez les « natifs », ni a fortiori chez les étrangers. La correction des erreurs constitue donc une activité linguistique authentique, utile et nécessaire pour tout travail de rédaction, même (surtout) chez les professionnels. Puisqu'il s'agit justement d'une activité authentique, le taux de validité semble élevé. Les

notes obtenues sont en étroite corrélation avec les résultats aux autres tests intégrés, et correspondent bien avec les capacités en expression écrite.

- 79 Le niveau de difficulté varie plus avec la correction des épreuves qu'avec le C-test, car le professeur peut contrôler librement la difficulté de chaque question. Mais étant donné la variété des questions possibles, le format peut être utile à tous les niveaux. Avec des questions portant sur les difficultés habituelles des étudiants de première année, les résultats à un test récent étaient de 9 à 31 (sur quarante), avec une moyenne de 18. Les notes sur vingt, calculées à l'aide d'un barème adéquat, sont souvent les mêmes qu'avec le C-test, les écarts ne dépassant pas un ou deux points. Les résultats permettent ainsi de confirmer le classement général déjà établi.
- 80 La notation est en principe objective, ce qui implique un taux élevé de fiabilité. Il existe pourtant quelques problèmes dont il faut tenir compte. On peut difficilement éviter des cas où il y a deux réponses admissibles, par exemple lorsqu'il serait possible soit d'ajouter un « s » à la troisième personne, soit de mettre le sujet du verbe au pluriel ; ou bien lorsqu'on pourrait soit ajouter un article défini, soit mettre un nom au pluriel. D'autre part, les candidats sont souvent amenés à proposer une modification dans une ligne où il n'y a pas de faute, sans que le résultat soit nécessairement faux non plus. Les correcteurs peuvent cependant se mettre d'accord à l'avance sur la marche à suivre, car le contexte linguistique, fourni par le discours continu, permet la plupart du temps de préférer une solution à l'autre.
- 81 Plus grave est le nombre de questions, limité à vingt par page environ, si le texte est imprimé en double interligne. Étant donné la place laissée à la chance, lorsqu'il y a si peu de questions, cet exercice ne peut être conseillé comme système unique d'évaluation ; il doit être complété par d'autres formats.
- 82 Il s'agit d'un test utile non seulement pour l'évaluation, mais aussi pour le travail en classe. On trouve justement des exercices semblables dans le *New Cambridge Advanced English* (Jones 1998), utilisés pour vérifier l'orthographe et la ponctuation. On en trouve également dans les *Sample Papers* pour le tout nouveau *Business English Certificate* (UCLES, Université de Cambridge). Bien qu'il y figure dans la partie « compréhension écrite », il semble en fait tout aussi utile comme exercice de préparation à l'expression. Il incite l'étudiant à relire son texte en détail, et le rend conscient de ses propres faiblesses, en lui fournissant des informations utiles, d'une manière simple et directe.
- 83 Selon l'hypothèse de la compétence unitaire en langue, les résultats des tests à l'écrit devraient permettre sous certaines conditions de prévoir de manière indirecte les capacités à l'oral. En effet, les tests intégrés font surtout appel aux attentes et aux inférences pragmatiques, nécessaires pour toute activité linguistique authentique. On ne s'étonne donc pas de constater une corrélation entre les capacités à l'écrit et à l'oral. Cependant, la corrélation n'implique pas une relation nécessaire. S'il n'y a pas de test prévu, spécifique à l'oral, l'étudiant ne sera pas incité à travailler sur cet aspect de la langue. On a vu que le taux de fiabilité des exposés et des entretiens/interrogations n'est pas acceptable, et que les résultats à ses tests ne sont pas assez précis. On a donc proposé d'utiliser la dictée comme test intégré.

Dictée

- 84 Certes, lorsqu'on a affaire à un petit groupe d'étudiants, la dictée est très facile à organiser, mais pour des effectifs plus importants, divisés en plusieurs groupes, ou dans un amphithéâtre, il est nécessaire de prévoir des cassettes ou des haut-parleurs, afin d'assurer l'égalité de traitement. Puisqu'il s'agit d'encourager le travail sur la langue orale, l'effort supplémentaire peut se révéler utile.
- 85 Pour beaucoup, la dictée n'est plus qu'un exercice scolaire et suranné, depuis longtemps passé de mode. Cette fausse impression n'est pas sans poser des problèmes au niveau de la validité apparente, du moins aux yeux des candidats. Certes, si le texte est lu trop lentement, mot par mot, l'exercice paraîtra simpliste. En revanche, s'il est lu à une vitesse normale, par un anglophone, par phrase entière ou par dizaine de mots (ou *a fortiori* s'il s'agit d'un enregistrement authentique), il devient vite difficile de reconnaître les mots qu'on croyait connaître. Le problème du découpage des mots devient alors très difficile, ce qui montre qu'il ne s'agit pas d'un simple test d'orthographe. Il s'agit en fait d'un test de mémoire à court terme ; en effet, puisque le candidat ne peut pas réécouter les phrases à volonté, il doit les garder en mémoire le temps de les écrire. Il est donc obligé de tenir compte en même temps du sens et de la forme, en intégrant ses connaissances aux deux niveaux. Il s'agit d'un véritable test intégré, dont la validité réelle est tout à fait acceptable.
- 86 À première vue, la notation semble objective, ce qui indique en principe une fiabilité tout aussi acceptable. Le nombre de questions correspond au nombre de mots dans le texte, ce qui semble largement suffisant pour réduire au minimum l'élément de chance. Cependant, la notation implique souvent un jugement subjectif. Comme c'était le cas de la procédure cloze, il faut pouvoir distinguer par exemple entre un candidat qui trouve le bon mot mais qui fait une faute d'orthographe, et un autre qui n'a apparemment rien compris du tout. Pire, l'orthographe phonétique, qui certes vaut mieux que rien, pourrait en même temps signifier un manque total de compréhension. La fiabilité de cet exercice ne semble donc pas évidente.
- 87 La dictée permet de distinguer entre les forts et les faibles, mais pas d'établir des distinctions fines. Il est nécessaire de choisir un texte qui convient aux étudiants, et de le lire à une vitesse adaptée. Or, un texte qui convient aux niveaux faibles ne conviendra pas forcément aux niveaux forts, et réciproquement. Malgré ses faiblesses, on peut affirmer que la dictée présente des effets en retour favorables. Il s'agit en effet de la seule façon pratique d'évaluer la capacité réceptive à l'oral de manière objective. Sans cette sanction, l'oral ne sera jamais pris au sérieux par les étudiants. Il faut néanmoins reconnaître que, si la dictée est un instrument adéquat pour tester les capacités en compréhension, elle ne permet pas de savoir si le candidat sait parler en langue étrangère ou non. Pour cela, il faudrait un test intégré qui donnerait l'occasion aux étudiants de s'exprimer à l'oral.

Lecture à haute voix

- 88 Le fait de comprendre la langue orale est évidemment un préalable à toute tentative d'expression. Pourtant, si les résultats à la dictée permettent de prévoir de manière indirecte les capacités à l'expression orale, cela n'incite pas forcément les étudiants à s'y

entraîner. Il serait incontestablement utile de disposer d'une forme de test qui permettrait d'évaluer avec plus d'objectivité et avec plus de précision les compétences en phonétique.

- 89 Certes, la lecture à haute voix constitue un exercice difficile, même en langue maternelle. Il suffit cependant d'entendre quelques mots pour distinguer le « natif » de l'étranger. Sur un paragraphe d'une centaine de mots, cet exercice devrait permettre l'évaluation des étudiants en fonction des critères linguistiques objectifs et directement observables. Il ne semble pas impossible de définir des critères pertinents, en rapport avec les règles élémentaires de la phonétique, sans toutefois exiger des connaissances théoriques à ce sujet. On pourrait tenir compte non seulement du débit et de la clarté, mais aussi de la prononciation des voyelles, des diphtongues, des formes réduites, de l'intonation, et ainsi de suite. Il semble d'ailleurs indispensable de travailler avec du discours continu, plutôt qu'avec des mots ou des phrases isolés. En effet, le placement de l'accent tonique et des accents secondaires dépend des enchaînements et du sens.
- 90 Malheureusement, un tel projet se heurte immédiatement à plusieurs difficultés. La prononciation peut varier, non seulement selon les accents régionaux, mais aussi selon les différentes interprétations possibles. Cela oblige à juger non pas les sons bruts produits individuellement mais plutôt les oppositions systématiques qui assurent la cohérence de la présentation. La notation ne peut donc pas être purement objective ; elle nécessite le jugement d'un expert, de niveau langue maternelle, compétent en phonétique. Pourtant, même les experts ne sont pas d'accord entre eux sur l'importance relative des différents points à évaluer. Pire, la grammaire de l'oral est moins enseignée que la syntaxe, et est donc moins connue des enseignants comme des élèves. L'élaboration d'un barème adéquat semble donc difficile.
- 91 En outre, sur le plan purement pratique, la notation en temps réel semble presque irréalisable. Certes, en écoutant n'importe quel candidat, on peut relever tout de suite un certain nombre de fautes, plus ou moins graves, ce qui rend possible une évaluation impressionniste et qui permet de donner des conseils pédagogiques. Cependant, pour obtenir une notation objective et chiffrée, à partir d'un fragment de discours, il serait préférable de pouvoir relever toutes les imperfections. L'enregistrement systématique semble donc indispensable. En effet, même s'il était possible de souligner chaque syllabe mal prononcée dès la première écoute, il ne serait toujours pas possible de préciser le type d'erreur visée. Pour simplifier la tâche, on pourrait proposer de se mettre d'accord à l'avance sur les points importants, dont il faudrait tenir compte en priorité. Mais d'une part, l'accord est loin d'être évident, et de l'autre, si le test porte uniquement sur les points difficiles, il ne serait plus abordable pour les étudiants faibles (majoritaires), qui risquent alors d'obtenir trop souvent la note de zéro.
- 92 En l'absence à l'heure actuelle dans la quasi-totalité des Facultés du temps et du matériel nécessaire pour l'enregistrement systématique des prestations, il faut donc renoncer pour le moment à l'utilisation de ce type d'exercice comme test d'évaluation.

Conclusion

- 93 La recherche pertinente, confirmée par l'expérience sur le terrain, montre l'insuffisance des systèmes couramment utilisés pour l'évaluation linguistique des spécialistes des autres disciplines.

- 94 Malgré tous les efforts de la part des enseignants, leurs étudiants prendront toujours plus au sérieux les examens que les cours ; ils travaillent toujours avec plus de concentration lorsqu'il s'agit de préparer un examen (et a fortiori pendant cet examen). Au lieu de déplorer cette attitude en fin de compte raisonnable, il faut essayer d'en profiter.
- 95 Il faut éviter de travailler sur des mots ou des phrases isolés, hors contexte ; de tels exercices ne permettent pas de tenir compte du contenu du cours, et n'entraînent pas à l'utilisation des connaissances en situation authentique. Pour l'évaluation objective des compétences en langue, il faut également éviter les tests fondés sur l'expression libre ; ces exercices sont de toute façon trop difficiles pour la plupart des inscrits. En revanche, les tests fondés sur le discours continu permettent de cumuler les avantages des deux approches. Ces tests permettent de travailler sur l'interface entre les niveaux pragmatique et sémantique, afin de vérifier (et de mesurer) les compétences linguistiques générales ; ils font ainsi appel aux inférences pragmatiques, ce qui nécessite l'utilisation des connaissances en contexte réel.
- 96 Les différents tests sur le discours continu donnent à peu près les mêmes résultats, et permettent d'établir sensiblement les mêmes classements, ce qui semble corroborer à la fois la validité du test et la fiabilité de la notation. Toutefois, ces tests ont chacun des effets en retour spécifiques, légèrement différents ; ils permettent de travailler sur des problèmes linguistiques différents. Il convient donc d'utiliser autant que possible tout l'arsenal des différents formats à disposition. Ceci aura en même temps l'effet de minimiser l'élément de chance inhérente à tout système d'examen.
- 97 La notation objective permet d'établir des distinctions fines à tous les niveaux, en fournissant des informations très précises, ce qui facilite le diagnostic des faiblesses, et le travail en autonomie guidée. Il semble en effet que le travail individuel avec ce type d'exercice puisse souvent remplacer avantageusement de nombreuses séances de travail en groupe TD classique.
- 98 Bien sûr, le fait de reconnaître le taux de validité plus élevé des exercices construits à partir du discours continu n'implique pas le refus d'enseigner la grammaire, parfois nécessaire en 2e voire 3e cycles, même lorsque les objectifs adoptés portent en principe sur la communication. De même, reconnaître la plus grande fiabilité des tests intégrés n'empêche nullement l'utilisation des dissertations, des exposés, ou de la discussion en classe ; ces exercices, qui correspondent à l'objectif final du cours, sont indispensables pour le traitement des sujets en rapport avec la discipline principale, ainsi que pour l'étude de la culture et des institutions du pays étranger.
- 99 Néanmoins, pour l'attribution d'une note unique sur vingt, surtout en ce qui concerne les (nombreux) étudiants moyens et faibles, il y a de bonnes raisons de préférer le travail sur le discours continu. Les effets en retour favorables justifient l'utilisation de ces tests « intégrés », à la fois pour l'évaluation et pour l'apprentissage. Puisque la notation est purement objective, ils sont donc particulièrement indiqués pour le télé-enseignement ou pour l'auto-évaluation en dehors du cours. La commodité d'organisation et la plus grande facilité de notation, par rapport aux examens classiques, permettent en outre une utilisation plus fréquente, en contrôle continu. Du point de vue pédagogique, ceci devrait être à l'avantage de tous.

BIBLIOGRAPHIE

- Alderson, C. & B. North (dir.). 1991. *Language Testing in the 90s —the communicative legacy*. Londres : Modern English Publications.
- Arena, L.A. (dir.). 1990. *Language Proficiency —defining, teaching and testing*. New York : Plenum.
- Bachman, L. F. 1989. *Fundamental Considerations in Language Testing*. Oxford : Oxford University Press.
- Bailey, K.M. 1996. « Working for washback: A review of the washback concept in language testing ». *Language Testing* 13, 257-79.
- Brumfit, C.J. & K. Johnson (dir.). 1979. *The Communicative Approach to Language Teaching*. Oxford : Oxford University Press.
- Cohen, A.D., Segal, M. & R.W. Bar-Siman-Tov. 1984. « The C-test in Hebrew ». *Language Testing* 1/2, 221-25.
- Coleman, J.A. 1996. *Studying Languages —the proficiency, background and motivation of students of foreign languages in the UK and Europe*. Londres : CILT.
- Connelly, M. 1997. « Using C-tests in English with post-graduate students ». *English for Specific Purposes* 16/2, 139-50.
- Culhane, T., Klein-Braley, C. & D.K. Stevenson (dir.). 1984. *Practice and Problems in Language Testing*. Colchester : University of Essex.
- Ducrot, O. 1972. *Dire et ne pas Dire*. Paris : Hermann.
- « Focus on Evaluation and measurement », 1992. *Proceedings of the National Research Symposium on Limited English Proficient Student Issues*, Washington.
- Hood, M.A.G. 1990. « The C-test, a viable alternative to the use of the cloze procedure in testing? ». In Arena L.A. (dir.), *Language Proficiency —defining, teaching and testing*. New York : Plenum, 173-190.
- Howatt, A.P.R. 1984. *A history of English Language Teaching*. Oxford : Oxford University Press.
- Hughes, A. 1989. *Testing for Language Teachers*. Cambridge : Cambridge University Press.
- Isani, S. 1996. « Les épreuves de langues dans les concours de la fonction publique —domaines, critères et objectifs ». *ASp*, 11-14, 351-61.
- Jafapur, A. 1995. « Is C-testing superior to Cloze? ». *English for Specific Purposes* 12/2, 194-214.
- Jones, Leo. 1998. *New Cambridge Advanced English* (new ed.). Cambridge : Cambridge University Press.
- Klein-Braley, C. 1984. « Advance prediction of difficulty with C-tests ». In Culhane, Klein-Braley & Stevenson (dir.), *Practice and Problems in Language Testing*. Colchester : University of Essex.
- Klein-Braley, C. 1985. « A Cloze-up on the C-test —a study in the content validation of authentic tests ». *Language Testing* 2/1, 76-104.
- Klein-Braley, C. 1997. « The C-test in the context of reduced redundancy testing ». *Language Testing* 14/1, 47-84.

- Klein-Braley, C. & U. Raatz. 1984. « A survey of research on the C-test ». *Language Testing* 1/2, 134-46.
- Morrow, K. 1979. « Communicative language testing – revolution or evolution? ». In Brumfit & Johnson (dir.), *The communicative Approach to Language Teaching*. Oxford : Oxford University Press.
- Oller, J.W. 1979. *Language Tests at School*. Londres : Longman.
- Oller, J.W. (dir.). 1983. *Issues in language testing research*, Rowley, Mass : Newbury,
- Pollit, A. 1991. « Giving students a sporting chance —assessment by counting and assessment by judging ». In Alderson & North (dir.), *Language testing in the 90s —the communicative legacy*, London : Modern English Publications, 46.
- Raatz, U. 1984. « The factorial validity of C-tests ». In Culhane, Klein-Braley & Stevenson (dir.), *Practice and Problems in Language Testing*. Colchester : University of Essex.
- Raatz, U. « Better theory for better tests ». *Language Testing* 2/1, 60-75, 1985.
- Spolsky, B. 1995. *Measured Words*. Oxford : Oxford University Press.
- Taylor, W.L. 1953. « Cloze procedure: A new tool for measuring readability ». *Journalism Quarterly* 30, 415-33.
- Weir, C.J. 1990. *Communicative Language Testing*. Hemel Hempstead : Prentice Hall.
- Weir, C.J. 1993. *Understanding and Developing Language Tests*. Hemel Hempstead : Prentice Hall.

ANNEXES

Annexe 1

High taxes and welfare benefits are the real obstacles to job creation in France. This is what the government should be thinking about, rather than dubious schemes for sharing existing work. High benefits for those out of work mean there is often little benefit from taking a job. The minimum wage and hefty labour taxes have driven up the cost to employers of taking on new workers. This is exacerbated by complex restrictions on hiring and firing. (*Financial Times* 1997)

Annexe 2

Under certain circumstances there are few hours in life more agreeable than the hour dedicated to the ceremony known as afternoon tea. There are circumstances in which, whether you partake of tea or not —some people of course never do— the situation is in itself delightful. Those that I have in mind in beginning to unfold this simple history offered an admirable setting to an innocent pastime. (Henry James, *Portrait of a Lady*, p. 5).

“Yes, of course, if it's fine to-morrow”, said Mrs Ramsey. “But you'll have to be up with the lark”, she added.

To her son these words conveyed an extraordinary joy, as if it were settled the expedition were bound to take place, and the wonder to which he had looked forward, for years and years it seemed, was, after a night's darkness and a day's sail, within touch. (Virginia Wolf, *To the Lighthouse*, p. 9).

Stately, plump Buck Mulligan came from the stairhead, bearing a bowl of lather on which a mirror and a razor lay crossed. A yellow dressing-gown, ungirdled, was sustained gently

behind him by the mild morning air. He held the bowl aloft and intoned: "Introibo ad altare Dei". (James Joyce, *Ulysses*, p. 9).

NOTES

1. Voir par exemple la bibliographie de Spolsky (1995).
2. En tout cas, depuis l'existence des examens écrits. Selon le *Northcote/Trevelyan Report* (1853), « *It is notorious that the examinations for Trinity fellowships have, directly and indirectly, done much to give a direction to the studies of Cambridge and of all the university schools who are the feeders of Cambridge* ».
3. Weir (1993 : 154) fait état des divergences allant largement au-delà de dix points sur vingt, constatées lorsque vingt jeunes enseignants « natifs » ont chacun corrigé huit copies (fournies en facsimile en annexe à son livre).
4. Malgré les complications introduites par la méthode analytique, il ne semble pas que celle-ci a permis d'améliorer sensiblement le taux de fiabilité, par rapport à la méthode holistique utilisée par des correcteurs expérimentés (Spolsky 1995 : 322).
5. Howatt (1984) rappelle l'existence, à côté des méthodes « grammaire/traduction », de méthodes dites « notionnelles » (J. Wilkins 1668), situationnelles (Miège 1695), « naturelles » (Sauveur 1874), « directes » (Passy 1899, Berlitz 1882) ; sans oublier la méthode (non scolaire) préconisée par le grand pédagogue Comenius (1657).
6. Il semble trop simpliste d'évoquer ici la distinction entre la forme et le contenu, ou entre les connaissances « déclaratives » et « procédurales ». Selon la terminologie utilisée par Ducrot (1972), la phrase correspond à un concept abstrait, auquel est attribué une signification sémantique, alors que l'énoncé constitue un fait observable, associé à un sens intentionnel.
7. La corrélation étroite entre les résultats obtenus aux tests intégrés et aux tests plus détaillés ne constitue pas une preuve de l'hypothèse en question, même si cela semble la corroborer. La corrélation serait due plutôt à l'expérience linguistique, à l'enseignement reçu ou aux capacités non linguistiques (Oller 1983).
8. Selon Oller (1983 : Introduction), le terme « cloze » était choisi à l'origine pour marquer le rapport avec le concept psychologique de « *gestalt closure* » (Taylor 1953). Klein-Braley (1997) confirme à son tour que le « C » de « C-test » marque le rapport avec le mot « cloze ».
9. Hood (1990) relève pourtant quelques fautes chez les « natifs » (des étudiants américains) : il s'agit essentiellement de fautes d'orthographe.
10. Mis à part les cas d'absence totale de réponse et l'utilisation de mots inexistants ou sans aucun rapport, on trouve de nombreux contre-sens (*no* à la place de *new*) ; de simples fautes d'orthographe (**governement*) ; des fautes de vocabulaire (*minimal* à la place de *minimum*, *count* à la place de *cost*, *schema* à la place de *schemes*) et surtout de nombreuses fautes de morphologie et/ou de syntaxe (*employment* ou *employee* à la place d'*employers*, *where* ou *why* à la place de *what*, *exacerbating* ou *exacerbates* à la place d'*exacerbated*, *labors* à la place de *labour...*), sans parler des « s » à la troisième personne.

RÉSUMÉS

Les systèmes traditionnels utilisés pour l'évaluation des compétences linguistiques chez les spécialistes des autres disciplines sont peu satisfaisants, de plusieurs points de vue. Les tests fondés sur des questions ponctuelles ont un taux de validité très bas, alors que les tests fondés sur l'expression libre sont peu fiables. En revanche, les tests intégrés, fondés sur le discours continu, comme la procédure cloze, le C-test, la lecture des épreuves, la dictée, et peut-être la lecture à haute voix, semblent cumuler les avantages des deux types de tests, tout en évitant les défauts. En outre, ils semblent plus utiles du point de vue pédagogique, leurs effets en retour favorables justifiant ainsi leur utilisation en cours.

The testing methods traditionally used for the evaluation of the linguistic competence of students specialised in other disciplines can be shown to be unsatisfactory from various points of view. Tests based on discrete point items have low validity, while exercises involving free expression have poor reliability. On the other hand, integrated tests, based on continuous discourse, like the cloze procedure, the C-test, proof-reading, dictation, and possibly reading aloud, seem to combine the advantages of both types of test, while avoiding the related drawbacks. In addition, they seem to be more useful from the pedagogical point of view. Their favourable washback is demonstrated by the fact that they can be used advantageously as classroom exercises.

INDEX

Keywords : evaluation, reliability, test, validity, washback

Mots-clés : effet en retour, évaluation (linguistique), fiabilité, test, validité

AUTEUR

ROSS CHARNOCK

Ross Charnock a préparé sa thèse à l'EHESS, sous la direction du Professeur Oswald Ducrot. Il est responsable depuis 1990 du Service Langues Vivantes à la Faculté de Droit, de Sciences Économiques et de Gestion, à l'Université de Rouen. Il a également enseigné en LEA et à l'École Supérieure d'Électricité. En plus de ses travaux en linguistique et en linguistique appliquée, il a publié de nombreux articles sur la musique classique dans la revue internationale spécialisée *The Strad*. Il assure actuellement la traduction anglaise d'un ouvrage spécialisé sur l'histoire de l'archet en France. charnock@dauphine.fr