Corpus d'apprenants et recherche linguistique

Casani Emanuele

CORPUS D'APPRENANTS ET RECHERCHE LINGUISTIQUE. UNE ÉTUDE SUR L'ITALIEN.

Emanuele CASANI[a]
emanuele.casani@unive.it
[a]*Dipartimento di Studi Linguistici e Culturali Comparati, Università Ca' Foscari Venezia Ca' Bembo, 30123 Venezia, Italia*

1. Introduction

Dès le lendemain de la publication du Cadre Européen Commun de Référence pour les Langues (CECRL), le Conseil de l'Europe a encouragé l'utilisation de corpus de langue authentique pour le développement d'outils d'aide au Cadre décrivant les compétences linguistiques et communicatives dans les différentes langues européennes (Council of Europe, 2005 ; Hulstijn, 2007). Toutefois, les réponses à cet appel ont été peu nombreuses et limitées aux langues européennes largement utilisées, à l'exception du *Norsk Profil* (Carlsen, 2013) norvégien. Pour la langue italienne, ces réponses se sont matérialisées par la publication du *Profilo della Lingua Italiana* (Spinelli & Parizzi, 2010) et, plus récemment, du corpus *MERLIN*[1], qui est utilisé dans l'étude décrite ici (Casani, 2020). Cette étude illustre une application possible de la méthodologie d'analyse d'erreurs à un corpus d'apprenants (CA) italien. Le but est de décrire la compétence morphosyntaxique d'un échantillon représentatif d'apprenants adultes d'italien Langue Étrangère (LE) pour les niveaux élémentaires et intermédiaires du CECRL.

1.1. CA et recherche linguistique

Dès le début des années 1990, la linguistique acquisitionnelle et la linguistique de corpus ont fait converger leurs études dans la recherche sur les CA (Leech, 1998 ; Tan, 2005), vastes et systématiques recueils électroniques de productions d'apprenants de langue seconde (L2) et LE (Granger, 2002 ; Leech, 1998).

De la linguistique de corpus, la recherche sur les corpus d'apprenants a adopté l'approche quantitative des mécanismes d'analyse, tels que l'annotation. De la linguistique acquisitionnelle, elle a hérité les méthodes d'analyse contrastive et d'analyse d'erreurs qui, mises en œuvre avec les possibilités de la recherche informatisée sur les corpus, ont généré un outil puissant pour l'étude quantitative et qualitative de l'apprentissage des langues. Travailler sur un corpus d'apprenants permet généralement de mettre en évidence les domaines dans lesquels les apprenants manifestent une sous-utilisation ou un abus des fonctions langagières dans une langue par rapport aux locuteurs natifs grâce à la méthodologie contrastive, et d'acquérir des connaissances sur les erreurs commises par les étudiants d'une certaine langue grâce à la méthodologie d'analyse d'erreur (Leech, 1998, Granger, 2002). Dans la première approche, il est possible de travailler directement sur les corpus ; alors que dans la seconde, un type spécifique de codage et d'annotation de l'erreur est requis. Pour cette raison, les CA sont particulièrement utiles quand ils sont étiquetés, c'est-à-dire quand toutes les erreurs du corpus ont été annotées avec un système normalisé de classification des erreurs, qui enrichit les données d'informations linguistiques. Les CA diffèrent des collections de données couramment utilisées par les chercheurs en linguistique acquisitionnelle et en didactique des langues pour deux raisons principales: ils sont informatisés et peuvent donc être analysés à l'aide d'un large éventail d'outils logiciels linguistiques assurant un traitement rapide et efficace des données; ils sont très vastes et fournissent donc une base beaucoup plus fiable pour la

[1] https://merlin-platform.eu.

description de la langue des apprenants que les collections utilisées précédemment (Granger, 2003). Si un corpus de 200 000 mots peut être considéré déjà très vaste pour la linguistique acquisitionnelle, des corpus de plusieurs centaines de millions de mots sont désormais devenus la norme (Granger, 2003).

1.2. Classifications des CA

La conception d'un CA peut se dérouler sur la base de critères linguistiques (mode, médium, genre ou sujet des textes); critères liés aux tâches (longitudinales/transversales, spontanées/préparées), critères relatifs aux apprenants (par exemple L2 ou LE, âge, sexe, langue maternelle, expérience à l'étranger etc.) (Tono, 2003). Certains critères de classification courants sont :

- la langue cible: la plupart des CA se réfèrent à l'anglais L2/LE. Le nombre de CA pour les autres langues est encore limité mais en augmentation (voir figure 1) ;
- le medium: les CA peuvent être composés de textes écrits ou parlés. Ces derniers sont beaucoup plus difficiles à remplir et donc moins communs ;
- langue maternelle : les données peuvent provenir d'apprenants de langue maternelle identique ou différente ;
- compétence dans la langue cible: certains CA rassemblent des textes d'apprenants du même niveau, d'autres incluent des textes de différents niveaux ;
- critères d'annotation: de nombreux CA ne contiennent que des données brutes, éventuellement corrigées, sans annotation linguistique; certains incluent l'annotation des parties du discours (Hana *et al.*, 2010).

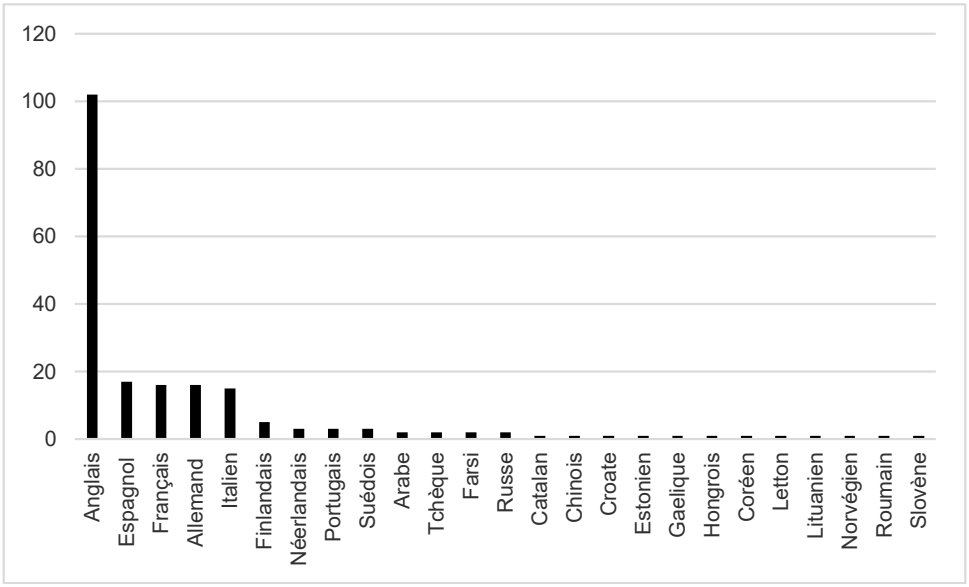

Figure 1 – CA existants classés selon la langue cible (Université Catholique de Louvain, 2018)

1.3. CA italiens

Comme on le voit dans la figure 1, 52% du total de 198 corpus existants[2] est dédié à l'anglais, suivi à grande distance de 9% de corpus espagnols et de 8% des corpus français et allemands.

[2] Données de l'Université Catholique de Louvain.

L'Italien est en quatrième position (7%) avec un nombre de 15 corpus dont 7 comprennent uniquement l'italien et les autres sont multilingues[3].

Malgré le vaste corpus de recherches sur l'acquisition de la L2, l'italien est actuellement largement sous-représenté en termes d'application de méthodologies quantitatives et statistiques aux données des apprenants, de conception et annotation des corpus d'apprenants, de ressources de traitement automatique des langues naturelles appliquées à la recherche sur les apprenants, et d'applications didactiques de la recherche sur les corpus d'apprenants (Spina, 2017). Martelli (2008) a analysé les erreurs d'un corpus de productions écrites d'apprenants d'italien LE. Andorno (2005) a étudié l'enchâssement des particules additives et restrictives dans l'énoncé verbal en italien L2. Gallina (2010 ; 2013) a utilisé le corpus *LIPS*[4] pour l'étude de l'acquisition du lexique (richesse lexicale et champs lexicaux) en italien L2. Konecny *et al.* (2016) ont classé les phrasèmes dans un CA d'italien L2 (*LEKO*[5]) et proposé un modèle d'enseignement mixte des phrasèmes (Zanasi *et al*., 2016). Bonsegna (2000) a étudié les faux amis italo-anglais dans la production écrite d'apprenants de niveau avancé d'italien LE.

La section suivante décrit une recherche menée par l'auteur (Casani, 2020) pendant le développement du projet *MERLIN*[6] (Université Technologique de Dresde, en partenariat avec l'Université Eberhard Karls Tübingen, l'*EURAC* (*European Academy*) de Bolzano/Bozen, et l'Université Charles IV Prague), un CA écrit trilingue (allemand, italien et tchèque) élicité par des apprenants de différentes langues maternelles.

2. Méthode

Nous avons analysé les textes écrits de 400 apprenants adultes d'italien LE (127 hommes et 273 femmes, âgés de 18 à 38 ans) du niveau A1 au niveau B2 du CECRL, de langue maternelle mixte avec une prévalence de germanophones (55%). Les autres langues maternelles sont le hongrois (17%), le polonais (13%), le français (10%), l'espagnol (1%), le portugais (1%) et d'autres (3%). Les textes constituent un sous-groupe de la section italienne du corpus MERLIN. Ils ont été élicités par des tests de certification du centre TELC[7] de Francfort, partenaire du gouvernement fédéral allemand et du projet MERLIN. Les tests ont été soumis aux procédures de contrôle de l'EALTA[8], conformément aux normes internationales de qualité en matière de tests. Les tâches demandées aux apprenants d'italien LE pour chaque niveau sont les suivantes :

A1
- aider un ami à la recherche d'un emploi;
- déplacer un rendez-vous;

A2
- proposer à un ami de lui rendre visite ;
- contacter un ami après une longue période ;
- informer des amis d'un cours de langue auquel on participe ;

B1
- aider un ami à la recherche d'un emploi après l'obtention de son diplôme ;

[3] Pour une brève description des CA italiens et de leurs possibilités d'exploitation pédagogique, voir Casani (2018). Pour une liste complète mise à jour des CA existants et des sites des projets respectifs, voir https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

[4] *Lessico dell'Italiano Parlato da Stranieri* http://www.parlaritaliano.it/~parole/index.php/en/corpora/653-corpus-lips.

[5] Lexemkombinationen und typisierte Rede im mehrsprachigen Kontext http://www.leko-project.org/.

[6] Mehrsprachige Plattform für die Europäischen Referenzniveaus https://merlin-platform.eu/.

[7] The European Language Certificates.

[8] European Association for Language Testing and Assessment.

- répondre à une lettre sur les plats typiques de son pays ;
- se renseigner sur un séjour d'étude ;
- répondre à une invitation de mariage ;

B2

- postuler pour un stage dans une entreprise ;
- parler de ses expériences d'apprentissage des langues ;
- rejoindre un projet de solidarité ;
- rejoindre les soirées de cuisine internationales ;
- répondre à une lettre d'une personne qui a des problèmes avec les chats ;
- porter plainte contre un hôtel.

Pour compiler le corpus, les productions écrites ont été extraites des tests originaux et réévaluées par deux évaluateurs formés à l'utilisation de deux grilles d'analyse conformes au CECRL. La première grille est similaire au tableau 3 du CECRL (Council of Europe, 2001 : 29-30), adaptée aux besoins des évaluateurs (Alderson, 1991) et opérationnalisée ; la seconde est l'échelle d'évaluation globale du répertoire linguistique général (Council of Europe, 2001 : 110). Le degré de fiabilité des évaluations a été vérifié à l'aide de la théorie classique des tests et de l'analyse multi-facettes de Rasch. Pour chaque critère d'évaluation, les productions ont reçu un score directement orienté sur les niveaux du CECRL, ce qui a permis d'obtenir un profil de compétence pour chaque production. Il convient de noter que, à la suite des réévaluations, seuls deux textes évalués comme B2 par *TELC* ont obtenu un jugement pleinement conforme aux niveau B2, de sorte que le B2 traité ici ne correspond pas à un niveau d'utilisateur pleinement indépendant mais à un niveau de première autonomie (B2.1). Les textes ont été transcrits à l'aide de *XML-mind* [9], un éditeur basé sur *XML*, par le personnel de *TELC*, qui a également veillé aux contrôles de fiabilité des transcriptions. Une fois la transcription terminée, les données ont été converties en *PAULA*[10], un *XML* conçu comme format d'échange pour les annotations linguistiques. Pour les annotations manuelles, on a utilisé l'outil *Falko*[11], un complément *Excel* développé par l'Université Humboldt de Berlin, pour les annotations des hypothèses cibles et l'outil *MMAX2*[12] pour les annotations multiniveaux des catégories linguistiques.

L'auteur a réalisé l'analyse décrite dans cette étude lors de la construction du corpus, à laquelle il a collaboré dans les années 2013 – 2014 à l'Université Technologique de Dresde. Les 400 textes, sélectionnés parmi ceux dans lesquels les annotations avaient été vérifiées, ont été importés dans *Annis*[13], un navigateur web gratuit pour la recherche et la visualisation de corpus linguistiques multiniveaux complexes, et analysés selon les 15 paramètres morphosyntaxiques suivants, extraits du schéma d'annotation *MERLIN*[14]:

- ordre syntaxique: erreurs d'ordre des constituants en phrases principales (*Mi molto dispiace au lieu de Mi dispiace molto) ou secondaires (*perché secondo me questo è un positivo gesto au lieu de perché secondo me questo è un gesto positivo);
- négation: erreurs d'omission (*lui viene mai più con noi in vacanza au lieu de lui non viene mai più con noi in vacanza), changement (*no ho molto tempo au lieu de non ho molto tempo) et position (*il mio luogo è non sviluppato

[9]https://www.xmlmind.com/
[10]Potsdamer AUstauschformat Linguistischer Annotationen purl.org/net/paula
[11]purl.org/net/Falko
[12]mmax2.net
[13]http://corpus-tools.org/annis/
[14]Les grilles d'évaluation *MERLIN,* informations détaillées sur la préparation et l'annotation du corpus, ainsi qu'une vaste bibliographie sont disponibles sur le site web du projet.

molto au lieu de il mio luogo non è molto sviluppato) de particules de négation ;

- valence: erreurs d'omission (*Hai superato con il massimo dei voti? au lieu de L'hai superato con il massimo dei voti?) ou addition (*vorrei che mi vi rimborsiate au lieu de vorrei che mi rimborsiate) de constituants obligatoires ;
- accord: erreurs d'accord sujet-verbe (**io vuole visitare Ville X*[15] au lieu de *io voglio visitare Città X*) ;
- flexions inexistantes: utilisation de formes non existantes dans les paradigmes de flexion des noms (**problemo* au lieu de *problema*), verbes (**Sposarei in bianco* au lieu de *Sposerei in bianco*) ou adjectifs (**una breva lettera* au lieu de *una breve lettera*) ;
- morphologie nominale: erreurs de genre (**i mani* au lieu de *le mani*), nombre (**Alle fine* au lieu de *Alla fine*) et cas (**persone come tu* au lieu de *persone come te*) dans les parties nominales de la phrase ;
- morphologie verbale: erreurs de mode (*So, che fra poco, avresti un proprio appartamento au lieu de So che fra poco avrai un proprio appartamento), temps (*Quando siamo ritornati, è possibile incontarti? au lieu de Quando saremo ritornati, è possibile incontrarti?) et aspect (*Il mio esame era abbastanza bene au lieu de Il mio esame è stato abbastanza buono) des verbes ;
- verbe principal: addition (*Andiamo tornare a casa au lieu de Andiamo a casa) et omission (*Un anno fa in Germania per una settimana au lieu de Un anno fa sono stato in Germania per una settimana) du verbe principal ;
- prédicats analytiques: erreurs dans la formation de prédicats complexes impliquant omission (**Michele e Luca andati in città* au lieu de *Michele e Luca sono andati in città*), addition (**Mentre ho leggevo il giornale* au lieu de *Mentre leggevo il giornale*), modification/substitution (**siamo potuti giocare* au lieu de *abbiamo potuto giocare*) de verbes auxiliaires, modaux, causatifs et de copulas ;
- conjonctions: erreurs d'omission (*Speriamo tu stai bene au lieu de Speriamo che tu stia bene), addition (*da due anni che sto imparando inglese au lieu de da due anni sto imparando l'inglese), substitution (*Sarà meglio quando uno di voi mi chiama au lieu de Sarà meglio se uno di voi mi chiama) et position (*e poi vuole fare la maturità anche au lieu de e poi vuole fare anche la maturità) des conjonctions ;
- pronoms clitiques et réfléchis: erreurs d'omission (*Vi scrivo per lamentare au lieu de Vi scrivo per lamentarmi), addition (*Come lo sai, avevo qualche problema au lieu de Come sai, avevo qualche problema), malformation (*se trova entro il Lago Maggiore au lieu de si trova entro il Lago Maggiore) et position (*puoi la chiamare au lieu de puoi chiamarla) dans les pronoms clitiques et réfléchis ;
- parties du discours: erreurs de substitution de parties du discours (**la città è molto bene* au lieu de *la città è molto bella*) ;
- prépositions : erreurs d'omission (*giocare tennis au lieu de giocare a tennis),

[15] Les noms de personnes et de villes ont été anonymisés.

addition (*Vorremmo di un parziale rimborso au lieu de Vorremmo un parziale rimborso) et substitution (*Le scrivo al riferimento au lieu de Le scrivo in riferimento) des prépositions ;

- articles: erreurs d'omission (*Lavoro anche in ristorante dei miei genitori au lieu de Lavoro anche nel ristorante dei miei genitori), addition (*In riferimento al questo annuncio au lieu de In riferimento a questo annuncio) et substitution (*i appartamenti au lieu de gli appartamenti) des articles.

Les analyses inférentielles de l'association entre niveau et type d'erreur et des différences de proportions d'erreurs entre niveaux de compétence ont été effectuées à l'aide de *IBM-SPSS24*.

3. Résultats

Le tableau 1 montre les proportions d'erreurs (intra-niveau) par niveau de compétence et catégorie grammaticale. L'association entre niveaux et catégories d'erreur est significative (χ^2 (39) = 116.017, $p < 0.001$). Les résultats de l'analyse *post hoc* (*Z*) sont représentés au moyen de couleurs différentes. Dans la même catégorie linguistique, les cases de couleur sombre indiquent des proportions d'erreur statistiquement plus élevées que celles de couleur claire ; les cases non peintes indiquent des proportions qui ne diffèrent pas de manière statistiquement significative ($p \leq 0.05$) des autres dans la même colonne.

Les tableaux 2, 3 et 4 montrent les proportions d'erreur (intra-niveau) groupées respectivement par morphologie nominale, morphologie verbale et construction syntaxique. Ces tableaux incluent les erreurs superficielles, c'est-à-dire les modifications des structures au niveau de la surface (addition, omission, changement, position). L'association entre niveaux et catégories d'erreur est statistiquement significative (*Fisher* = 275.411, $p < 0.001$; $V = 0.237$, $p < 0.001$). Les résultats de l'analyse *post hoc* (*Z*) ($p \leq 0.05$) sont représentés par les différentes couleurs (voir ci-dessus) dans les tableaux.

4. Courte discussion[16]

La plupart des catégories linguistiques ont tendance à la distribution normale des erreurs, avec un pic entre le niveau de survie et le niveau seuil. Ce cadre est compatible avec une introduction progressive des structures, qui sont absentes ou non-analysées dans le niveau A1, sont expérimentées dans les niveaux de survie et seuil, puis ont tendance à se consolider au fur et à mesure que l'interlangue évolue vers l'indépendance. Les exceptions sont certaines catégories qui présentent un maximum d'erreurs dans le niveau A1. Celles-ci diminuent dans l'A2 en proportion statistiquement significative pour la flexion inexistante des verbes, pour l'omission de parties des prédicats analytiques, pour les omissions de pronoms réfléchis et pour la substitution de parties du discours. La distribution des erreurs dans ces catégories tend alors à se normaliser aux niveaux suivants, sauf dans les prédicats analytiques, qui voient une nouvelle diminution numérique au B1 et une augmentation significative au B2. Dans le cas de telles tendances irrégulières, l'analyse de sous-catégories supplémentaires au niveau superficiel peut suggérer un ajustement possible des étiquettes d'annotation. Dans les prédicats analytiques, par exemple, seulement les omissions (principalement des copules et des verbes modaux) diminuent de manière significative dans le niveau B1, atteignant presque zéro, tandis que les malformations restent statistiquement constantes, ainsi que les ajouts, qui restent proches de zéro. Les erreurs de morphologie nominale (14,5%) sont présentes dans une proportion prépondérante par rapport à celles de morphologie verbale (8,8%) et restent statistiquement inchangées à tous les niveaux, contrairement aux erreurs verbales, qui

[16] Pour une discussion détaillée des données accompagnée d'exemples extraits du corpus, voir Casani (2020).

augmentent de manière statistiquement significative entre le niveau de découverte et le niveau de survie, parallèle à l'élargissement du répertoire des temps et des modes verbaux. Les plus grandes proportions d'erreurs se produisent dans les prépositions (20,7%) et les articles (15%), mais si les prépositions restent problématiques à tous les niveaux de compétence (avec une augmentation significative du nombre de substitutions au niveau A2), les articles sont soumis à une amélioration significative des omissions lors du passage au niveau seuil, ce qui conduit ensuite à une certaine tendance à la surutilisation au niveau B2. Le passage de phrases courtes juxtaposées à des expressions plus complexes entraîne une augmentation significative de l'omission des conjonctions au niveau A2, tandis que l'expérimentation de l'hypotaxe au niveau B2 entraîne une augmentation significative du nombre de substitutions de conjonctions. Il n'y a pas de variations significatives d'un niveau à l'autre en ce qui concerne les erreurs de distribution syntaxique (3,9%), de négation (0,8%), de valence (5%), d'accord sujet-verbe (2,9%) et d'ajout/omission du verbe principal (1,6%).

	Ordre synt.	Neg.	Val.	Acc. S/V	Flex. inex.	Morph. nomin.	Morph. verb.	Verbe princ.	Predic. analytiques	Conjonct.	Clit.+ refl.	Parties du disc.	Prep.	Art.
B2.1	3,2	0,6	4,1	3,8	3,5	16,4	9,8	1,9	6	6,9	6	3,8	17,7	16,4
B1	4,4	0,6	4,8	3,3	4,2	14,1	10,6	1,2	2,3	4,8	9,6	5	23,9	11,2
A2	4,1	1	4,1	3,3	3,3	15,6	11,1	0.8	3,3	6	3,9	3,9	21,6	17,9
A1	3,9	1	6,9	1,3	8	12,1	3,1	2,6	7,7	2,1	9,8	8,5	18,3	14,9

Tableau 1 - Erreurs par niveau et catégorie grammaticale (% intra-niveau)

	Flex. Inexist.	Aspect	Pred. analyt. (add)	Pred. analyt. (change)	Pred. analyt. (omiss)	Verbe princip. (add)	Verbe princip. (omiss)	Mode	Temps	Accord
B2.1	0	2,5	0	4,4	1,6	0,3	1,6	4,1	3,2	3,8
B1	1,9	1	0,4	1,7	0,2	0	1,3	6,5	3,1	3.3
A2	1,2	2,4	0	1,6	1,6	0	0,8	5,5	3,1	3,3
A1	4,4	0	0,3	4,1	3,3	0,3	2,3	2,6	0,5	1,3

Tableau 2 - Erreurs de morphologie verbale par niveau de compétence et catégorie linguistique (% intra-niveau)

	Flex. inexist. (nom)	Flex. inex. (adj)	Articl. (add)	Articl. (cha)	Articl. (omiss)	Clit. (add)	Clit. (chan)	Clit. (omiss)	Clit. (pos)	Refl. (add)	Refl. (chan)	Refl. (omiss)	Refl. (pos)
B2.1	3,5	0	7,6	1,9	6,9	1,3	0,6	3,2	0	0,6	0,3	0	0
B1	1,9	0,4	5,0	2,7	3,6	1,7	1,5	2,7	0,8	0	0,4	1,7	0,2
A2	1,8	0.2	4,9	1,0	11,4	1,6	0,6	1,0	0,2	0	0,2	0,2	0
A1	2,3	1,3	2,8	0,5	11,6	0,5	1,5	2,3	1,3	0	0,8	3,3	0

Tableau 3 - Erreurs de morphologie nominale par niveau de compétence et catégorie linguistique (% intra-niveau)

	Ordre syntax. (phrase princ.)	Ordre syntax. (phrase subordonnée)	Clitiques (pos.)	Réfléchis (pos)	Conjonctions (pos)	Negations (pos)
B2.1	1,9	1,3	0	0	0,6	0,3
B1	3,1	1,3	0,8	0,2	0,8	0
A2	3,3	0,8	0,2	0	1,4	0
A1	3,1	0,8	1,3	0	0,5	0,3

Tableau 4 – Erreurs de structure syntaxique par niveau de compétence et catégorie linguistique (% intra-niveau)

Les tableaux 1, 2, 3 et 4 peuvent être utilisés comme support pour l'évaluation de la morphosyntaxe des productions écrites d'apprenants d'italien LE ou pour l'étude des interlangues. L'utilisation peut être faite à un niveau plus général, en considérant uniquement le changement de couleur dans chaque catégorie linguistique. Dans la même colonne, la transition du gris foncé au gris clair indique une amélioration significative de la catégorie linguistique respective. Inversement, le passage du gris clair au gris sombre indique une détérioration significative de la catégorie linguistique concernée. La couleur blanche n'indique aucun changement significatif. À un niveau plus fin, les pourcentages peuvent donner une idée de la proportion d'erreurs pouvant être attendue à un certain niveau pour chaque catégorie linguistique. Les grilles peuvent être composées en fonction des besoins, en regroupant les catégories utiles, et mises en œuvre via l'analyse d'erreurs des CA. Les données statistiques peuvent atteindre un niveau de généralisation plus élevé grâce à la composition de grilles verbales combinant des descriptions quantitatives et qualitatives des erreurs, structurées selon les critères de formulation positive, de concision et de transparence du CECRL. En annexe on propose des descripteurs pour la compétence morphosyntaxique générale (annexe 1), pour la morphologie nominale (annexe 2) et verbale (annexe 3), ainsi que pour la construction syntaxique (annexe 4), résultats des analyses présentées dans cette étude. Ils peuvent être accompagnés d'exemples authentiques tirés des CA de référence à la fois pour les structures individuelles et pour les textes intégraux[1]. Ces descripteurs constituent une réponse possible aux recommandations du Conseil de l'Europe qui, depuis la publication du CECRL, a encouragé le développement d'outils d'aide au Cadre pour la description empirique des niveaux de compétence dans les différentes langues communautaires (voir § 1).

5. Conclusions

Au cours des trois dernières décennies, les corpus d'apprenants ont acquis un rôle de premier plan dans le domaine de l'acquisition de la L2/LE. À l'heure actuelle, il existe cependant encore peu d'études concernant l'application à la pratique réelle des résultats des CA italiens et l'italien est encore sous-représenté en termes de recherche appliquée aux CA.

Cette contribution illustre une application possible de la méthodologie d'analyse d'erreurs à un CA à travers une étude visant à la validation empirique des niveaux du CECRL pour la compétence morphosyntaxique en italien LE. Les résultats ont été synthétisés dans des grilles construites sur le modèle des descripteurs du CECRL. Ces grilles sont présentées en annexe comme une aide pour l'évaluation de la morphosyntaxe dans les productions écrites des apprenants d'italiens LE (Casani, 2020).

[1] Pour des exemples de différentes structures, voir Casani (2020). Pour consulter et télécharger les exemples ainsi que les textes complets, voir https://merlin-platform.eu/.

Références bibliographiques

ALDERSON, J. Charles (1991). Bands and scores, in *Language testing in the 1990s,* Alderson, J. Charles & North, Brian (Eds.). London: Modern English Publications and the British Council, 71-86.

ANDORNO, Cecilia (2005). Additive and restrictive particles in Italian as a Second Language. Embedding in the verbal utterance structure, in *The structure of learner varieties*, Hendriks, Henriëtte (Ed.). Berlin & New York: Mouton de Gruyter, 405–444.

BONSEGNA, Chiara (2000). Italian-English false friends in the written production of advanced Italian EFL learners: a corpus-based analysis. Unpublished Thesis. Torino: Università di Torino.

CARLSEN, Cecilie H. (2013). *Norsk Profil. Det europeiske rammeverket spesifisert for norsk. Et første steg*. Oslo: Novus forlag.

CASANI, Emanuele (2018). Corpus linguistics e didattica dell'italiano. I learner corpora nella classe di lingue, in *Nella classe di italiano come lingua seconda/straniera*, D'Angelo, Maria Carmela & Diadori, Pierangela (Eds). Firenze: Franco Cesati Editore, 157-167.

CASANI, Emanuele (2020). Valutare la competenza morfosintattica in italiano L2. Una validazione corpus-based dei livelli del QCER, in *Valutazione e misurazione delle produzioni orali e scritte in italiano lingua seconda*, Nuzzo, Elena, Santoro, Elisabetta & Vedder, Ineke (Eds). Firenze: Franco Cesati Editore.

COUNCIL OF EUROPE (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

COUNCIL OF EUROPE (2005). *Reference level Description for National and regional Languages. Guide for the production of RLD*. Strasbourg: Council of Europe Language Policy Division, DG IV.

GALLINA, Francesca (2010). The LIPS Corpus (Lexicon of Spoken Italian by Foreigners) and the acquisition of vocabulary by learners of Italian as L2, in *Papers from the Lancaster University Postgraduate Conference in Linguistics & Language Teaching*, *vol. 4: Papers from LAEL PG 2009*, Bota, Grace, Hargreaves, Helen, Chia-Chun, Lai & Rong, Rong (Eds). Lancaster, Lancaster University, 30-50.

GALLINA, Francesca (2013). The Lexicon of Spoken Italian by Foreigners: A study on the acquisition of vocabulary by L2 Italian learners between measures of lexical richness and lexical fields, in *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead. Proceedings of the First Learner Corpus Research Conference (LCR 2011)*, Granger, Sylviane, Gilquin, Gaëtanelle & Meunier, Fanny (Eds). Louvain-la-Neuve: Presses Universitaires de Louvain, 179–195.

GRANGER, Sylviane (2002). A bird's-eye view of learner corpus research, in *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, Granger, Sylviane, Hung, Joseph & Petch-Tyson, Stephanie (Eds). Amsterdam/Philadelphia: John Benjamins, 3-33.

GRANGER, Sylviane (2003). Error-tagged learner corpora and CALL: a promising synergy. Special issue on error analysis and error correction in computer-assisted language learning. *CALICO Journal, 20 (3)*: 465-480.

HANA, Jirka, ROSEN, Alexandr, SVATAVA, Skodov´ & STINDLOVÁ, Barbora (2010). Error-tagged Learner Corpus of Czech, in *Proceedings of the Fourth Linguistic Annotation Workshop, LAW 2010,* 11-19.

HULSTIJN, Jan H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal, 91*: 663-667.

KONECNY, Christine, ABEL, Andrea, AUTELLI, Erica & ZANASI, Lorenzo (2016). Identification and Classification of Phrasemes in an L2 Learner Corpus of Italian, in *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives* -

Fraseología computacional y basada en corpus: perspectivas monolingües y multilingües. Geneva: Tradulex, 533–542.
LEECH, Geoffrey (1998). Learner corpora: what they are and what can be done with them, in *Learner English on Computer,* Granger, Sylviane (Ed). London/New York: Longman, xiv-xx.
MARTELLI, Aurelia (2008). Error analysis and learner corpora: a study of errors in the written production by English students of Italian, in *Investigating English with corpora. Studies in honour of Maria Teresa Prat*, Martelli, Aurelia & Pulcini, Virginia (Eds). Milano: Polimetrica Publisher, 365–378.
SPINA, Stefania (2017). Learner Corpus Research and the acquisition of Italian as a second language: the case of the Longitudinal Corpus of Chinese Learners of Italian (LoCCLI), in *4th Learner Corpus Research Conference. LCR 2017. Book of Abstracts,* Abel, Andrea (Ed). Bolzano/Bozen: Eurac Research, 18-19.
SPINELLI, BARBARA & PARIZZI, Francesca (2010). *Profilo della lingua italiana*. Firenze: La Nuova Italia.
TAN, Melinda (2005). Authentic language or language errors? Lessons from a learner corpus, *ELT Journal, 59 (2)*, 26–34.
TONO, Yukio (2003). Learner corpora: Design, development and applications. In *Proceedings of the 2003 Corpus Linguistics Conference*, Archer, Dawn Elizabeth, Rayson, Paul, Wilson A. & McEnery, Tony (Eds). UCREL: Lancaster University, 800–809.
ZANASI, Lorenzo, KONECNY, Christine, AUTELLI, Erica & ABEL, Andrea (2016). L'insegnamento dei fraseologismi nell'italiano come lingua seconda: proposta di un modello didattico misto, in *Lingua al plurale: la formazione degli* insegnanti, De Marco, Anna (Ed). Perugia: Guerra, 171-180.

Annexes

Les descripteurs structurés à la suite des présentes analyses (Casani, 2020) sont mentionnés ci-dessous.

	Morphosyntaxe générale
B2.1	Il fait preuve d'une bonne maîtrise grammaticale. Les prédicats sont généralement complets, mais ils continuent à faire des erreurs, évidentes dans les constructions analytiques, en particulier avec les participes et les modaux. Ces constructions sont cependant généralement complètes. Les erreurs de morphologie nominale sont encore fréquentes, en particulier dans le genre, mais il ne commet pas d'erreurs pouvant provoquer des malentendus.
B1	Il utilise de manière raisonnablement correcte des formules de routine et des structures fréquemment utilisées, liées aux situations les plus prévisibles. Il formule des prédicats généralement complets dans lesquels des omissions des actants continuent à se produire, généralement dans les références pronominales, mais l'ajout de compléments superflus est rare. Malgré quelques erreurs, ce qu'il communique est généralement clair.
A2	Il utilise correctement certaines structures simples, dans lesquelles il exprime les prédicats de manière assez complète, omettant parfois les actants et commettant des erreurs systématiques de base. Par exemple, il confond les temps et les modes verbaux et il oublie de signaler les accords, mais les formes inexistantes diminuent et les parties du discours se distinguent généralement.
A1	Il possède une maîtrise limitée de structures grammaticales simples et de modèles syntaxiques simples dans un répertoire stocké. Il construit des prédicats partiels en omettant les actants et en ajoutant ceux qui sont superflus. Il utilise souvent des noms et des verbes de manière indistincte et expérimente leur flexion, créant souvent des formes personnelles.

Annexe 1 – Exemples de descripteurs de la compétence morphosyntaxique générale

	Morphosyntaxe nominale
B2.1	Bien que des difficultés persistent avec le genre et le nombre, il utilise des constructions généralement claires et plus complexes, en utilisant des références pronominales dans lesquelles il fait parfois des erreurs
B1	Ce qu'il communique est généralement clair. Il fait un usage intensif des articles, mais il a toujours des difficultés avec l'accord de genre et de nombre.
A2	Les formes inventées diminuent, bien que des erreurs de genre et de nombre persistent.
A1	Il fait un usage personnel de la flexion, quels que soient les accords.

Annexe 2 - Exemples de descripteurs de la morphologie nominale

	Morphosyntaxe verbale
B2.1	Il utilise toujours correctement des formes verbales existantes, même s'il fait encore des erreurs, par exemple en remplaçant des auxiliaires, ce qui ne compromet pas la compréhension du message. Des erreurs de temps et d'aspect persistent également, souvent dues à l'influence de la langue maternelle.
B1	Il fait un usage systématique de l'aspect verbal et exprime la modalité d'une manière assez claire, mais il continue de commettre des erreurs dans les temps, en particulier dans l'accord de formes perfectives.
A2	Il utilise presque toujours des formes verbales existantes, bien qu'il fasse encore des erreurs d'accord, des omissions d'auxiliaires et de modaux. Bien qu'avec des erreurs, il commence à exprimer des différences d'aspect et des modes différentes de l'indicatif.
A1	Il utilise souvent des formes verbales inexistantes.

Annexe 3 - Exemples de descripteurs de la morphologie verbale

	Construction syntaxique
B2.1	Il utilise des constructions assez correctes, aussi hypothétiques, dans lesquelles on trouve parfois des adverbes et des connecteurs dans une position non standard.
B1	Il utilise des constructions syntaxiques plus articulées, introduisant des phrases secondaires (objectives, temporelles et certaines relatives) dans lesquelles cependant l'ordre des constituants ne correspond parfois pas à la norme.
A2	Il dispose d'un petit répertoire linguistique dans lequel des constructions atypiques se produisent en raison d'échanges nom/adjectif, verbe/adverbe et position de la conjonction *anche*.
A1	Il utilise des phrases courtes, dans lesquelles les possibilités de constructions incorrectes sont plutôt limitées. Cependant, des échanges sujet/verbe et des positions incorrectes des adverbes (*molto, sempre, insieme*) se produisent.

Annexe 4 - Exemples de descripteurs de la construction syntaxique