Chapitre 8

Quelle place pour la notation automatique de productions écrites dans un test standardisé de français langue étrangère ?

Dominique CASANOVA[1], Alhassane AW[1],
Marc DEMEUSE[2]

1. Introduction

La notation des épreuves d'expression écrite et orale des tests de langue à forts enjeux présente un cout généralement élevé, du fait de la mobilisation d'évaluateurs humains qualifiés. De surcroit, celle-ci étant humaine, elle est empreinte de subjectivité et requiert en général des évaluations multiples pour garantir la fidélité des résultats, ce qui accroit d'autant les frais et le temps de traitement. Or, les épreuves d'expression écrite se déroulent de plus en plus fréquemment sur ordinateur, phénomène qui s'est accéléré lors de la pandémie de la COVID-19 pour limiter les échanges physiques entre personnes. Cela donne, aux organismes concepteurs de tests, l'opportunité de constituer des corpus de productions au format numérique, qui peuvent être exploités à des fins de recherche ou d'amélioration de la qualité de ces tests. Le développement des méthodes, outils et recherches en traitement automatique des langues et en intelligence artificielle rend également accessible l'élaboration de systèmes de notation automatique qui existaient jusqu'ici principalement en langue anglaise. C'est ainsi que *Le français des affaires* a conçu un premier prototype pour la notation automatique pour l'épreuve d'expression écrite d'un test standardisé de français langue étrangère.

La notation humaine et la notation automatique ne peuvent cependant pas être considérées comme équivalentes (Attali, 2013), quand bien

[1] Le français des affaires, CCI Paris Ile-de-France.
[2] Université de Mons (Belgique).

même elles conduiraient à des classements similaires (Bennet & Bejar, 1997). Le construit évalué par chacun des deux systèmes de notation diffère. Remplacer l'humain par la machine serait prendre une décision en négligeant l'évaluation de certains aspects de la compétence à écrire, plus complexes à évaluer automatiquement, comme l'utilisation de figures de rhétorique telles que l'ironie ou le second degré. Une réflexion doit donc être menée par les concepteurs de tests qui souhaitent intégrer une part de notation automatique dans leurs dispositifs d'évaluation. Cette réflexion porte sur l'usage envisagé de la notation automatique, sur les moyens de rendre compte de sa pertinence et sur son articulation possible avec l'évaluation humaine.

2. Évaluation humaine versus évaluation automatique

L'évaluation automatique est souvent présentée en opposition à l'évaluation humaine, qui est notoirement imparfaite. Mais l'évaluation automatique a ses propres lacunes : si elle offre la possibilité d'une évaluation objective, elle ne permet, aujourd'hui, d'appréhender que de façon médiocre certains aspects des écrits qui peuvent être caractéristiques du construit évalué.

2.1 L'évaluation par des humains, une activité hautement cognitive

Le processus d'évaluation est un processus complexe, qui mobilise un ensemble d'actions cognitives et métacognitives. Il se situe au sommet de la taxonomie de Bloom. La modélisation des processus cognitifs mobilisés reste à approfondir pour une meilleure compréhension de leur impact sur la notation. L'image générale qui se dégage des modèles proposés pour l'acte évaluatif dans le domaine médical (Gauthier et al., 2016) et dans le domaine des langues (Bejar, 2012 ; Han, 2016 ; Wolfe, 2005) est la suivante : l'évaluateur possède une représentation interne du modèle théorique de la compétence à évaluer et des degrés de maitrise de cette compétence, teintée par son expérience professionnelle et sociale, et sa compréhension des cadres de référence officiels ou auxquels il se réfère par ailleurs dans son activité.

Cette représentation s'élabore notamment lors des sessions de formation et de standardisation auxquelles il peut participer en amont des sessions d'évaluation. Il réactive, à la lecture de la grille d'évaluation, cette représentation et celle des critères à considérer dans le contexte du test qu'il doit évaluer. S'il est mis en présence du candidat ou peut observer ce dernier, il active inconsciemment ses filtres perceptifs et se fait une

représentation catégorielle (sociale, culturelle. . .) du candidat (Macrae & Bodenhausen, 2001). Le début de la performance vient renforcer ou activer, si l'évaluateur n'a pas eu l'opportunité d'observer le candidat au préalable, cette perception catégorielle dont il doit se départir pour tendre à l'objectivité.

Étonnamment, cette dimension n'est pas prise en compte dans les modèles précités propres à l'évaluation en langue. Pourtant, cette représentation du candidat par l'évaluateur à travers un ensemble de filtres catégoriels est difficilement évitable, y compris dans des épreuves d'expression écrite où l'évaluateur n'est pourtant pas en contact avec le candidat. Par exemple, si la tâche consiste en la rédaction d'une lettre formelle (pour formuler une demande, donner un avis, informer. . .), la phrase d'adresse au lecteur peut être très marquée culturellement (par un style très direct ou, au contraire, particulièrement ampoulé), ce qui ne manquera pas d'être relevé par l'évaluateur, même s'il sait que cela ne doit pas entrer en ligne de compte dans sa notation. De même, si la lettre est rédigée à la première personne, la présence d'adjectifs attributs ou de participes passés pourra informer l'évaluateur sur le sexe du rédacteur, qui notera cette information tout en sachant que sa prise en compte doit être limitée à la vérification du respect de l'accord en genre tout au long du texte.

L'observation permet à l'évaluateur de se construire une image mentale de la performance, perçue à travers les filtres des aspects qu'il considère pertinents pour l'évaluation de la compétence, image qu'il réajuste au fil de l'observation. Il confronte cette image à sa représentation interne de la compétence à évaluer et à des exemples de performances stéréotypées ou passées pour catégoriser l'information recueillie. Il intègre progressivement cette information pour quantifier la performance et finaliser sa prise de décision en la justifiant.

Cette étape d'intégration est fortement dépendante de l'instrumentation de l'évaluateur (le type de grille d'évaluation utilisé, par exemple analytique ou holistique, et les descripteurs qu'elle comporte) (Lumley, 2002) et de la nécessité ou non de justifier l'évaluation en la commentant. Tout au long du processus, l'évaluateur subit l'influence d'un ensemble de caractéristiques qui lui sont propres (sa vision du monde, ses connaissances générales), qui introduisent une variabilité non souhaitée dans ses jugements.

2.2 Les différences entre évaluateurs, source de variation des scores

La docimologie critique a depuis longtemps mis en évidence la variabilité des évaluations humaines (Leclercq et al., 2004 ; Martin, 2002).

Cette connaissance et la sensibilisation des évaluateurs à cette problématique n'ont cependant pas suffi à en réduire l'impact (Suchaut, 2008).

Pour cela, il faut être en mesure d'identifier plus précisément les sources de ces variations, notamment sur le plan cognitif. Gingerich et al. (2014) distinguent trois perspectives différentes d'appréhension des origines et des solutions envisageables à la question de la variabilité des jugements évaluatifs (dans le domaine médical). Selon la première perspective, les variations sont principalement dues à des comportements qui peuvent évoluer au moyen d'actions de formation et d'une meilleure instrumentation. De nombreux travaux ont cependant montré l'effet limité de la formation sur la variation des scores, que ce soit pour l'évaluation des compétences langagières (Lumley & McNamara, 1995 ; Weigle, 1998) ou dans d'autres cadres, comme l'évaluation des performances au travail (Landy & Farr, 1980). La deuxième perspective voit dans ces variations le résultat des limitations de la cognition humaine et du fait que les évaluateurs sont prompts à être influencés par leur contexte immédiat. Les capacités de mémoire de travail étant réduites, l'information est rapidement perdue, à moins d'être traitée et rattachée aux structures de connaissance de l'évaluateur pour être retenue et exploitée à des fins de notation (van Merriënboer & Sweller, 2010). Par ailleurs, l'être humain a aisément tendance à produire des jugements comparatifs et donc, à être influencé par la performance précédente (effet de contraste). La troisième perspective concerne davantage des situations dont la complexité ou la spécificité nécessitent ou justifient le recours à l'expérience individuelle de l'évaluateur, comme dans le cas de l'évaluation sur le lieu de travail, dans le domaine médical (Gingerich et al., 2014). Dans de telles situations, non standardisées du fait de leur emprise avec le réel, différents évaluateurs peuvent former des interprétations différentes, mais également légitimes et pertinentes, avec pour conséquence des différences de notation.

La première perspective a largement été explorée dans le domaine des langues étrangères. La formation des évaluateurs n'a montré qu'un impact limité sur la réduction de variabilité interévaluateurs (Lumley & McNamara, 1995 ; Weigle, 1998). Si elle semble réduire les tendances extrêmes à la sévérité ou à l'indulgence et favoriser la stabilité des évaluateurs, c'est-à-dire la constance avec laquelle ils font preuve de sévérité ou d'indulgence, ces derniers ne deviennent pas pour autant interchangeables. Les retours (in)formatifs individuels, quoiqu'appréciés par les évaluateurs, semblent également avoir un impact mitigé (Elder et al., 2005). La formation et l'accompagnement restent utiles et nécessaires, mais ils sont loin de remédier à la présence d'écarts de notation. En favorisant la stabilité des évaluateurs, la formation et l'accompagnement permettent toutefois de modéliser plus efficacement les biais de sévérité et d'en tenir

compte dans l'expression d'un score ajusté (Linacre, 1989), mais non parfait, d'autres facteurs intervenant dans la variabilité des résultats. Pour aller au-delà, il faut mieux comprendre les processus cognitifs mobilisés lors de l'évaluation et pourquoi ils sont mobilisés différemment selon les évaluateurs.

Une grande attention a notamment été portée aux éléments sur lesquels les évaluateurs fondent leur jugement. Différentes études ont montré que ces derniers avaient tendance à se concentrer sur des aspects différents de la performance ou avaient une interprétation différente des critères d'évaluation ou des exigences, en dépit de la formation reçue et de leur expérience en évaluation (Ang-Aw & Goh, 2011 ; Ince, 2022 ; Orr, 2002). Les grilles d'évaluation sont censées guider les évaluateurs dans les aspects de la performance à considérer pour l'évaluation, mais ceux-ci accordent plus ou moins d'importance aux différents critères. Dans le cas d'évaluations basées sur des échelles holistiques, cela peut conduire à des décisions très différentes (Barkaoui, 2010), mais cela impacte également les évaluations recourant à des grilles analytiques (Eckes, 2008). Même lorsque les évaluateurs semblent s'accorder sur les aspects à prendre en compte pour l'évaluation, des différences de notation peuvent être mises en évidence du fait d'une interprétation différente de certaines caractéristiques de la performance (Brown et al., 2005). Une explication avancée par Han (2016) est que les évaluateurs, s'appuyant sur leurs expériences personnelles, professionnelles et culturelles, ont une représentation ancrée de ce qui compose le construit de l'épreuve. Cette connaissance est stockée dans leur mémoire à long terme. En dépit de la formation qu'ils peuvent recevoir, la représentation qu'ils se font des critères et des exigences du test est susceptible d'être influencée par cette connaissance ancrée, à laquelle ils accèdent également lors de leur prise de décision. D'autres études ont cherché à mettre en évidence, à travers l'analyse de rapports verbaux, la fréquence de mobilisation de différents processus cognitifs lors du processus d'évaluation (Wolfe, 1997, 2005 ; Wolfe et al. 1998).

Dans l'analyse des facteurs pouvant être à l'origine de ces différences, une emphase particulière a été mise sur les caractéristiques individuelles des évaluateurs, leur niveau d'expertise en évaluation et leur formation à l'acte évaluatif (Weigle, 2002). Eckes (2008) a montré que les variables contextuelles de l'évaluateur (comme l'âge, le nombre de langues étrangères parlées, le nombre d'années passées à l'étranger, le nombre d'années d'activité en tant qu'évaluateur et le nombre de sessions d'évaluation auxquelles il a participé) expliquent en partie les différences de profil de notation. Wolfe (2005) a également montré que, selon leur degré d'expertise, des évaluateurs présentaient des différences dans les informations qu'ils prenaient en compte pour l'évaluation et leur traitement.

Néanmoins, les processus mobilisés dépendent également de la personnalité de l'évaluateur et de son style cognitif. Scott et Bruce (1995) ont mis en évidence l'existence de cinq types principaux de stratégie de prise de décision, qui peuvent influer sur le jugement de l'évaluateur. Selon son style cognitif, un évaluateur peut notamment être plus à l'aise avec une grille holistique qu'avec une grille analytique. Barkaoui (2010) a mis en évidence que l'utilisation de grilles différentes pouvait avoir un effet plus important que l'expérience des évaluateurs sur leur comportement lors de la prise de décision et sur les aspects de la copie auxquels ils portent leur attention. Une grille analytique semble préférable pour des évaluateurs peu expérimentés du fait qu'elle contribue à fixer leur attention sur la tâche et les critères d'évaluation, à alléger la charge cognitive en ne laissant pas aux évaluateurs la responsabilité de pondérer l'importance des différents critères dans leur jugement et à améliorer leur consistance interne.

Les facteurs à prendre en considération sont donc multiples, les modes de fonctionnement sont complexes et ne peuvent être qu'en partie révélés au moyen des rapports verbaux. Par exemple, Isaacs et Trofimovich (2010) ont mis en évidence que la compétence musicale d'évaluateurs novices avait un impact sur la notation du critère d'accentuation de productions orales de personnes dont l'anglais n'était pas la langue maternelle, particulièrement pour les locuteurs de compétence faible en anglais. On peut également s'interroger sur la systématicité de l'influence des caractéristiques individuelles et sur leur perméabilité à des facteurs contextuels. Rappelons par exemple que les résultats donnés par un évaluateur humain à une même copie, à deux occasions différentes suffisamment éloignées dans le temps pour neutraliser l'effet de mémoire, ne sont pas toujours identiques. Il semble donc difficile de trouver, à court terme, un remède à la variabilité des évaluations humaines.

2.3 Les limites de l'évaluation humaine

L'évaluation humaine est sujette à la variabilité, ce qui a un impact direct sur la fidélité des épreuves à base de performance. La validité des jugements peut également être mise en question, aussi bien quand on considère la notation d'une épreuve d'expression écrite sans contact visuel ou auditif avec le candidat que lorsque l'épreuve est passée sur ordinateur (sans confrontation à une écriture manuscrite), évaluée au moyen d'une grille analytique à échelles descriptives (précisant les critères à considérer pour l'évaluation et donnant des indicateurs de la performance attendue aux différents échelons), par des évaluateurs formés et bénéficiant de sessions de standardisation.

Il est en effet difficile de savoir si ce qui est à l'œuvre dans la prise de décision d'un évaluateur correspond, effectivement, à ce qui est attendu par le concepteur du test. Quelle représentation réelle l'évaluateur a-t-il des différents critères au moment de la notation ? Dans quelle mesure note-t-il, réellement indépendamment, les différents critères ? Quelle est sa représentation des standards de niveau pour chacun des critères ? Comment relie-t-il ses observations à ces standards lors de la notation ? Qu'est-ce qui garantit qu'il ne donne pas une importance exagérée à un critère donné en modérant les notes qu'il délivre aux autres critères ou qu'il n'est pas sensible à un aspect particulier de la copie, comme la précision orthographique ou les tournures stylistiques utilisées, au détriment des autres observations ? Autant de questions qui invitent à la prudence dans l'exploitation des résultats notés par des évaluateurs même qualifiés.

2.4 L'évaluation par les machines, une mécanique algorithmique

Contrairement à l'évaluation humaine, l'évaluation par les machines est une mécanique algorithmique, qui ne laisse pas de place à une interprétation au-delà de ce qui a été prévu par l'algorithme. La machine ne comprend pas le texte et ne peut donc pas l'interpréter. L'algorithme va toujours rechercher les mêmes informations, qu'il combinera de la même manière pour chacune des productions afin d'aboutir à un résultat qui sera toujours le même pour une copie donnée, tant que le modèle n'aura pas été mis à jour ou entrainé avec de nouvelles données. Les informations à extraire peuvent être identifiées de sorte à être toutes pertinentes au regard du construit mesuré et le nombre d'informations différentes prises en considération peut dépasser celui des évaluateurs humains, qui ont un fonctionnement plus global.

Les systèmes de notation automatique suivent souvent une chaîne de traitement comportant plusieurs phases (Lim et al., 2021). La première phase, après récupération du texte produit par le candidat, est une phase de prétraitement, durant laquelle le système procède à une standardisation typographique de la copie, identifie les mots du texte qui ne figurent pas dans son dictionnaire et les remplace par les mots que le candidat a le plus probablement voulu écrire (étape de normalisation). L'algorithme de normalisation est un composant essentiel, surtout dans le cas de copies rédigées en langue étrangère et donc susceptibles de comporter un nombre élevé d'erreurs morphologiques. De sa qualité dépendra celle du traitement syntaxique subséquent, qui reconstruira d'autant mieux la logique de la phrase, qu'il reconnaitra les mots la constituant.

La seconde phase consiste, pour les systèmes reposant sur un apprentissage automatique, en l'extraction quantitative d'attributs ou « caractéristiques textuelles ». L'enjeu est d'extraire des informations quantitatives

pertinentes pour rendre compte de la qualité de la copie. Ce peut être en rapport avec le thème du sujet proposé au candidat (par exemple, le nombre de mots se situant dans le champ lexical du sujet), avec le type d'écrit attendu (variété des marqueurs de discours pertinents pour la tâche considérée), le développement thématique (variété des articulateurs logiques), les mécanismes utilisés pour maintenir la cohésion et la cohérence du texte ou encore des éléments liés à la syntaxe et à la correction lexicale et grammaticale.

La phase de notation proprement dite communique cet ensemble d'attributs quantifiés en entrées d'un modèle préalablement entrainé pour produire les résultats de la notation (score, niveau, degré de certitude. . .). La constitution du modèle est l'autre pierre angulaire du système. Elle consiste à trouver la combinaison, souvent non linéaire, entre les différentes variables quantitatives pour prédire au mieux le score ou le niveau de la copie, en se basant sur un historique large de copies évaluées par des humains, pour lesquelles le niveau de confiance dans l'évaluation est élevé. Différents types d'algorithmes peuvent être exploités à cette fin (régressions linéaires multiples, régressions logistiques ordinales, forêts d'arbres aléatoires, séparateurs à vaste marge, réseaux neuronaux. . .), selon qu'il s'agit de délivrer un score ou un niveau et selon la taille de l'échantillon d'apprentissage.

2.5 Les limites supposées de la notation automatique

Selon Cori (2020, p. 203), « les ordinateurs ne comprennent rien à nos langues, ce qui ne les empêche pas de nous rendre des services, de nous apporter une aide dans l'accomplissement de tâches diverses et variées relatives à nos productions langagières. ».

Est-il nécessaire de comprendre le contenu des textes produits par les candidats pour pouvoir les évaluer ? Dans le contexte du *Test d'évaluation de français*, où les tâches proposées sont des tâches communicatives, destinées à des lecteurs humains, la réponse semble affirmative. Pourtant de nos jours, des articles de presse sont produits automatiquement par un algorithme, alors qu'ils s'adressent à des lecteurs humains, dans une démarche communicative (Raynaud & Didier, 2018), et il est parfois difficile de s'en rendre compte.

Les humains interagissent également de plus en plus souvent avec des robots artificiels, qui les aident à formuler ou résoudre un problème. L'interaction s'effectue sans que le robot ne « comprenne » réellement l'échange langagier : il analyse le texte produit pour détecter l'intention du locuteur et les mots porteurs de sens, afin d'interroger une base de connaissances et restituer l'information la plus pertinente. Si la réalisation d'une tâche d'évaluation nécessite que le candidat inscrive son texte

dans un champ lexical en lien avec la thématique proposée et qu'il mette en œuvre des fonctions langagières prévisibles, la « machine » saura en rendre compte.

Cette limite peut toutefois s'avérer lorsqu'on ne s'intéresse pas simplement à la capacité du candidat à produire un discours organisé en mobilisant de manière pertinente une variété de ressources linguistiques en réponse à une tâche donnée, mais que le contenu précis du texte produit revêt une importance particulière. On peut alors penser qu'un expert saura porter un jugement plus approprié (Laurier & Diarra, 2008). De même, une copie rédigée avec une approche très originale, qui pourra être perçue positivement par un évaluateur humain, risque d'être évaluée de manière erronée par le système de correction si la base sur laquelle il a effectué son apprentissage comporte peu de textes de ce type. Il s'agit là d'une situation marginale, mais qui montre l'intérêt de conserver une évaluation humaine aux côtés de l'évaluation automatique. Cela permet de ne pas pénaliser à tort des productions originales. Une autre possibilité serait de doter le système de notation automatique de la capacité à identifier des copies s'éloignant fortement de sa base d'apprentissage pour les adresser à un correcteur humain (van Dalen et al., 2015). Enfin, si la maitrise des codes linguistiques était suffisante pour caractériser la compétence à rédiger un texte, un test à réponses fermées ciblant ces connaissances serait probablement plus efficace qu'une épreuve de rédaction. Pour Attali (2013) et Deane (2013), une machine ne peut pas réellement comprendre un écrit, ne le lit pas comme le lirait un humain et ne peut pas en interpréter le sens. En conséquence, les scores ne peuvent pas être interprétés de la même manière, l'ordinateur n'ayant pas accès au sens, contrairement à l'évaluateur humain.

D'autres critiques formulées à l'encontre des systèmes de notation automatique (Deane, 2013) concernent le comportement du candidat, qui peut différer s'il sait qu'il va être évalué par un système automatique de notation. Dans un contexte d'apprentissage, le fait que son écrit s'adresse à « une machine » et non à un humain peut entrainer un manque de motivation. Dans un contexte à fort enjeu tel que celui dans lequel nous nous situons, le candidat pourra chercher à tromper le système pour obtenir une surévaluation de sa compétence (McGee, 2006), notamment en tenant compte du fait que les systèmes de notation automatique sont réputés accorder plus d'importance que les humains à la longueur des textes produits (Pereleman, 2014, Kumar et al., 2017). Pour y remédier, les concepteurs de tests peuvent développer des modèles de détection de copies atypiques qui ne devraient pas être corrigées par les systèmes de notation automatique classique (Higgins et. al., 2004). Il faut toutefois être conscient que des stratégies similaires sont déjà présentes dans les dispositifs reposant sur la notation humaine où, par exemple, les candidats recourent parfois à des schémas rédactionnels et des structures

syntaxiques mémorisés qu'ils se contentent de contextualiser pour les adapter à la thématique de la tâche proposée. C'est d'ailleurs une des sources de divergence dans la notation humaine, les évaluateurs pouvant apprécier différemment la part d'apport personnel dans la copie reflétant la compétence réelle du candidat. Le comportement d'un individu en situation de test diffère en général de son comportement dans la vie réelle.

Enfin, selon Dean (2013), il ne faut pas se laisser abuser par les corrélations élevées entre les scores délivrés par des systèmes de notation automatique et des humains, du fait de la relation forte existant entre l'aisance à produire un texte et la capacité à mobiliser des ressources cognitives pour traiter des problèmes d'ordre conceptuel ou rhétorique. Pour Dean (2013), les systèmes de correction automatique fournissent peu de preuves directes de leur capacité à apprécier la force argumentative ou l'efficacité rhétorique d'un écrit, ce qui est problématique si ces éléments font partie du construit évalué. Les travaux de Kumar et al. (2017) accréditent cette thèse. Ces derniers ont montré qu'il était possible de concevoir un système d'évaluation automatique rudimentaire qui, en ne considérant que cinq attributs (reflétant la correction orthographique, la précision grammaticale, la similarité sémantique des phrases consécutives du texte, la connectivité et la diversité lexicale) et en s'appuyant sur une régression linéaire multiple pour prédire les scores, était capable de rivaliser avec les systèmes de notation automatique commerciaux. Il est dès lors légitime de s'interroger sur la prise en compte, par les systèmes de notation automatique, d'habiletés linguistiques de plus haut niveau. Le système SAGE, auquel ont contribué Zupanc et Bosnic (2017), qui se distingue des systèmes précédents par sa capacité à intégrer des attributs relatifs à la cohérence sémantique des textes et à la cohérence de l'énoncé, s'est d'ailleurs montré plus performant que ces systèmes commerciaux, ce qui renforce l'idée que cette dimension est prise en compte par les évaluateurs humains, mais pas suffisamment par les programmes.

Ainsi, les systèmes de notation automatique sont encore imparfaits, ont notamment des difficultés à accéder au sens et à intégrer les dimensions relevant de l'esprit critique dans l'évaluation (Deane, 2013). Mais des progrès sont réalisés continument et, à défaut d'envisager le remplacement des évaluateurs humains par des ordinateurs, les systèmes de notation automatique ont sans doute une place à trouver aux côtés des évaluateurs humains.

3. Évolutions récentes dans le domaine de la notation automatique

Dans un précédent ouvrage, Laurier et Diarra (2008) ont décrit plusieurs systèmes de notation automatique en langue anglaise (*Project Essay*

Grader – PEG, *Intelligent Essay Assessor* – IEA, *IntelliMetric*[3] et *e-rater*). Au-delà de son intérêt historique, cette présentation mettait en évidence la pluralité des approches proposées jusqu'alors. Ces quatre systèmes commerciaux ont participé en 2012, aux côtés de cinq autres, à une compétition baptisée *kaggle* (*Automated Student Assessment Prize* – ASAP) et destinée à favoriser l'émergence de solutions de notation automatique de productions écrites d'étudiants (Shermis, 2014). En dépit des réserves qui peuvent être émises sur les données utilisées (types d'écrits et notation humaine) et sur les résultats (Pereleman, 2013, 2014), l'initiative a ravivé l'intérêt pour la notation automatique. En effet, les organisateurs ont par la suite rendu publiques les données de la compétition et les performances des systèmes de notation automatique mis en concurrence. Cela a permis à d'autres acteurs, notamment universitaires, d'accéder à des échantillons conséquents de productions écrites au format numérique, étiquetées par leurs évaluations, matériau qui était jusqu'alors l'apanage des grands organismes de tests. Cela leur a également donné la possibilité de comparer les performances des systèmes qu'ils ont par la suite développés à l'état de l'art de 2012 (Zupanc & Bosnic, 2015). Cette émulation a favorisé l'innovation et l'émergence de systèmes plus performants (selon les résultats obtenus à partir des données du concours).

Notre intention dans cette partie n'est pas de faire une revue de l'existant. Nous renvoyons pour cela le lecteur à différentes revues en langue anglaise, sur lesquelles nous nous appuyons pour informer des tendances qui se dégagent (Zupanc & Bosnic, 2015 ; Hussein et al., 2019 ; Ke & Ng, 2019 ; Uto, 2021).

Un premier constat, à la lecture de Zupanc et Bosnic (2015), est l'apparition dans les années 2000 de systèmes de notation automatique dans au moins douze autres langues que l'anglais. Pour le français, l'article cite le programme Apex (système d'aide à la préparation d'examens) (Lemaire & Dessus, 1999), qui s'appuie sur l'analyse sémantique latente. La numérisation des tests, la dynamique actuelle autour de l'apprentissage automatique et la présence de nombreux articles et outils en libre accès, notamment pour le traitement automatique des langues et l'apprentissage automatique, ne peuvent qu'encourager cette tendance, même si l'accès à des productions notées reste limité.

Ce foisonnement fait également apparaitre de nouvelles méthodes pour l'extraction d'attributs, la classification des copies en niveaux ou la prédiction d'un score. Au-delà de la qualité des attributs extraits, la performance des systèmes de notation devient fortement dépendante des modèles et algorithmes utilisés pour la prédiction (régressions logistiques ordinales, forêts d'arbres aléatoires, machines à support de vecteurs,

[3] Désormais disponible en plusieurs langues.

ordonnancement, réseaux neuronaux. . .). La pluridisciplinarité devient alors une des clés du succès d'un projet. Une illustration détaillée d'un projet de ce type peut être trouvée dans Yannakoudakis (2013), l'université de Cambridge ayant fait une entrée remarquée dans le domaine de la notation automatique[4].

Le développement des plateformes de formation à distance encourage, par ailleurs, le développement d'outils de notation capables de produire un retour formatif (feed-back) sur la qualité linguistique des écrits produits, si possible en temps réel, notamment pour accompagner les étudiants ou candidats à la préparation d'examens (Gutierrez et al., 2012 ; Lemaire & Dessus, 1999 ; Rich et al., 2013). Les principaux systèmes commerciaux de notation automatique sont ainsi souvent adossés à une plateforme d'entrainement ou de formation : *Criterion* pour *e-rater*, *WriteToLearn* pour *IEA*, *MyAccess !* pour *Intellimetric* (Zupanc & Bosnic 2015), *Write&Improve et Speak&Improve* pour le système de notation automatique du test *Linguaskill*, conçu par *Cambridge English Assessment.* Ces plateformes de formation leur permettent de collecter massivement des productions de candidats qui alimentent les recherches menées en vue d'améliorer le système de notation automatique.

Selon Ke et Ng (2019), Taghipour et Ng (2016) sont les premiers à avoir proposé un système de notation automatique s'appuyant sur des réseaux neuronaux, suivi de près par Alikaniotis et al. (2016). Ils ont ainsi ouvert la voie à une multiplicité de systèmes neuronaux, qu'Uto (2021) classifie en quatre catégories selon qu'ils proposent une évaluation holistique ou multitraits[5] et selon qu'ils sont liés à un sujet spécifique ou applicables à des sujets différents. Alors qu'une grande partie du travail des concepteurs de systèmes de notation automatique consistait jusque-là à identifier des attributs pertinents et à en programmer manuellement l'extraction (Ke & Ng, 2019), les réseaux neuronaux holistiques prédisent directement le score des individus sur la base de la

---

[4] L'université de Cambridge a regroupé en 2013 une équipe multidisciplinaire autour de l'institut virtuel ALTA (*Automated Language Teaching and Assessment*) pour développer la recherche et proposer des solutions opérationnelles dans le domaine de l'évaluation et de l'enseignement automatiques. Depuis quelques années, leurs travaux portent surtout sur l'évaluation automatique de l'expression orale, comme l'indique la part importante de leurs publications sur ce thème, dont la plupart sont accessibles librement sur le site https://aclanthology.org/.

[5] Ici, un trait représente un aspect caractéristique de la compétence à évaluer. Une évaluation multitraits restitue des résultats pour un ensemble de traits, qui peuvent ensuite être combinés en un résultat global. L'évaluation au moyen de grilles analytiques peut être considérée comme une évaluation multitraits, chaque critère d'évaluation renvoyant à un trait particulier du construit, qui permet d'établir un profil de compétences. Encore faut-il pour cela que les différents critères soient évalués de manière indépendante.

séquence de mots du texte. Le rôle de l'ingénieur ou du chercheur est alors de concevoir une architecture qui permettra au réseau neuronal, par apprentissage supervisé (c'est-à-dire connaissant la note délivrée à chacune des copies de l'échantillon d'apprentissage), de déterminer de son propre chef les attributs qu'il est pertinent d'extraire pour prédire au mieux les résultats. Une telle architecture comporte généralement plusieurs niveaux (ou « couches ») et on parle par exemple de table de consultation, de couche de convolution, de couche récurrente, de réseau de mémoire à long terme, de couche à activation sigmoïdale (Taghipour & Ng, 2016). Les lecteurs non familiers avec la notion de réseaux neuronaux seront sans doute étonnés de voir que la conception de systèmes de notation automatique évolue ainsi vers la définition d'architectures technologiques susceptibles de capturer des attributs potentiellement intéressants pour la tâche d'évaluation. Il est clair que des efforts importants sont à fournir pour rendre de tels systèmes interprétables et explicables, sans quoi ils seront difficilement acceptés par la communauté éducative.

La revue proposée par Uto (2021) présente vingt-six systèmes de notation automatique à base de réseaux neuronaux. Vingt-deux systèmes proposent une évaluation holistique et quatre seulement proposent une évaluation multitraits. Les raisons pour lesquelles une majorité des systèmes développés concernent l'évaluation holistique sont notamment d'ordre pratique. Il existe davantage de corpus étiquetés avec des scores humains holistiques dans le domaine public (notamment depuis la compétition *kaggle*) que des corpus comportant des notes détaillées par trait. Or, les réseaux neuronaux ont besoin de grandes quantités de données d'entrainement pour être performants (Ke et Ng, 2019), ce qui est une limitation à leur essor, notamment dans des langues autres que l'anglais (ou sans doute le chinois). Cependant une évaluation holistique est souvent peu satisfaisante dans un contexte d'apprentissage, où l'étudiant a besoin de savoir quels aspects de son écrit il peut améliorer. Il est donc probable que, pour un temps encore, l'extraction artisanale d'attributs perdure dans la mise au point de systèmes opérationnels, probablement aux côtés de réseaux neuronaux plus à même de capturer efficacement certains attributs spécifiques. Une autre limite des systèmes à base de réseaux de neurones est que beaucoup d'entre eux sont entrainés sur des sujets (*prompts*) spécifiques (vingt sur vingt-six dans la recension d'Uto, 2021). Pour corriger des copies concernant un autre sujet, il faut alors procéder à un nouvel entrainement du modèle et donc, disposer de copies étiquetées par un score pour ce sujet. D'une part, c'est un processus couteux et, d'autre part, il induit que les poids attribués à chacun des attributs dans l'établissement du résultat (et donc le construit) varieront d'un sujet à l'autre (Attali, 2013) alors qu'un des atouts de la notation automatique était jusqu'alors justement sa constance dans l'importance accordée

à chacun des aspects de la compétence à évaluer, contrairement à l'évaluation humaine.

4. Le système de notation automatique du Français des affaires

Le *Français des affaires*, établissement de la Chambre de commerce et d'industrie de Paris Ile-de-France, conçoit et diffuse un *Test d'évaluation de français – TEF*, utilisé notamment dans des démarches d'accès au territoire, de résidence ou d'acquisition de la nationalité dans plusieurs pays francophones. *Le français des affaires* a amorcé en 2019 un projet de conception d'un système de notation automatique, dans une perspective exploratoire. Il s'interroge désormais sur les usages possibles d'un tel système de notation automatique dans le cadre de son activité d'évaluation en langue française.

4.1 L'épreuve d'expression écrite du TEF

L'épreuve d'expression écrite du TEF a commencé à être proposée sur ordinateur en janvier 2018 et son utilisation a été généralisée au cours de l'année 2020. Cette épreuve comporte, dans son format complet, deux tâches distinctes : la première tâche évalue la capacité à transmettre des informations, via un récit, alors que la seconde évalue la capacité à argumenter. Pour certaines versions du test, seule la seconde tâche est proposée aux candidats. Nous nous situons donc dans le contexte d'un test préexistant, dont les tâches de l'épreuve d'expression écrite n'ont pas été sélectionnées en vue d'une notation automatique.

Chaque production écrite est systématiquement évaluée par deux évaluateurs, de manière indépendante, lesquels utilisent pour cela une grille d'évaluation analytique à échelles descriptives. La numérisation de l'épreuve a ainsi permis la constitution d'un corpus de productions écrites de candidats pour lesquelles *Le français des affaires* dispose des deux notes individuelles délivrées par les évaluateurs pour chacun des critères, ainsi que du score final délivré à la performance (qui peut résulter d'un arbitrage lorsque les deux évaluations initiales diffèrent fortement).

La grille d'évaluation comporte deux critères pragmatiques, qui concernent la capacité à réaliser chacune des deux tâches, ainsi que trois critères linguistiques, qui évaluent la syntaxe, le lexique et la cohérence/cohésion. L'évaluation de chacun des critères est transcrite en une note allant de 0 à 10 (soit onze modalités). Les notes aux critères sont combinées selon un système fixe de pondérations de façon à exprimer un score à l'épreuve. L'échelle des scores et subdivisée en sept niveaux

principaux : un niveau <A1 et les niveaux A1 à C2 du Cadre européen commun de référence (désormais CECR).

Les résultats des candidats ayant passé l'épreuve complète d'expression écrite du TEF au cours de l'année 2021 (soit environ 35.000 copies) sont distribués de façon asymétrique : 8 % des candidats ont un niveau A1 ou A2, 59 % un niveau B1 ou B2 et 33 % un niveau C1 ou C2. Les notes attribuées par les évaluateurs aux candidats à chacun des critères sont fortement corrélées entre elles. La corrélation de Spearman entre les séries notes de chacun des deux critères pragmatiques est de 0,88. Les corrélations entre notes impliquant un critère pragmatique et un critère linguistique varient entre 0,88 et 0,93. Les corrélations entre les notes correspondant aux critères linguistiques sont les plus élevées (0,93 à 0,95). Une analyse en composantes principales des cinq séries de notes montre que le premier facteur explique à lui seul 92,1 % de la variance. Ces corrélations élevées confèrent à chacun des critères un pouvoir prédictif fort concernant le score final de la copie. Chaque critère renvoie lui-même à différents mécanismes langagiers (Il existe, par exemple, différentes façons d'assurer la cohérence et la cohésion d'un texte.), mais là encore on peut supposer que si une note était attribuée à chacun des mécanismes mobilisables, un facteur commun prépondérant se dégagerait de l'ensemble des notes. Il semble donc raisonnable de penser que, en comptabilisant tout un ensemble de caractéristiques présentes dans la copie, il est possible de prédire efficacement le score qui sera délivré par un jury d'évaluateurs humains.

Les scores délivrés par les deux évaluateurs d'une même copie ne sont toutefois pas toujours identiques et le niveau associé peut différer. La corrélation de Pearson entre les scores délivrés par les évaluateurs aux copies de l'année 2021 était de 0,738 avant arbitrage. L'accord exact des classements (même niveau CECR délivré par les évaluateurs) était de 45 % et l'accord adjacent (même niveau ou un niveau d'écart) de 92 %. Lorsqu'on considère les notes attribuées à chacun des critères, l'accord exact entre les deux évaluateurs (même note) varie entre 25 % et 27 % selon les critères et l'accord adjacent (notes différant au plus de 1 point) varie entre 62 % et 65 %. Pour la mise au point de son outil de notation automatique, *Le français des affaires* a fait le choix de ne considérer que les copies pour lesquelles les deux évaluateurs attribuaient un niveau identique.

4.2 L'extraction d'information à partir des textes

Le Test d'évaluation de français s'adressant majoritairement à un public dont le français n'est pas la langue maternelle, les textes produits peuvent comporter un nombre important d'erreurs morphologiques. De telles erreurs ont un impact important sur la qualité de l'annotation des

copies par l'outil d'analyse syntaxique (Pour l'analyse syntaxique, nous avons utilisé la librairie *udpipe* de R, qui recourt aux modèles du projet Dépendances Universelles.) (Nivre et al., 2018). L'étape de normalisation des textes est donc cruciale, notamment pour les copies de niveau faible et le simple usage d'un correcteur automatique comme *Hunspell* s'avère insuffisant. *Le français des affaires* s'est appuyé sur les travaux de Bergé (2007) pour encoder phonétiquement un dictionnaire ainsi que les textes produits. Cela permet, en utilisant une distance entre les encodages phonétiques (comme la distance de Levenshtein), de suggérer, à partir du dictionnaire, des mots proches phonétiquement des mots erronés saisis par les candidats. Un modèle d'apprentissage automatique a été développé pour choisir, parmi ces propositions et celles de *Hunspell*, la plus pertinente en s'appuyant sur un ensemble de variables extraites du mot saisi par le candidat. La fréquence des erreurs morphologiques est un indicateur de la compétence orthographique et, le nombre total de mots différents, un premier indicateur de la richesse lexicale du candidat.

L'analyse syntaxique des textes normalisés permet de récupérer un ensemble d'informations grammaticales concernant les mots utilisés, comme la catégorie de mot (ou partie du discours – *part of speech*), le genre, le nombre, le temps verbal (s'il s'agit d'un verbe) et les relations de dépendance entre les constituants de la phrase. De telles caractéristiques sont porteuses d'information tant isolément que lorsqu'elles sont mises en relation. L'utilisation de conjonctions de subordination sera plus fréquente dans les modèles de niveau avancés et il est possible de vérifier le respect de règles d'écriture au sein de la copie, comme l'accord en genre et en nombre. Cet étiquetage va permettre le calcul de variables de différentes catégories. Certaines variables, comme la fréquence des erreurs morphologiques, sont autosuffisantes. D'autres variables se réfèrent à des règles d'écritures fixes, comme l'accord en genre et en nombre. D'autres variables se réfèrent à des listes de références préétablies, comme la diversité des temps verbaux existants, ou à des listes plus ouvertes constituées manuellement, comme les groupes de mots marquant l'expression d'une opinion, ou encore des listes de références externes comme l'outil FLELex (François et al., 2014). La liste FLELex a été établie à partir d'un corpus de textes issus de manuels d'apprentissage du français langue étrangère à différents niveaux du CECR. Elle fournit une fréquence d'apparition dans ces manuels d'un vaste ensemble de mots à chacun des niveaux CECR. A cela s'ajoutent des variables qui s'obtiennent en mettant en relation différentes parties du texte, comme la similarité entre les mots de phrases successives pour rendre compte de la cohérence du texte, ou entre les mots du texte et les mots du sujet pour vérifier que le texte produit s'inscrit dans le champ lexical du sujet. D'autres outils

non encore exploités, comme ALSI (Loignon, 2021) ou FABRA (Wilkens et al., 2022), devraient permettre à l'avenir d'ajouter de nouvelles variables.

Enfin il est possible de constituer d'autres types de variables en réservant une partie du corpus de copies à la création de modèles de langue à base de n-grammes (Jurafsky & Martin, 2020) pour chacun des niveaux du CECR. Les modèles à base de n-grammes (où les textes sont transformés en séquences de n mots ou de n caractères contigus) sont fréquemment utilisés pour l'identification de l'auteur d'un texte (Kešelj et al., 2003). Il s'agit de constituer un modèle par auteur en stockant une table comportant chacun des n-grammes présents dans les textes de référence produits par cet auteur (échantillon d'apprentissage) avec sa fréquence d'apparition, ce qui permet de calculer, pour tout nouveau texte, après décomposition en n-grammes, la probabilité qu'il ait été généré par chacun des auteurs modélisés. Dans notre cas, les modèles n-grammes permettent d'identifier la probabilité que l'auteur de la copie soit d'un niveau CECR donné. Différents modèles n-grammes sont utilisés, lesquels portent soit directement sur les mots soit, après analyse syntaxique automatique, sur les catégories de mots. La fréquence d'apparition des mots complexes dans un texte sera généralement plus élevée dans les modèles de niveau avancé, de même que les conjonctions de subordination. Les modèles varient également selon la taille des n-grammes, des unigrammes rendant compte de fréquence d'apparition des mots (ou catégories de mots) indépendamment du contexte, alors que des bigrammes comptabilisent les fréquences d'apparition de paires de mots et renseignent sur l'utilisation de collocations ainsi que sur l'organisation de la phrase.

4.3 La notation et ses limites

Une fois les variables recueillies, elles peuvent servir à entrainer un modèle de prédiction du score ou du niveau de la copie à partir d'un échantillon d'apprentissage. Afin que le modèle fasse sa prédiction sur la base du texte uniquement et que cette dernière ne soit pas influencée par la distribution de la compétence dans la population, il est important de veiller à ce que l'échantillon d'apprentissage comporte un nombre comparable de copies pour chacun des niveaux. Cette condition étant difficilement réalisable au vu de l'asymétrie des niveaux dans l'échantillon de départ, qui comporte peu de copies de niveau A1 ou A2, *Le français des affaires* a intégré à l'échantillon d'apprentissage des copies de niveau A1 et A2 provenant des versions du test ne proposant que la seconde tâche (argumentative), en attribuant aux variables spécifiques de la première tâche (non traitée par les candidats) une valeur identique à celles obtenues pour les variables spécifiques de la seconde tâche. Une

fois le modèle entrainé, il peut être appliqué à de nouvelles copies pour en prédire le résultat.

Nous avons utilisé le reste du corpus comme échantillon de test pour évaluer les performances des différents modèles testés. Les modèles basés sur des machines à supports de vecteurs et les forêts d'arbres aléatoires sont ceux qui ont montré la meilleure capacité de prédiction. Ils ont permis de retrouver, sur l'échantillon de test, le niveau CECR délivré par les évaluateurs dans 76 % des cas et l'écart n'a été de 2 niveaux CECR ou plus que dans moins de 1 % des cas.

Ces résultats sont encourageants, mais ne peuvent pas être généralisés à l'ensemble des copies. En effet, les échantillons retenus tant pour l'apprentissage que pour le test sont constitués de copies pour lesquelles les deux évaluations humaines initiales étaient de niveau identique. Or il se peut que les copies, pour lesquelles les évaluations humaines sont en désaccord, correspondent plus fréquemment à des copies atypiques que le système de notation automatique aura plus de difficultés à évaluer précisément. De surcroit, l'échantillon de test comportait davantage de copies de niveau B1 et B2, alors que la prédiction est moins bonne pour les niveaux extrêmes.

Par ailleurs, les variables auxquelles les modèles de classification accordent le plus d'importance sont des variables lexicales, à savoir les probabilités des modèles unigrammes portant sur les lemmes pour les niveaux <A1 à B1, le taux de mots du texte qui ont été reconnus (c'est-à-dire qui figuraient dans le dictionnaire), le nombre de mots différents présents dans la copie (les mots non identifiés ayant tous été remplacés par un même code). Viennent ensuite majoritairement les variables se rapportant aux trigrammes sur les catégories de mots, qui se rapportent à la syntaxe. Les attributs auxquels le modèle accorde le moins d'importance sont ceux qui sont en rapport avec les critères pragmatiques et la cohérence du texte. Cela questionne la validité du construit évalué, les aspects pragmatiques étant sous-considérés alors qu'il s'agit des critères de notation qui ont un poids prépondérant dans l'élaboration des scores à partir des grilles de notation.

C'est pourquoi *Le français des affaires* oriente désormais ses travaux vers une prédiction de la note délivrée à chacun des critères, ce qui permettra de combiner les notes prédites en un score en utilisant le même système de pondération que pour la notation humaine. Cela nécessite de réorganiser les corpus pour n'exploiter, pour chaque critère, que les copies dont les notes délivrées par les deux évaluateurs ne diffèrent de pas plus de 1 point, la somme des deux notes variant entre 0 et 20. Il faudra également procéder à une phase de sélection des variables pertinentes à considérer pour chacun des critères, en limitant le nombre de variables communes à plusieurs d'entre eux. Cette approche par critère permettra

d'identifier les aspects de la compétence que le système de notation automatique est le plus capable d'évaluer (à priori les critères relatifs au lexique et à la syntaxe) et ceux pour lesquels les variables actuellement recueillies sont insuffisantes pour porter un jugement proche du jugement humain (probablement les critères pragmatiques).

Cette estimation des forces et faiblesses pourra orienter l'usage qui sera fait de l'outil. Pour les raisons évoquées plus haut, il semble illusoire de vouloir remplacer, dans le contexte d'un test à forts enjeux et à visée communicative, les évaluateurs humains par un système de notation automatique. Mais un tel système doit pouvoir trouver sa place aux côtés des évaluateurs humains (Davis & Papageorgiou, 2021). S'il s'avère particulièrement fidèle pour l'évaluation de certains critères, il pourrait être utilisé pour préremplir les grilles d'évaluation pour ces derniers. Les évaluateurs seraient alors invités à se concentrer sur l'évaluation des autres aspects de la langue et à ne modifier les choix du système de notation automatique que lorsque leur perception de la performance pour ces critères est fortement différente. Un autre usage possible serait d'exploiter les résultats pour rendre compte de la sévérité relative avec laquelle les évaluateurs humains notent certains critères, en bénéficiant d'une comparaison directe avec le résultat attribué par le système de notation automatique.

6. Conclusion

Le développement de systèmes de notation automatique s'est beaucoup focalisé sur sa capacité à reproduire des scores humains, dans le but de les substituer, à terme, à l'évaluation humaine. Mais pour atteindre un tel objectif, il faudrait montrer que les deux systèmes de notation mesurent le même construit. Or, on ne comprend pas finement les modes de fonctionnement de l'évaluateur humain. Deux évaluateurs humains peuvent aboutir à un même score en considérant différents traits ou en appréciant différemment un même ensemble de traits. Les efforts de formation visent justement à harmoniser les pratiques pour s'assurer que le construit est suffisamment bien respecté. La similarité entre scores n'est donc pas une garantie suffisante : il faut comprendre comment la machine aboutit à une note, en s'appuyant sur quels attributs, avec quelles pondérations et s'assurer que ces attributs offrent une couverture suffisante du construit mesuré. Or aujourd'hui la machine n'est pas réellement en mesure de comprendre et d'interpréter un texte. Pas plus sans doute qu'elle n'est en mesure d'apprécier l'originalité d'un écrit ni la pensée critique du rédacteur. Il semble donc préférable de s'orienter vers une répartition des rôles entre l'homme et la machine (Attali, 2013), du

moins lorsque le test prétend évaluer d'autres aspects que le simple respect de la mécanique langagière.

Compte tenu de la charge cognitive de l'évaluation, l'humain ne peut raisonnablement prendre en considération qu'un nombre limité d'aspects dans son évaluation. De ce fait, les critères qui sont proposés dans les grilles d'évaluation sont relativement globaux et imprécis, ce qui fait qu'on ne sait pas réellement sur quels attributs l'évaluateur humain s'appuie pour noter un critère et en quelles proportions. En réduisant son champ d'intervention aux aspects du construit que le système de notation automatique ne sait pas traiter efficacement, il serait en mesure d'apprécier différents aspects de haut niveau de la compétence à écrire au lieu de les amalgamer. La combinaison homme/machine fait donc la promesse d'une évaluation plus riche, avec une meilleure représentation du construit d'expression écrite. Une profonde réflexion sur ce construit en lien avec la répartition des rôles entre l'homme et la machine nous semble primordiale.

Dissocier ainsi les rôles, c'est aussi opter pour une approche par traits, qui présente plusieurs autres avantages. Cela permet notamment d'expliquer plus facilement à la communauté éducative ce qu'évalue le système de notation automatique et de contrôler l'importance prise par chacun des traits dans l'évaluation finale. Le niveau inférieur, qui correspond à la sélection et la pondération des attributs pour aboutir à la notation du trait, pourra quant à lui être plus complexe par le nombre d'attributs considérés, l'algorithme de notation utilisé et l'importance attribuée à chacun des attributs, mais devra rester interprétable et explicable. Cette capacité d'interprétation ouvre aussi la voie à un retour formatif plus riche à destination du rédacteur du texte. Intégrer cette dimension formative dans le développement d'un système de notation automatique peut d'ailleurs être un garde-fou qui aidera à privilégier la compréhension des mécanismes de notation à l'efficacité brute de boites noires.

En conclusion, plutôt que d'envisager la notation automatique sous son aspect purement économique ou comme solution aux failles de l'évaluation humaine et d'opposer ainsi l'homme à la machine, il convient sans doute de la penser comme une aide à l'évaluation et à l'apprentissage. Elle permettrait aux évaluateurs humains de se concentrer sur les aspects de l'écrit, pour lesquels leur appréciation a la plus forte valeur ajoutée, c'est-à-dire ceux qui mobilisent leur capacité d'inférence et leur expérience de lecteur et sont sans doute les plus stimulants intellectuellement. Ainsi toutes les parties prenantes pourraient tirer profit d'un tel système, qui devrait toutefois faire l'objet d'une étude de validation approfondie avant d'être utilisé dans des situations à enjeux élevés.

Références

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. Dans K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the AssociationVol. 1Long Papers* (pp. 715–725). Association for Computational Linguistics. https://doi.org/10.18653/v1/p16-1068

Ang-Aw, H., T., & Chuen Meng Goh, C. (2011). Understanding discrepancies in rater judgement on national-level oral examination tasks. *RELC journal, 42*(1), 31–51. https://doi.org/10.1177/0033688210390226 *for Computational Linguistics :* .

Attali, Y. (2013). Validity and reliability of automated essay scoring. Dans M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (pp. 181–198). Routledge.

Barkaoui, K. (2010). Variability in ESL essay rating processes: the role of the rating scale and rater experience ESL essay rating processes. *Language Assessment Quarterly, 7*(1), 54–74. https://doi.org/10.1080/15434300903464418

Bejar, I. I. (2012). Rater cognition: implications for validity. *Educational Measurement Issues and Practice, 31*(3), 2–9. https://doi.org/10.1111/j.1745-3992.2012.00238.x

Bennett, R. E., & Bejar, I. I. (1997). Validity and automated scoring: It's not only the scoring. *ETS Research Report Series, 1997*, i-30. https://doi.org/10.1002/j.2333-8504.1997.tb01734.x

Bergé, E. (2007). *Phonetic for the french language via "SOUNDEX FR"-algorithm*. https://github.com/voku/phonetic-algorithms/blob/master/src/voku/helper/PhoneticFrench.php

Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on english-for-academic-purposes speaking tasks. *ETS Research Report Series, 2005*, i-157. https://doi.org/10.1002/j.2333-8504.2005.tb01982.x

Casanova, D. (2021). *Optimiser l'arbitrage grâce à la notation automatique ?* [Communication orale]. *ALTE 1st International Digital Symposium*. https://www.alte.org/DigitalSymposium2021-videos

Cori, M. (2020). *Le traitement automatique des langues en question. Des machines qui comprennent le français ?* Cassini.

Davis, L., & Papageorgiou, S. (2021). Complementary strengths ? Evaluation of a hybrid human- machine scoring approach for a test of oral academic english. *Assessment in Education: Principles, Policy & Practice, 28*(4), 437–455. https://doi.org/10.1080/0969594X.2021.1979466

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002

Eckes T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155–185. https://doi.org/10.1177/0265532207086780

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005) Individual feedback to enhance rater training : Does it work? *Language Assessment Quarterly, 2*(3), 175–196. https://doi.org/10.1207/s15434311laq0203_1

François, T., Gala, N., Watrin, P., & Fairon, C. (2014). FLELex : a graded lexical resource for French foreign learners. Dans N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (Eds.), *LREC'14 Proceedings of the 9th International Conference on Language Resources and Evaluation* (pp. 3766–3773) European Language Resources Association (ELRA).

Gauthier, G., St-Onge, C., & Dory, V. (2016). Synthèse et conceptualisation des processus cognitifs du jugement évaluatif de l'enseignant clinicien. *Pédagogie Médicale, 17*(4), 261–267. https://doi.org/10.1051/pmed/2017014

Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014). Seeing the 'black box' differently: assessor cognition from three research perspectives. *Medical Education, 48*, 1055–1068. https://doi.org/10.1111/medu.12546

Gutierrez, F., Dou, D., Fickas, S., & Griffiths, G. (2012). Providing grades and feedback for student summaries by ontology-based information extraction. Dans X. Chen, G; Lebanon, H. Wang & M. J. Zaki (Eds.), *CIKM'12 Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 1722–1726). Association for Computing Machinery. https://doi.org/10.1145/2396761.2398505

Han, Q. (2016). Rater cognition in L2 speaking assessment: a review of the Literature. *Studies in Applied Linguistics & TESO: Vol. 16*(1), 1–24. https://doi.org/10.7916/salt.v16i1.1261

Higgins, D., Burstein, J., Marcu, D., & Gentile, C. (2004). Evaluating multiple aspects of coherence in student essays. Dans K. Toutanova (Ed.), *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL* (pp. 185–192). Association for Computational Linguistics. https://aclanthology.org/N04-1024=

Hussein M.A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: a literature review. *PeerJ Computer Science* 5 : e208. https://doi.org/10.7717/peerj-cs.208

Ince, E. (2022). *Le jugement des examinateurs dans le cas de l'épreuve d'expression orale du TEF* [Thèse de doctorat, Université de Montréal]. Papyrus. https://doi.org/1866/27537

Isaacs, T., & Tromfimovich, P. (2010). Falling on sensitive ears ? The iof musical ability on extreme raters' judgments of L2 pronunciation. *TESOL Quarterly, 44*(2), 375–386. https://onlinelibrary.wiley.com/doi/abs/10.5054/tq.2010.222214

Jurafsky, D., & Martin, J. (2020). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition* (3rd Edition draft). https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf

Ke, Z., & Ng, N. (2019). Automated essay scoring: a survey of the state of the Art. Dans *Proceedings of the 28th International Joint Conference on Artificial Intelligence* (IJCAI-19) (pp. 6300–6308). https://doi.org/10.24963/ijcai.2019/879

Kešelj, V., Peng, F., Cercone, N., & Thomas, C. (2003). N-gram-based author profiles for authorship attribution. Dans *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING, 3*, (pp. 255–264). https://www.researchgate.net/publication/2872982_N-Gram-Based_Author_Profiles_For_Authorship_Attribution

Kumar, V., Fraser, S., N., & Boulanger, D. (2017). Discovering the predictive power of five baseline writing competences. *Journal of Writing Analytics, 1*, 176–226. https://doi.org/10.37514/JWA-J.2017.1.1.08

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72–107. https://doi.org/10.1037/0033-2909.87.1.72

Laurier, M., D., & Diarra, L. (2008). L'apport des technologies dans l'évaluation de la compétence à écrire. Dans J.-G. Blais (Ed.), *Évaluation des apprentissages et technologies de l'information et de la communication. Enjeux, applications et modèles de mesure* (pp. 77–104). Presses de l'Université Laval.

Leclercq, D., Nicaise, J., & Demeuse, M. (2004). Docimologie critique : des difficultés de noter des copies et d'attribuer des notes aux élèves. Dans M. Demeuse (Ed.), *Introduction aux théories et aux méthodes de la mesure en sciences psychologiques et en sciences de l'éducation* (pp. 273–292). Les éditions de l'Université de Liège. https://hal.science/hal-00844778

Lemaire, B., & Dessus, Ph. (1999). APex, un système d'aide à la préparation d'examens. *Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation*, *6*(2), 409–415. https://doi.org/10.3406/stice.1999.1637

Lim, C. T., Bong, C. H., Wong, W. S., & Lee, N. K. (2021). A comprehensive review of automated essay scoring (AES) research and development. *Pertanika Journal of Science and Technology, 29*(3), 1875–1899. https://doi.org/10.47836/pjst.29.3.27

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. MESA Press.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias : implications for training, *Language Testing, 12*, 54–71. https://doi.org/10.1177/026553229501200104

Lumley, T. (2002). Assessment criteria in a large-scale writing test : what do they really mean to the raters ? *Language Testing, 19*(3). https://doi.org/10.1191/0265532202lt230oa

Loignon, G. (2021). *Une approche computationnelle de la complexité linguistique par le traitement automatique du langage naturel et l'oculométrie*. [Thèse de doctorat, Université de Montréal]. Papyrus. https://doi.org/1866/26189

Macrae, C. N., & Bodenhausen, G. V. (2001). Social cognition : Categorical person perception. *British Journal of Psychology, 92*(1), 239–255. https://doi.org/10.1348/000712601162059

Martin, J. (2002). Aux origines de la « science des examens » (1920–1940). *Histoire de l'éducation, 94*, 177–199. https://doi.org/10.4000/histoire-education.817

McGee, T. (2006). Taking a spin on the intelligent essay assessor. Dans P. Freitag Ericsson et R. H. Haswell (Eds.), *Machine Scoring of Student Essays : Truth and Consequences?* (pp. 79–92). Utah State University Press.

Orr, M. (2002). The FCE Speaking test: using rater reports to help interpret test scores. *System 30*(2), 143–154. https://doi.org/10.1016/S0346-251X(02)00002-7

Pereleman, L. (2013). Critique of Mark D. Shermis & Ben Hammer, "Contrasting state-of-the-art automated scoring of essays: Analysis". *Journal of Writing Assessment, 6*(1). https://escholarship.org/uc/item/7qh108bw

Pereleman, L. (2014). When "the state of the art" is counting words. *Assessing Writing, 21*, 104–111. https://doi.org/10.1016/j.asw.2014.05.001

Raynaud, P., & Didier, I. (2018). Production automatique de textes : l'IA au service des journalistes. *La revue des médias*. https://larevuedesmedias.ina.fr/production-automatique-de-textes-lia-au-service-des-journalistes

Rich, C. S., Schneider, M. C., & D'Brot, J. M. (2013). Applications of automated essay evaluation in West Virginia. Dans M. D. Shermis et J. Burstein (Eds.), *Handbook of Automated Essay Evaluation : Current Applications and New Directions* (pp. 99–123). Routledge.

Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences.* Éditions du renouveau pédagogique.

Shermis, M. D. (2014). State-of-the-art automated essay scoring : competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*(1), 53–76. https://doi.org/10.1016/j.asw.2013.04.001

Scott, S. G., & Bruce, R. A. (1995). Decision-making style : the development and assessment of a new measure. *Educational and Psychological Measurement, 55*(5), 818–831. https://doi.org/10.1177/0013164495055005017

Suchaut, B. (2008). La loterie des notes au bac : un réexamen de l'arbitraire de la notation des élèves. *Les Documents de Travail de l'IREDU*. https://shs.hal.science/halshs-00260958v2

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. Dans J. Su, K. Duh & X. Carreras (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp.1882–1891). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1193

Uto, M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika, 48,* 459–484. https://doi.org/10.1007/s41237-021-00142-y

van Dalen, R. C., Knill, K., & Gales, M. (2015). Automatically grading learners' english using a Gaussian process. *Workshop on Speech and Language Technology in Education*. ISCA. https://www.repository.cam.ac.uk/handle/1810/249186

van Merriënboer, J. J. G., & Sweller, J. (2010). Cognitive load theory in health professional education : design, principles and strategies. *Med Educ, 44*(1), 85–93. https://doi.org/10.1111/j.1365-2923.2009.03498.x

Weigle S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263–287. https://doi.org/10.1177/026553229801500205

Weigle, S. C. (2002). *Assessing Writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Wilkens, R., Alfter, D., Wang, X., Pintard, P., Tack, A., Yancey, K., & François, T. (2022). FABRA : French aggregator-based readability assessment toolkit. Dans N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk & S. Piperidis (Eds.), *LREC 2022 Proceedings of the thirteenth international conference on language resources and evaluation*. European Language Resources Association https://aclanthology.org/2022.lrec-1.130

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4*, 83–106. https://doi.org/10.1016/S1075-2935(97)80006-2

Wolfe, E. (2005). Uncovering rater's cognitive processing and focus using think-Aloud protocols. *Journal of Writing Assessment, 2*(1), 37–56. Hampton Press Inc. https://escholarship.org/uc/item/83b618ww

Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication, 15*, 465–492. https://doi.org/10.1177/0741088398015004002

Yannakoudakis, H. (2013). *Automated assessment of English-learner writing*. University of Cambridge, Computer Laboratory, TR-842. https://doi.org/10.48456/tr-842

Nivre, J., Abrams, M., Agic, Z., Ahrenberg, L. et al. (2018). *Universal dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (UFAL)*, Faculty of Mathematics and Physics, Charles University. https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2895

Zupanc, K., & Bosnic, Z. (2015). Advances in the field of automated essay evaluation. *Informatica, 39*, 383–395. https://www.proquest.com/docview/1783257662

Zupanc, K., & Bosnic, Z. (2017). Automated essay evaluation with semantic analysis. *Knowledge-Based Systems, 120,* 118–132. http://dx.doi.org/10.1016/j.knosys.2017.01.006