BABYLONIA 2|2024

Corpora d'apprendenti di lingue straniere

Corpus d'élèves en langue étrangère

Korpora von Fremdsprachen-Lernenden

Students corpora in foreign languages

Corpus da students e studentas da lungatgs jasters

Rivista per l'insegnamento e l'apprendimento delle lingue

Zeitschrift für Sprachunterricht und Sprachenlernen

Revue pour l'enseignement et l'apprentissage des langues

Rivista per instruir ed emprender linguatgs

A Journal of Language Teaching and Learning

WWW.BABYLONIA.ONLINE


France
LONDON
GERMANY

Corpora d'apprendenti di lingue straniere
Che cosa ci rivelano corpora d'apprendenti sull'apprendimento delle lingue straniere?

Corpus d'élèves en langue étrangère
Que nous révèlent les corpus de production d'élèves sur l'apprentissage des langues ?

Korpora von Fremdsprachen-Lernenden
Was verraten uns Korpora von Lernenden über Fremdsprachenlernen?

Students corpora in foreign languages
What do corpora of students' productions tell us about Foreign Language Learning?

Corpus da students e studentas da lungatgs jasters
Tgei constateschan corpus da students e studentas partenent igl emprender da lungatgs jasters?

Responsabili della parte tematica:
Karine Lichtenauer & Anita Thomas

Babylonia
Rivista svizzera per l'insegnamento delle lingue

Trimestrale plurilingue
edito dalla
Associazione Babylonia Svizzera
cp 120, CH-6949 Comano
ISSN 1420-0007
no 3/anno XXX/2023

Con il sostegno di

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI
Département fédéral de l'intérieur DFI
Dipartimento federale dell'interno DFI
Departament federal da l'intern DFI
Bundesamt für Kultur BAK
Office fédéral de la culture OFC
Ufficio federale della cultura UFC
Uffizi federal da cultura UFC

Repubblica e Cantone Ticino

Corpora d'apprendenti di lingue straniere
Corpus d'élèves en langue étrangère
Korpora von Fremdsprachen-Lernenden
Students corpora in foreign languages
Corpus da students e studentas da lungatgs jasters

6 **Editoriale della redazione**

4 **Introduzione**
Karine Lichtenauer
Anita Thomas

14 **Learner corpora to meet learners' individual needs**
Gaëtanelle Gilquin

20 **Auf dem Weg zu einer großen Datenbasis für Deutsch als L2 – das Projekt DAKODA**
Interview mit Katrin Wisniewski und Matthias Schwendemann

26 **Learner corpus research meets foreign language education: Examples from the Swiss Learner Corpus SWIKO**
Nina Hicks
Thomas Studer

36 **«und dann isses aber trotzdem manchmal anders wie man spricht» – Verschmelzungsformen in der gesprochenen Wissenschaftssprache von Studierenden mit Deutsch als L1 und L2**
Matthias Schwendemann
Franziska Wallner

42 **Générer des principes didactiques à partir d'un corpus**
Gwendoline Lovey

52 **« Je ne sais pas », « Je n'sais pas », « Ch'sais pas », « Chais pas » … Quelle place pour la variation phonique dans l'enseignement/apprentissage du français langue étrangère ?**
Isabelle Racine

BABYLONIA

2|2024

58 **Les corpus comme input et comme output : l'exemple des marqueurs bon et bien**
Anita Thomas
France Rousset

64 **'Frapper' ou 'jeter sur', comment choisir ? Apport de l'analyse de corpus expérimentaux**
Mireille Copin
Inès Saddour

72 **Un corpus de productions écrites en français langue étrangère – matériau pour mieux comprendre les choix lexicaux d'apprenants multilingues**
Christina Lindqvist

78 **Teaching and learning foreign languages: insights from classroom corpus research**
Rosamond Mitchell
Florence Myles

86 **Graded readers in the EFL classroom – the example of Mary Shelley's *Frankenstein***
Janina Liechti

92 **Impressum**

prìmula *(non com.* ***prìmola****) s. f. [lat. scient. Primula, dal lat. mediev. primula, der. di primus «primo»]. – Genere di piante primulacee con alcune centinaia di specie erbacee quasi tutte perenni, in prevalenza delle zone montane e subalpine dell'emisfero settentrionale*

Editoriale

Corpus d'apprenant.es et premiers bourgeonnements

Après son numéro sur les variétés linguistiques en classe de langue, Babylonia s'intéresse à la diversité des productions d'apprenant.es. Après l'émerveillement du promeneur du dimanche découvrant les prairies en ville, nous vous invitons donc à suivre la botaniste travaillant avec acharnement pour classifier, étudier, et conserver les traces des premiers bourgeonnements linguistiques.

En ce début d'automne, nous souhaitons paradoxalement attirer votre attention vers l'une des premières fleurs du printemps : la primevère. Cette plante aux couleurs vibrantes symbolise le renouveau et la croissance, mais elle est aussi particulièrement vulnérable : sécheresse, fluctuation de températures, disponibilité des nutriments... autant de défis pour son épanouissement.

La rédaction de Babylonia s'engage avec une nouvelle rubrique, appelée Primula, à promouvoir la nouvelle génération d'enseignant.es/chercheur.ses en mettant en valeurs les travaux de mémoire d'étudiant.es en HEP. Vous trouverez notre premier fleuron dans ce numéro.

Nous vous souhaitons une belle lecture !

Korpora von Fremdsprachenlernenden und erste Knospen

Nach der Ausgabe über Sprachvarietäten im Sprachunterricht widmet sich Babylonia der Vielfalt der Produktionen von Lernenden. Nach dem Staunen des sonntäglichen Spaziergängers, der die Wiesen der Stadt erkundet, laden wir Sie ein, der Botanikerin zu folgen, die hart daran arbeitet, die ersten sprachlichen Knospen zu klassifizieren, zu studieren und zu bewahren.

Paradoxerweise möchten wir zu Beginn des Herbstes Ihre Aufmerksamkeit auf eine der ersten Blumen des Frühlings lenken: die Primel. Diese farbenfrohe Pflanze symbolisiert Erneuerung und Wachstum, ist aber auch besonders anfällig: Wassermangel, Temperaturschwankungen, Verfügbarkeit von Nährstoffen... all dies sind Herausforderungen für ihre Blüte.

Die Babyloniaredaktion hat sich mit einer neuen Rubrik namens Primula verpflichtet, die neue Generation von Lehrenden und Forschenden zu fördern, indem sie Abschlussarbeiten von PH-Studierenden hervorhebt. In dieser Ausgabe finden Sie unsere erste "Primel".

Wir wünschen eine gute Lektüre!

BA BY

Corpus di apprendenti e primi boccioli

Dopo il suo numero sulle varietà linguistiche nelle classi di lingue, Babylonia è interessata alla diversità delle produzioni di studenti/studentesse. Davanti allo sguardo meravigliato del passeggiatore della domenica alla scoperta dei giardini della città, vi invitiamo a seguire il botanico che si è dato da fare per classificare, studiare e conservare le tracce dei primi boccioli linguistici.

All'inizio dell'autunno, vorremmo attirare paradossalmente la vostra attenzione su uno dei primi fiori della primavera: la primula. Questa pianta dai colori vivaci simboleggia il rinnovamento e la crescita, ma è allo stesso tempo particolarmente vulnerabile: siccità, sbalzi di temperatura, disponibilità di nutrienti... tante sfide da affrontare per garantire lo sviluppo.

La redazione di Babylonia lancia una nuova sezione, chiamata Primula, che promuove la nuova generazione di docenti/ricercatori mettendo in evidenza il lavoro di tesi degli studenti HEP. Troverete il nostro primo fiore all'occhiello in questo numero.

Ci auguriamo che la lettura sia di vostro gradimento!

Corpus da studentas e students ed emprems brumbels

Suenter l'ediziun davart las varietads linguisticas ell'instrucziun da lungatgs s'interessescha Babylonia alla diversitad da producziuns da studentas e students. Suenter il smarvegl dil spassegiader dalla dumengia ch'examinescha ils praus dil marcau, envidein nus Vus da suandar la botanista che lavura cunscienziusamein vida classificar, studegiar e conservar ils emprems brumbels linguistics.

Paradoxamein vulin nus render attent all'entschatta digl atun ad ina dallas empremas flurs dalla primavera: la primula. Questa plontina da tuttas colurs simbolisescha la regiuvinaziun e la creschientscha, denton eis ella era specialmein vulnerabla: munconza d'aua, fluctuaziuns da temperatura, disponibilitad da substanzas nutritivas... tut quei ei sfidas per sia fluriziun.

La redacziun da Babylonia ei s'obligada cun ina nova rubrica, numnada Primula, da promover la nova generaziun da persunas d'instrucziun e perscrutadras e perscrutaders cun metter en evidenza lavurs da finiziun da studentas e students da scolas aultas da pedagogia. En questa ediziun anfleis Vus nossa emprema "primula".

Nus giavischein ina buna lectura!

LO NIA

Introduzione

CORPUS D'APPRENANT-ES : UN « BASQUETTE » DE RESSOURCES POUR L'ENSEIGNEMENT DES LANGUES ÉTRANGÈRES

LEARNERS' CORPORA: A "BASQUETTE" OF RESSOURCES FOR FOREIGN LANGUAGE TEACHING

Karine Lichtenauer
| Université de Genève
Anita Thomas
| Université de Fribourg

Quel meilleur moyen que de regarder les réalisations des élèves pour comprendre la manière dont leurs apprentissages progressent ?

La linguistique de corpus appliquée aux corpus d'apprenant-es de langue peut être vue comme une méthode de recherche empirique proche de l'expérience des enseignant-es de langue : écouter, lire, observer ce que font les élèves pour mieux comprendre leurs cheminements et les écarts de leurs productions d'avec les usages les plus courants. D'autant plus que le travail sur les corpus permet de renouveler la description de la langue et de sortir des carcans normatifs : L'étude de corpus présentant la langue telle qu'elle est utilisée par ses locuteurs permet de relativiser les normes présentées dans les grammaires et donne une place légitime à la langue parlée. Ainsi, les corpus permettent de documenter de manière concrète l'usage de la langue que ce soit comme L1 ou L2 et donner un apport spécifique à la didactique des langues.

What better way to understand how students learn a foreign language than to look at what they actually do with it?

When applied to corpora of language learners, corpus linguistics can be seen as an empirical research method that shows numerous similarities to the experience of language teachers: listening to students, reading and observing what they do in order to analyse their development and the gaps between their language and more common formulations. Indeed, working with corpora lets us revisit the description of language and break free from normative constraints: investigating corpora presenting the language as used by its speakers questions the norms presented in grammars and gives a legitimate place to spoken language. Corpora provides thus concrete documentation of language use, be it L1 or L2, and makes a specific contribution to language teaching.

Landure & Boulton définissent un *corpus linguistique* comme un « grand ensemble de textes authentiques représentatif d'une variété donnée quelle qu'elle soit et exploitable sous forme électronique » (Landure & Boulton 2010 : 57, selon la définition de McEnery et al. 2006 :5). Dans ce numéro de Babylonia toutefois, les textes ne sont pas représentatifs d'une « variété donnée » qui serait partagée par une communauté linguistique, mais plutôt d'interlangues, individuelles, non systématiques et en évolution constante. Les corpus d'apprenant-es dont il sera question dans ce numéro reflètent ainsi la progression des apprentissages, ils fournissent des indices sur les compétences langagières et communicatives des élèves, mais aussi sur leurs attitudes et comportements, sur l'efficacité de dispositifs didactiques et bien plus encore. Les articles publiés ici montrent bien le foisonnement des intérêts de recherche, des méthodes de recueil des données, des fonctions et des applications didactiques de chaque corpus.

Nous aurions pu procéder à une catégorisation des articles par types de corpus selon le médium (écrit, oral – avec ou sans vidéos), le genre de texte (interactions enseignant-e/élèves ou entre élèves, production écrite individuelle, production orale de discours scientifique, etc.), la langue (corpus monolingues, bilingues, multilingues) ou encore la taille (grand corpus, ensemble de plusieurs corpus, corpus restreints pour des questions de recherche particulières). A une telle catégorisation de corpus, une perspective plus orientée vers les types d'applications des résultats des recherches – vers l'enseignement des langues – nous a semblé favorable.

Le numéro s'ouvre sur deux articles qui montrent les enjeux de la recherche sur des corpus d'apprenant-s et les domaines d'applications des résultats. Tout d'abord, **Gaëtanelle Guilquin** présente des projets de recherche actuels basés sur des corpus d'apprenant-es et l'évolution du champ au cours des dernières années, notamment en termes d'applications didactiques concrètes. Cet état des lieux éclaire ainsi les bénéfices que les enseignant-es de langue, les concepteurs-trices de matériel pédagogique et autres acteurs-actrices peuvent tirer des recherches sur le sujet mais aussi des corpus existants pour mieux répondre aux besoins indi-

Dans ce numéro de Babylonia toutefois, les textes ne sont pas représentatifs d'une « variété donnée » qui serait partagée par une communauté linguistique, mais plutôt d'interlangues, individuelles, non systématiques et en évolution constante. Les corpus d'apprenant-es dont il sera question dans ce numéro reflètent ainsi la progression des apprentissages, ils fournissent des indices sur les compétences langagières et communicatives des élèves, mais aussi sur leurs attitudes et comportements, sur l'efficacité de dispositifs didactiques et bien plus encore.

Landure & Boulton define a linguistic corpus as 'a large set of authentic texts which are representative of any given language variety and exploitable in electronic form' (Landure & Boulton 2010 : 57[1], following the definition of McEnery et al. 2006 : 5). In this issue of Babylonia, however, the data is not representative of a 'given variety' shared by a linguistic community, but rather of individual, non-systematic and constantly evolving interlanguages. Thus, learner corpora discussed in this issue reflect the progression of learning. But not only does it provide clues about learner interlanguage and communication skills, it also informs us about attitudes and learning practices, about the effectiveness of teaching methods, and much more! The various articles illustrate the wide range of research interests, data collection methods, as well as the purpose of each corpus and the possible uses in language teaching.

We could have categorised the contributions by type of corpus according to the medium (written, oral – with or without videos), type of text (interaction between learners or with the teacher, individual written production, oral production of scientific discourse, etc.), language (monolingual, bilingual, multilingual corpora) or size (large corpus, set of several corpora, restricted corpora for particular

Anita Thomas est professeure en Français Langue Étrangère à l'Université de Fribourg et actuellement directrice de l'Institut de plurilinguisme. Ses recherches et son enseignement portent sur le rôle de l'input et de son traitement en classe de langue ainsi que chez les apprenant·e·s L2. Elle s'intéresse depuis près de 20 ans à l'utilisation des corpus comme ressource pour l'enseignement et la recherche.

Karine Lichtenauer est chargée d'enseignement en didactique de l'allemand à l'IUFE de l'Université de Genève.

1 Translated by the authors.

Explorer des corpus dont la simple présentation aiguise les envies de recherche.

viduels des apprenant-es. Les bénéfices ne vont toutefois pas sans défis à relever, comme le confirment **Katrin Wisniewski & Matthias Schwendemann** dans l'interview qu'ils nous ont accordée. Ils partagent leur expérience du projet DAKODA (Datenkompetenzen in DaF/DaZ: Exploration sprachtechnologischer Ansätze zur Analyse von L2-Erwerbsstufen in Lernerkorpora des Deutschen) qui vise à intégrer et rendre accessible un grand nombre de corpus d'apprenant-es d'allemand langue étrangère ou seconde dans un répositoire. Spécificité des corpus en fonction des questions de recherche, réflexions sur les métadonnées et le travail d'annotations, exigences éthiques du respect de l'anonymat, possibilités d'utilisation ou perspectives pour la recherche, l'interview nous fait ainsi entrer dans le détail du travail de corpus avant de donner un ensemble de conseils à celles et ceux qui souhaitent se lancer dans l'aventure des corpus d'apprenant-es.

La deuxième partie de ce numéro est consacrée aux facteurs qui influencent le développement des compétences linguistiques des apprenant-es de langue. **Thomas Studer & Nina Hicks** nous présentent le corpus suisse des apprenant-es SWIKO, un corpus multilingue de jeunes en langue de scolarisation et en langue étrangère, ainsi que deux utilisations concrètes : l'une pour mieux comprendre les corrélations entre types de tâche et réalisations des élèves ; l'autre pour une exploitation didactique de SWIKO en classe de langue pour soutenir l'apprentissage de la négation en allemand langue étrangère. En milieu universitaire, cette fois, **Matthias Schwendemann & Franziska Wallner** étudient la réalisation des contractions phonologiques en allemand L2 dans les discours académiques oraux produits par des étudiant-es à l'aide du corpus GeWiss (Gesprochene Wissenschaftssprache kontrastiv). Outre les comparaisons avec un corpus de productions académiques produites par des locuteurs-trices dont l'allemand est la langue première, l'analyse des productions d'apprenant-es montre des différences entre locuteur-trices qui étudient en Allemagne et à l'étranger.

Gwendoline Lovey montre comment l'étude d'un corpus audio d'interactions en classe entre quatre enseignantes à l'école primaire et leurs élèves respectifs peuvent donner des indications sur

research questions). We preferred, however, to concentrate on the types of applications of the scientific results – to concentrate on language teaching.

The issue opens with two articles providing deep insights into research with learner corpora and showing how the results can be useful to language teaching in a large sense. **Gaëtanelle Guilquin** presents current research projects on learner corpora as well as the evolution of the field in recent years with an emphasis on practical applications. This overview sheds light on the benefits to language teachers, designers of teaching materials and other stakeholders from this research and from existing corpora so to better meet the individual needs of learners. However, the benefits are not without their challenges, as **Katrin Wisniewski & Matthias Schwendemann** confirm in their interview. They share their experience with the DAKODA project (Datenkompetenzen in DaF/DaZ: Exploration sprachtechnologischer Ansätze zur Analyse von L2-Erwerbsstufen in Lernerkorpora des Deutschen): the creation of an open-access repository collecting a large amount of data from learners of German.

They explore characteristics of corpora according to the type of research question, metadata and annotation, issues on anonymisation requirements and the possible ways of exploiting the repository in research or in teaching. They include a series of tips for those wishing to embark on the adventure of learner corpora.

In the second part of this issue, we turn to the factors having an impact on the improvement of learner language skills. **Thomas Studer & Nina Hicks** present the Swiss learner corpus SWIKO, a multilingual young learner corpus in both the language of schooling and a foreign language. Using two examples, they illustrate how the corpus can be used: first to investigate the correlations between types of task and learners' achievements; and then to use SWIKO directly in the language classroom when teaching, here for negation in German as a foreign language lessons.

In German-speaking academic settings, **Matthias Schwendemann and Franziska Wallner** focus on phonological contractions in oral academic discourse

l'efficacité des pratiques enseignantes. Elle en dégage des pratiques d'introduction et d'accompagnement des activités propres à encourager les élèves à formuler des productions plus riches. **Rosamond Mitchell & Florence Myles** présentent des résultats de recherche sur la base du corpus d'interaction en classe de langue Learning French composé d'enregistrements audio et vidéo de 33 leçons de français langue étrangère à l'école primaire en Grande-Bretagne. Leurs analyses visent, d'une part, le rôle de la fréquence des mots dans l'input des enseignant-es et dans la langue français et, d'autre part, l'influence du niveau d'engagement perçu des élèves sur l'apprentissage de nouveaux mots.

L'interlangue des élèves et le développement de leurs compétences langagières seront au centre de la troisième partie, qui illustre bien la variété des phénomènes linguistiques étudiés que les corpus d'apprenant-es révèlent. En effet, les quatre contributions qui la composent visent à mieux comprendre le développement des compétences lexicales, pragmatiques ou phonologiques des élèves – en explorant des corpus dont la simple présentation aiguise les envies de recherche.

Isabelle Racine se penche tant sur la réception que sur la production de caractéristiques phonologiques du français oral sur la base des corpus PFC (Phonologie du Français Contemporain) et IPFC (Inter-Phonologie du Français Contemporain). La pragmatique est ensuite à l'honneur avec le corpus DiCoi (digitalisation – corpus – interaction) à partir duquel **Anita Thomas & France Rousset** étudient l'apprentissage, réceptif et productif, de marqueurs discursifs en français. Leur étude longitudinale du développement des compétences des élèves comprend une intervention didactique qui engage les apprenant-es à observer des usages spontanés en L1.

Les articles suivants portent notre attention sur les choix lexicaux des élèves, mais par des méthodes d'investigation très différentes. Avec des méthodes d'élicitations issues de la psycholinguistique et des analyses des champs conceptuels fondées sur la sémantique cognitive, **Mireille Copin & Inès Saddour** proposent de travailler sur des corpus de type expérimental pour des phénomènes sémantiques complexes. De tels corpus

produced by students with German as a Foreign or Second language using the GeWiss corpus (Gesprochene Wissenschaftssprache kontrastiv). In addition to comparisons with a corpus of academic productions produced by speakers of German as a first language, the analysis of learner productions shows differences between speakers studying in Germany or abroad.

Gwendoline Lovey analyses an audio corpus of classroom interactions between four primary school teachers and their respective learners to find out how efficient different teaching practices are. She deduces from this how to introduce or scaffold foreign language activities in order to help learners producee more complex language.

Rosamond Mitchell & Florence Myles share research results also based on a classroom interaction corpus: Learning French collects 33 lessons of French filmed in English primary schools. They focus on the role of word frequency in teacher input and in the French language and investigate how learners' perceived level of commitment affects their learning of new words.

Exploring corpora whose mere presentation whets the appetite for research

The third part of this issue deals with learners' interlanguage and the development of their language skills, and thus illustrates the variety of linguistic phenomena revealed by learner corpora. The four contributions seek a better understanding of the development of learner lexical, pragmatic or phonological skills by exploring corpora whose mere presentation whets the appetite for research.

Qu'il s'agisse du développement des compétences grammaticales, lexicales, pragmatiques ou phonologiques des élèves ou encore des facteurs-clés ayant un impact sur l'ampleur et la qualité de leurs réalisations linguistiques, les auteur-es de ce numéro montrent ce que les corpus d'élèves leur ont dévoilé et les implications directes de ces connaissances pour une amélioration des pratiques enseignantes ou pour la rédaction de supports didactiques.

permettent d'observer les différences et similarités dans les choix lexicaux des apprenant-es de français langue étrangère et de francophones, ce qui aide les enseignant-e-s à mieux comprendre comment enseigner (et corriger) le lexique. **Christina Lindqvist** interroge un grand corpus de productions écrites en FLE par des élèves plurilingues du secondaire en Suède. L'analyse détaillée des transferts lexicaux, et, plus généralement, des choix lexicaux des élèves, font émerger une connaissance en profondeur de leurs compétences lexicales plurilingues. L'article se termine sur des pistes didactiques pour travailler sur un tel corpus dans une approche plurilingue.

Qu'il s'agisse du développement des compétences grammaticales, lexicales, pragmatiques ou phonologiques des élèves ou encore des facteurs-clés ayant un impact sur l'ampleur et la qualité de leurs réalisations linguistiques, les auteur-es de ce numéro montrent ce que les corpus d'élèves leur ont dévoilé et les implications directes de ces connaissances pour une amélioration des pratiques enseignantes ou pour la rédaction de supports didactiques. Les exemples d'utilisation des corpus en classe, avec le matériel didactique, illustrent par ailleurs la variété des recherches actuelles sur le *Data-Driven-Learning*.

Nous vous souhaitons une lecture inspirante !

Isabelle Racine looks at both the reception and production of phonological features in spoken French, using the PFC (Phonologie du Français Contemporain) and IPFC (InterPhonologie du Français Contemporain) corpora. Pragmatics then takes centre stage in the contribution by **Anita Thomas & France Rousset**. They investigate the receptive and productive learning of discourse markers in French with the corpus DiCoi (digitalisation - corpus - interaction). Their longitudinal study of the development of students' skills includes an intervention that engages learners in observing spontaneous uses in L1.

The next articles both focus on learners' lexical choices, but with very different methods of investigation. With a corpus elicited with psycholinguistic methods and analysed along cognitive semantic lines, **Mireille Copin & Inès Saddour** demonstrate how to work on experimental types of corpora to address complex semantic phenomena. They observe differences and similarities in the lexical choices of students of French as a Foreign Language on one hand, and Francophones on the other in order to understand how to teach (and correct) these phenomena.

Christina Lindqvist examines a large corpus of written production by plurilingual secondary school learners of French as a Foreign Language in Sweden. The detailed analysis of lexical transfers and, more generally, of the pupils' lexical choices, reveals in-depth knowledge of their plurilingual lexical skills. The article concludes with teaching suggestions for working with such a corpus in a plurilingual approach.

Whether they study the development of learner grammatical, lexical, pragmatic, or phonological skills or the key-factors impacting on quantity and quality of learner linguistic achievements, the authors of this issue show us what learner corpora have revealed to them as well as practical uses of this knowledge for improving teaching practices or creating teaching materials.

The examples of the use of corpora in the classroom, along with teaching materials, also illustrate the variety of current research on *Data-Driven-Learning*.

We wish you inspiring reading!

Whether they study the development of learner grammatical, lexical, pragmatic, or phonological skills or the key-factors impacting on quantity and quality of learner linguistic achievements, the authors of this issue show us what learner corpora have revealed to them as well as practical uses of this knowledge for improving teaching practices or creating teaching materials.

Landure, Corinne & Alex Boulton (2010). « Corpus et autocorrection pour l'apprentissage des langues », ASp [En ligne], 57 | 2010, consulté le 5 octobre 2022. URL : http://journals.openedition.org/asp/931

McEnery, T., R. Xiao et Y. Tono (2006). *Corpus-Based Language Studies: An Advanced Resource Book*. Londres : Routledge.

Tema

LEARNER CORPORA TO MEET LEARNERS' INDIVIDUAL NEEDS

Les corpus d'apprenants, comme d'autres corpus, ont surtout permis de faire des généralisations sur des populations entières. Ils peuvent cependant être exploités à des fins pédagogiques de manière plus différenciée et inclusive, en montrant comment des apprenant·es avec un profil spécifique utilisent (ou sont susceptibles d'utiliser) la langue cible. Une telle approche peut s'appuyer sur les métadonnées de corpus d'apprenants existants ou sur des données de corpus recueillies parmi ses propres étudiant·es. Les résultats issus de l'analyse de ces corpus peuvent aider à développer du matériel et des activités pédagogiques sur mesure pour répondre aux besoins de groupes d'apprenant·es particuliers ou d'apprenant·es individuel·les, y compris des activités d'apprentissage sur corpus *(data-driven learning)*, grâce auxquelles les étudiant·es peuvent faire des découvertes sur leur propre utilisation de la langue cible.

Gaëtanelle Gilquin
| UCLouvain

Gaëtanelle Gilquin is Professor of English Language and Linguistics at the University of Louvain, Belgium. She has co-edited the Cambridge handbook of learner corpus research and is the coordinator of several learner corpus projects, including the Louvain International Database of Spoken English Interlanguage.

Learner corpora: generalizing trends

Learner corpora started to be collected in the 1990s, with the aim of providing linguists (including lexicographers) with large electronic databases of authentic language produced by second/foreign language (L2) learners. One of the earliest learner corpora, the International Corpus of Learner English (ICLE), was first published in 2002 and contained some 2.5 million words of L2 English written by learners from 11 different mother tongue (L1) backgrounds (Granger et al., 2002).

Since then, learner corpora have kept growing. The EF-Cambridge Open Language Database (EFCAMDAT), for example, another learner corpus of written English, is currently made up of some 50 million words and represents almost 200 nationalities (based on Shatz, 2020). Spoken learner corpora tend to be smaller, which reflects the time and effort needed to collect and transcribe speech, but the largest spoken learner corpora now come close to 5 million words (4.2 million words for the Trinity Lancaster Corpus (TLC), see Gablasova et al., 2019).

The increasingly large size of learner corpora is usually seen as a welcome development, since large learner corpora are likely to better represent certain learner populations and include more instances of specific linguistic phenomena than small learner corpora. Generalizations made on the basis of large learner corpora therefore tend to be more reliable. This is an important feature for many studies, because learner corpora, similarly to other corpora, have mostly been used to establish what is frequent in language and common to a majority of writers/speakers.

Gilquin et al. (2007), for example, is a guide included in the second edition of the *Macmillan English Dictionary for Advanced Learners*, which is meant to help learners of English produce better academic and professional writing. In this guide, distinctive learner usage is only

The across-the-board approach usually adopted in learner corpus research may not be fully satisfactory to teachers who aim to meet learners' individual needs.

mentioned if it is frequent in the learner corpus used (ICLE) and typical of a majority of the learner groups represented in the corpus (the groups being defined by the learners' L1). Thus, the guide includes a 'Be careful' note about *of course*, because many learners from different L1 groups overuse *of course* in academic writing. By contrast, the overuse of *in fact* is not mentioned in the guide, because it is a feature that mainly characterizes French- and Italian-speaking learners.

Generalizations can be made at more specific levels than that of 'all learners of a target language'. For instance, O'Keeffe & Mark (2017), focusing on grammatical structures used correctly in learner English, adopt several criteria to ensure the widespread use of the structures: they should be frequent in the learner corpus, spread across a range of learners from several L1 families, occur in different registers/tasks, etc. However, a distinction is drawn between learners with different proficiency levels, so that the "grammatical competence statements" (O'Keeffe & Mark, 2017: 457), which show what learners are usually able to do, apply to learners with a given proficiency level, e.g. "Can use the affirmative form of the past perfect simple" (O'Keeffe & Mark, 2017: 476) for learners with a B1 level according to the Common European Framework of Reference for Languages (CEFR).

Such generalizing trends have provided many valuable insights into learner language (see, e.g., Granger et al., 2015) and have led to useful teaching applications. The analysis of the TLC, for example, has helped develop classroom activity worksheets available via the TLC Hub (https://cass.lancs.ac.uk/trinity-lancaster-corpus/). These worksheets highlight strategies – some successful, others less successful – which are often employed by learners in the TLC. Thus, it is shown that successful C1-C2 speakers (along the CEFR scale) use *I don't agree* and *I agree but...* more often than *I disagree* or *I can't agree* to express disagreement (Brezina, 2017). However, such an across-the-board approach may not be fully satisfactory to teachers who aim to meet learners' individual needs.

Differentiated and inclusive instruction with existing learner corpora

Learner language is known to be very heterogeneous, being affected by a large variety of factors such as the learner's L1, knowledge of additional languages, exposure to the target language, but also task, timing, access to reference tools, etc. In a language classroom, learners are therefore likely to use the target language in distinct ways. This is all the more so in mixed classrooms, which have become more common recently, partly as a result of educational policies. Mixed-age classrooms, mixed-ability classrooms, ethnically mixed classrooms, etc. bring together learners from various backgrounds, with distinctive characteristics, diverse aspirations, and so on. If teachers want to offer differentiated and inclusive instruction, as they are often encouraged to do, they need to adapt to the diversity of the classroom and seek to cater for the individual needs of each of their students.

Learner corpora, many of which are publicly available,[1] can help with this, because they can provide information about the language behaviour of learners with specific profiles. If a classroom includes students from a range of L1 backgrounds, different existing (sub)corpora can be exploited that represent each of these L1 backgrounds. Even if teachers are not familiar with these L1s and therefore not aware of the particular challenges that speakers of these L1s may face when trying to learn the target language, they will be able to find out about these in the learner corpora.

1 See https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html for a list of learner corpora (Centre for English Corpus Linguistics, 2024).

For example, most learners of English find it hard to use high-frequency verbs (such as *make*, *take*, or *give*) appropriately, because these verbs are highly polysemous and in some cases the choice of the verb is largely arbitrary (compare *give a talk* and *make a comment*). However, learners with different L1s tend to produce distinct non-standard combinations with these verbs. Huiping & Yongbing (2014) observe that, in the International Corpus of Crosslinguistic Interlanguage (ICCI), Chinese learners often use *make* with the collocate *party* (e.g. *make a birthday party*). They explain this by the fact that the equivalent of *make* in Chinese, *zuo*, can be used with the meaning of "causing (a birthday party, banquet, etc.) to take place" (Huiping & Yongbing, 2014: 268). Austrian learners, by contrast, produce no such collocations (ibid.). Based on an analysis of *make* in the French and Swedish ICLE subcorpora, Altenberg & Granger (2001: 180) point to combinations that also seem to be caused by transfer from the L1 and hence tend to be specific to these L1 groups, e.g. *make a poll* (instead of the more standard *carry out/conduct a poll*) for French-speaking learners and *make harm* (instead of *do harm*) for Swedish learners.

Learner corpora can help teachers offer differentiated and inclusive instruction by providing them with information about the language behaviour of learners with specific profiles.

Such findings can help provide students with pedagogical activities that address their language specificities (according to their L1 or some other feature), for instance in the form of L1-influenced collocations to be corrected. These targeted activities can stimulate interesting classroom discussions, where students explain how the equivalent word or structure is used in their L1, so that everybody can learn about each other's L1 and become more aware of crosslinguistic variation and also more respectful of differences.

Customizing pedagogical materials and activities according to the diversity of the classroom on the basis of existing learner corpora is made possible by the rich metadata that most learner corpora contain. In the Louvain International Database of Spoken English Interlanguage (LINDSEI; Gilquin et al., 2010), for example, each learner interview is described in terms of 23 variables, including learners' L1, how long they have been learning English, and how much time they have spent in an English-speaking country.

Each learner corpus comes with its own set of variables, some of which may be particularly relevant in certain teaching contexts. Thus, the Process Corpus of English in Education (PROCEED; Gilquin, 2022) is a learner corpus made up of argumentative essays as well as data showing the process through which these essays were composed (keylog files and screencast videos). Its metadata include information about learners' possible neurodivergence, e.g. whether they were diagnosed with dyslexia or ADHD. Using data from this corpus, it would for example be possible to discover the successful strategies of high-functioning dyslexics writing in L2 English (see Radar & Gilquin, forthcoming) and present these as potential models to dyslexic students who struggle with L2 writing tasks.

While the first compiled learner corpora were mostly made up of written English produced by advanced learners, over the years learner corpora have diversified, representing more varied target languages, L1 backgrounds, proficiency levels, etc. (see Gilquin & Granger, forthcoming). Yet, not all possible student profiles will have their corresponding learner corpora. Less commonly taught languages and less typical learners (including heritage language learners or learners with special educational needs), in particular, are not well represented among existing learner corpora.

In addition, if teachers want to target a very specific profile, they may end up with a small sample, even if they use a large learner corpus to start with (see Callies, 2015: 52). Thus, although the current version of ICLE contains over 5.5 million words, of which almost 500,000 were written by close to 1,000 Chinese-speaking learners, there is only one text produced by a Chinese-speaking

learner with Italian as an attested L2, corresponding to 455 words. When existing learner corpora do not provide sufficient or sufficiently relevant data, teachers can turn to data collection among their own students.

Individual tailoring with local learner corpora

Most learner corpora exploited for teaching purposes are "learner corpora for delayed pedagogical use" (Granger, 2009: 21). This means that they are compiled among a certain learner group (usually by academics or publishers) and exploited later among a different learner group (e.g. through the use of a textbook which was written with the help of the learner corpus data). However, teachers can also rely on "learner corpora for immediate pedagogical use" (ibid.), that is, corpora that are collected among the learners who will benefit from their pedagogical exploitation. Such corpora are called "local learner corpora" (Seidlhofer, 2002), because they represent a local learner group, typically the students of the teacher who is in charge of the corpus compilation.

Local learner corpora are usually collected as part of the day-to-day teaching activities. Whenever the students have to complete a writing task, the texts can be added to the corpus. The same is true of spoken tasks, although the time necessary for the transcription of speech may be an obstacle to the compilation of spoken local learner corpora. Provided they are made in a principled way, the teacher's corrections can be integrated into the corpus too, having the function of 'error tagging' and making it possible to retrieve certain error types automatically (e.g. all incorrect uses of the auxiliary *can*) and generate statistics (e.g. learners' progress in spelling over the weeks).

Learner corpora collected from one's own students are usually quite small, but they are truly representative of the students' language production. These corpora can be analyzed to bring to light patterns that are characteristic of the students' L2 usage. These observations can then be turned into pedagogical materials or activities that target the students' specific needs. Rankin & Schiftner (2011), for instance, show how the analysis of a local learner corpus of German-speaking learners of English revealed an overuse and predominantly non-standard use of the marginal preposition *concerning* (e.g. *Alberto showed no real progress concerning grammar*). On the basis of this finding, they prepared targeted exercises, including sentences taken from the corpus in which the students were required to find an alternative to the non-standard uses of *concerning*.

> An individual local learner corpus reveals the linguistic features typical of a learner's idiolect and makes it possible to offer tailor-made feedback and instruction to the learner.

Each learner has their unique way of using the L2, which is the result of multiple factors, such as the type and amount of input that they have received, their capacity to remember and reproduce words or structures that they have been exposed to, or their creativity in applying language patterns. Using learner corpus data produced by learners with a similar profile, even from the same classroom, can therefore only provide an approximation of a learner's language system. For many purposes, this approximation will be good enough – and, in any case, better than across-the-board generalizations. However, in some pedagogical contexts, it may be desirable to get to the uniqueness of each learner. This can be done by compiling local learner corpora made up of data produced by individual students.

An individual local learner corpus comprises texts in L2 produced by one and the same learner. The examination of such a corpus reveals the linguistic features typical of the learner's idiolect. This information makes it possible to offer tailor-made feedback and instruction to the learner. Importantly, this approach avoids exposing students to learner language features (including errors) that may not apply to them – one of the main criticisms levelled at the use of learner corpora in pedagogy (see, e.g., Flowerdew, 2001). Thanks to individual local learner corpora, students can also situate themselves in relation to the whole group, find

out what aspects of language they already master, and see what progress they have made. Although traditional graded assignments may allow for this too, texts brought together in the form of a corpus can be queried using the tools and techniques of learner corpus research, which can facilitate analysis and give access to more precise and accurate information than would otherwise be available, such as word frequencies, collocations, or keywords (that is, words distinctive of the corpus at hand as compared to a reference corpus).

Using learner corpora in different contexts

Learner corpora can be used by different educational actors and for different functions. They can help language testers develop tests that are suited to learners with special characteristics (e.g. learners with speech disorders) and set fair and achievable standards for them. Teaching materials writers can produce resources that take better account of learners' realities (what they can already do, what they have difficulty with, etc.) and offer contents that are adapted to certain learner groups (e.g. beginners or Spanish-speaking learners). Teachers, provided they have received training in corpus linguistics, can also integrate learner corpora into their teaching routine. Given their students' profiles, they can select the most relevant information and cater for the learners' specific needs by providing them with tailor-made activities. Students can also be given direct access to learner corpus data and be encouraged to make discoveries about learner language themselves, through so-called data-driven learning (see Gilquin & Granger, 2022). With the right type of learner corpus (which could be a corpus of their own language production or a corpus of language produced by learners with similar characteristics), they can embrace their own individuality and become responsible for their differentiated learning.

The way learner corpora can be exploited for pedagogical purposes will vary according to the context. With beginners, for example, highly rated texts from learner corpora can be a source of relatively simple sentences and an achievable target to be used as a model. With more advanced students, learner corpora compared with corpora of expert language can help raise awareness of fossilized errors (Nesselhauf, 2004). In data-driven learning, the teacher's role may be more or less prominent depending on the students' degree of autonomy and their observation and abstraction skills.

Despite the value of (local) learner corpora in highlighting what exactly each learner needs and hence promoting pedagogical differentiation and inclusion, one should not underestimate the difficulty of the endeavour. Offering individualized teaching to one's students clearly requires more time and effort than one-size-fits-all teaching, and the more one seeks to take the diversity of the classroom into account (that is, the closer one aims to get to the uniqueness of learners' language systems), the more work will be involved. Having to carry out extensive analyses on existing learner corpora or compiling learner corpora for this purpose may put too heavy a burden on the shoulders of teachers who are often already overburdened. However, even modest incursions into the realm of learner corpora and pooling of learner-corpus-based resources among teachers whose students have similar profiles should, when combined with other differentiated teaching practices, contribute to more inclusive classrooms, in which all learners feel respected and valued for their differences.

Through data-driven learning, students can embrace their own individuality and become responsible for their differentiated learning.

References

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics*, *22*(2), 173-195.

Brezina, V. (2017). Pragmatic functions: Expressing disagreement. Learning from Assessment: CEFR level C1 – Activity worksheet 1. Available at: https://www.trinitycollege.com/resource/?id=7979.

Callies, M. (2015). Learner corpus methodology. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 35-55). Cambridge University Press.

Centre for English Corpus Linguistics. (2024). Learner Corpora around the World. Louvain-la-Neuve: Université catholique de Louvain. Available at: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

Flowerdew, L. (2001). The exploitation of small learner corpora in EAP materials design. In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small corpus studies and ELT: Theory and practice* (pp. 363-379). John Benjamins.

Gablasova, D., Brezina, V., & McEnery, T. (2019). The Trinity Lancaster Corpus: Development, description and application. *International Journal of Learner Corpus Research*, *5*(2), 126-158.

Gilquin, G. (2022). The *Process Corpus of English in Education*: Going beyond the written text. *Research in Corpus Linguistics*, *10*(1), 31-44. Available at: http://ricl.aelinco.es/first-view/174-Article%20Text-1066-1-10-20210407.pdf.

Gilquin, G., De Cock, S., & Granger, S. (2010). *The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM*. Presses universitaires de Louvain.

Gilquin, G., & Granger, S. (2022). Using data-driven learning in language teaching. In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics. Second edition* (pp. 430-442). Routledge.

Gilquin, G., & Granger, S. (Forthcoming). L2 English learner varieties. In R. Reppen, L. Goulart, & D. Biber (Eds.), *The Cambridge handbook of English corpus linguistics. Second edition*. Cambridge University Press.

Gilquin, G., Granger, S., & Paquot, M. (2007). Improve your writing skills (Writing sections). In M. Rundell (Editor in chief), *Macmillan English dictionary for advanced learners. Second edition* (pp. IW1-IW29). Macmillan Education.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 13-32). John Benjamins.

Granger, S., Dagneaux, E., & Meunier, F. (2002). *International Corpus of Learner English. Handbook and CD-ROM*. Presses universitaires de Louvain.

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge handbook of learner corpus research*. Cambridge University Press.

Huiping, Z., & Yongbing, L. (2014). A corpus study of most frequently used English verbs by Chinese beginner learners from a conceptual transfer perspective. *International Journal of Corpus Linguistics*, *19*(2), 252-279.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In J. McH. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 125-152). John Benjamins.

O'Keeffe, A., & Mark, G. (2017). The English Grammar Profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, *22*(4), 457-489.

Radar, L., & Gilquin, G. (Forthcoming). A corpus study of dyslexic university students' L2 writing processes. *Revue de linguistique et de didactique des langues*.

Rankin, T. & Schiftner, B. (2011). Marginal prepositions in learner English: Applying local corpus data. *International Journal of Corpus Linguistics*, *16*(3), 412-434.

Seidlhofer, B. (2002). Pedagogy and local learner corpora: Working with learning-driven data. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 213-234). John Benjamins.

Shatz, I. (2020). Refining and modifying the EFCAMDAT: Lessons from creating a new corpus from an existing large-scale English learner language database. *International Journal of Learner Corpus Research*, *6*(2), 220-236.

Intervista

AUF DEM WEG ZU EINER GROSSEN DATENBASIS FÜR DEUTSCH ALS L2 – DAS PROJEKT DAKODA

Interview mit **Katrin Wisniewski** und **Matthias Schwendemann**

Katrin Wisniewski
| Universität Leipzig
Matthias Schwendemann
| Universität Leipzig

Ihr Projekt DAKODA hat zum Ziel, eine grosse Anzahl von Lernerkorpora des Deutschen in ein Repositorium zu integrieren und korpusübergreifend zu verknüpfen. Zu nennen wären hier das ZISA-Korpus (Clahsen et al., 1983), die FALKO-Korpusfamilie (Hirschmann et al., 2022), das MERLIN-Korpus (Wisniewski et al., 2013), das deutsch-ungarische Lernerkorpus DULKO (Beeh et al., 2021) oder auch die beiden Schweizer Korpora DiGS (Diehl et al., 2000) und SWIKO (Karges et al., 2022) – die Liste liesse sich fortsetzen. Was sind dabei die grössten Herausforderungen?

Katrin Wisniewski: Ein Mangel an Herausforderungen besteht nicht. Diese betreffen ganz unterschiedliche Ebenen im Projekt.

Zunächst einmal ist die Herausforderung, eine publizierbare Datenbasis zusammenzustellen, nicht zu unterschätzen. In DAKODA bitten wir ja Inhaber:innen öffentlicher und nicht öffentlicher L2-Korpora darum, uns ihre Daten für unser geplantes Dashboard zur Verfügung zu stellen. Das Dashboard soll eine Auswahl an Suchfunktionen anbieten. Zusätzlich möchten wir andererseits die Daten, wie von Ihnen schon angesprochen, in einem Repositorium bereitstellen und, falls rechtlich möglich, auf verschiedenen Wegen zum Download anbieten.

Nun mangelt es in keiner Weise an der Kooperationsbereitschaft von Kolleg:innen – ganz im Gegenteil! Hier sind wir nahezu ausnahmslos auf sehr positive Reaktionen gestoßen und unseren Kolleg:innen äußerst dankbar für ihre Offenheit. Allerdings ist die Publikation von Lernerkorpora an zahlreiche datenschutz- und lizenzrechtliche Bedingungen geknüpft. Wurden diese Faktoren bei der Korpuserstellung nicht schon mitbedacht und kein entsprechendes Einverständnis der Lernenden eingeholt und keine entsprechende Datenanonymisierung durchgeführt, können die Korpora tatsächlich der Forschungsgemeinschaft nicht mehr zur Verfügung gestellt werden. Bei älteren

Korpora haben wir außerdem einen nicht unerheblichen ‚Schwund' festgestellt: Einige Lernerkorpora sind Datenhacks zum Opfer gefallen, regalweise Transkripte wurden bei zu räumenden Büros vernichtet. Viele Korpusdaten wurden nie digitalisiert; in einigen publizierten Korpora fehlen Daten, oder sie sind nicht vollständig anonymisiert usw.

Eine weitere Herausforderung liegt darin, dass sehr viele Lernerkorpora des Deutschen in vielen verschiedenen, uneinheitlichen und nicht kompatiblen Formaten vorliegen und in unterschiedlicher Weise aufbereitet sind. Wenn sich zum Beispiel die Definition für ein Token oder einen Satz in Korpora unterscheidet und man dann Analysen, die auf solchen Einheiten beruhen, über die Korpora hinweg durchführt, muss das zu Folgefehlern führen. Deshalb bringen wir alle Lernerkorpora in ein einheitliches Format und führen bestimmte Vorverarbeitungsschritte neu und für alle Datensätze in gleicher Form durch.

Um ein Dashboard zu erstellen, das die Suche über verschiedene Lernerkorpora hinweg erlaubt, müssen zudem die darin enthaltenen Metadaten aufeinander abgestimmt sein. Metadaten sind «Daten über Daten», also beispielsweise Informationen über das Alter oder die L1 von Lernenden. Auch hier herrscht in der Lernerkorpuslandschaft eine große Vielfalt. Einige Korpora haben sehr wenige Metadaten, andere sehr viele. Auch ist die Definition selbst von scheinbar eindeutigen Variablen wie „L1" oft unterschiedlich. Wir arbeiten deshalb an einem großen gemeinsamen Metadatenschema für alle involvierten Korpora.

Die bisher genannten Herausforderungen betreffen sämtlich Fragen aus dem Bereich der Open Science. Immer häufiger wird – sinnvollerweise – von Fördergebern in Projekten ein Forschungsdatenmanagement verlangt, das die sogenannten FAIR-Prinzipien berücksichtigt. Forschungsdaten sollten demnach gut auffindbar (*findable*), zugänglich (*accessible*), kompatibel (*interoperable*) und nachhaltig bzw. wiederverwendbar (*reusable*) sein. Hier hat die Lernerkorpuslandschaft des Deutschen noch etliche Herausforderungen zu meistern. Ein Problem sehen wir in der geberseitigen Forderung (aber auch dem forschungsseitigen Wunsch) nach nachhaltigen Daten, ohne dass gleichzeitig eine dauerhafte Förderung bzw. eine stabile übergeordnete, institutionell angebundene Infrastruktur für Lernerkorpora besteht.

«Wir bringen alle Lernerkorpora in DAKODA in ein einheitliches Format und stimmen die Metadaten aufeinander ab. So ermöglichen wir korpusübergreifende Analysen mit großen Datenmengen.»

Die verschiedenen Korpora wurden auch generiert, um Daten für unterschiedliche Forschungsfragen zu liefern, und können demnach stark variieren. Wie geht das Projekt DAKODA damit um?

Katrin Wisniewski: Auf inhaltlicher Ebene geht es im Kern des DAKODA-Projekts darum, auszuprobieren, wie gut eine automatische Analyse von Verbstellungsphänomenen in Lernersprache funktioniert, genauer gesagt eine Analyse der sogenannten Erwerbsstufen, wie sie in der *Processability Theory* beschrieben wurden (Pienemann, 1998, 2005). Automatische Analysen von Lernersprache sind aber sehr anspruchsvoll, vor allem bei beginnenden Lernenden. Deshalb trägt DAKODA auch dezidiert explorativen Charakter: Eine „perfekte" Annotation ist computerbasiert nicht möglich. Dennoch interessiert es uns herauszufinden, welche Dinge mit dem Computer besser oder aber weniger gut funktionieren, auch um zukünftigen Entwicklungen hier vielleicht ein wenig den Weg zu bahnen.

Eine Voraussetzung für das Funktionieren (automatischer) Annotationen ist, dass man die interessierenden Phänomene (bei uns: Erwerbsstufen) sehr klar und detailliert linguistisch definiert. Mit Vagheit konfrontiert man automatische Analysesysteme nämlich besser nicht. Dabei ist uns aufgefallen, dass trotz der enormen Verbreitung der syntaktischen Erwerbsstufen teils noch nie genau beschrieben wurde, wie sich die Stufen eigentlich im Detail definieren. Von daher geht der automatisierten Annotation derzeit eine umfassende linguistische Arbeit voraus.


© Anke Steinberg/Universität Leipzig

Matthias Schwendemann ist wissenschaftlicher Mitarbeiter in den Bereichen Linguistik und Angewandte Linguistik am Herder-Institut der Universität Leipzig. Seine Arbeitsschwerpunkte in Forschung und Lehre liegen in den Bereichen Lexikologie, Wissenschaftssprache und Erwerb und Entwicklung des Deutschen als Fremd- und Zweitsprache sowie der Analyse von Lernersprache. Derzeit ist er zudem Mitarbeiter im BMBF-geförderten Drittmittelprojekt DAKODA.


© Swen Reichhold/Universität Leipzig

Katrin Wisniewski hat die Gerhard-Helbig-Professur für Deutsch als Fremd- und Zweitsprache am Herder-Institut der Universität Leipzig inne und forscht zu Grammatikerwerb und Sprachdiagnostik. Sie leitet das DAKODA-Projekt und war auch für das MERLIN-, das DISKO- und das MIKO-Korpus verantwortlich.

An wen werden sich dieses Repositorium und das von Ihnen beschriebene Dashboard hauptsächlich richten? Und für welche Verwendungszwecke kann man diese jeweils konsultieren?

Matthias Schwendemann: Kurz gesagt: Aus unserer Sicht wird bei DAKODA für alle etwas Interessantes dabei sein. Wir haben tatsächlich ein sehr breites und diverses Zielpublikum im Blick, also beispielsweise Spracherwerbsforscher:innen oder Wissenschaftler:innen, die sich grundsätzlich mit Lernersprache und/oder mit sprachlicher Variation auseinandersetzen, aber auch und nicht zuletzt Sprachdidaktiker:innen. Zusätzlich kann DAKODA aufgrund der für das Deutsche als L2 relativ großen Datenmenge auch für Computerlinguist:innen relevante Einsichten bereithalten. Eine weitere interessante Forschungsperspektive ergibt sich zudem dadurch, dass in DAKODA nicht nur Daten von L2-Lernenden, sondern auch Daten von Menschen mit Deutsch als Erstsprache zur Verfügung gestellt werden. Und zuletzt ist es natürlich so, dass die vielen Korpora aus DAKODA zu sehr unterschiedlichen Zwecken und vor dem Hintergrund sehr vielfältiger Erkenntnisinteressen erhoben wurden. Wer sich also bereits in der Vergangenheit aus bestimmten Gründen für spezifische Korpora interessiert hat, die jetzt auch in DAKODA zu finden sind, wird nun neue Zugangswege zu diesen Daten erhalten und vor allem auch die Möglichkeit, diese Daten mit anderen Lernerkorpusdaten in Verbindung zu setzen und zu vergleichen.

«Wir erhoffen uns, dass durch DAKODA bestimmte Forschungsfragen nochmal oder auch ganz neu an eine große Datenbasis gestellt werden können und so die empirische Wissensbasis der Zweitspracherwerbsforschung erweitert werden kann.»

Diese sehr vielfältige Ausrichtung zeigt sich in der auch von Katrin Wisniewski gerade angesprochenen Konzeption der verschiedenen DAKODA-Werkzeuge deutlich: dem Repositorium einerseits und dem interaktiven Dashboard andererseits. Einige der Funktionalitäten des Dashboards (z.B. ein Filtern der Korpusdaten nach sprachlichen Strukturen auf bestimmten Erwerbsstufen der Processability Theory) werden natürlich sehr eng an unsere konkreten Forschungsfragen angelehnt sein und daher vor allem für Spracherwerbsforscher:innen relevant sein, die sich für Erwerbsstufen im Bereich der Wortstellung interessieren. Gleichzeitig wird das Dashboard aber die Möglichkeit bieten, ganz unterschiedliche Korpora anhand ihrer Metadaten zu durchsuchen und so besser kennenzulernen und miteinander in Beziehung zu setzen. Die Aufbereitung der zahlreichen Metadaten in den verschiedenen Korpora mit dem Ziel einer größeren Einheitlichkeit ist gleichzeitig auch ein sehr umfangreiches Teilprojekt von DAKODA. Nutzer:innen des Dashboards wird es dann möglich sein, in den DAKODA-Daten etwa nach Lernenden mit ganz spezifischen Eigenschaften (Alter, Erstsprache(n), Sprachniveau, Sprachbiografie etc.) zu suchen und etwa die Texte dieser Lernenden vergleichend zu analysieren.

Was das Repositorium angeht, so erfüllt dieses aus unserer Perspektive mehrere sehr wichtige Funktionen. Auf einer grundsätzlichen Ebene sind Daten innerhalb eines Repositoriums öffentlich und damit prinzipiell zugänglich und in vielen Fällen auch downloadbar. Das DAKODA-Repositorium wird so interessierten Wissenschaftlern eine ganze Reihe an Möglichkeiten zur weiteren Arbeit mit den Korpusdaten geben. Gleichzeitig sorgt die Ablage von Daten in einem Repositorium im besten Fall dafür, dass Daten an mehreren Orten gesichert sind und so die Gefahr von Datenverlusten reduziert wird.

Letztlich erhoffen wir uns, dass durch DAKODA von vielen Wissenschaftler:innen aus unterschiedlichen Forschungsgebieten bestimmte Forschungsfragen nochmal oder auch ganz neu an eine große Datenbasis gestellt werden können und so die empirische Wissensbasis der Spracherwerbsforschung erweitert werden kann. Dies ist besonders auch deshalb der Fall, weil im Rahmen von DAKODA einige Korpora zum allerersten Mal überhaupt veröffentlicht und zur Verfügung

gestellt werden (z.B. das MULTILIT-Korpus (Schroeder/Schellhardt, 2015) oder auch das umfangreiche Chinesische Deutschlernerkorpus (Wu/Li, 2022)).

Werden auch nach Abschluss des Projekts neue Korpora in die Korpussammlung integriert?

Katrin Wisniewski: Wie oben schon gesagt, müssen Korpora je erst in ein einheitliches Format überführt werden, bevor sie in die DAKODA-Datenbasis aufgenommen werden können. Auch eine Übernahme in das Metadatenschema ist nötig. Dafür sind jedoch personelle Ressourcen nötig, die nach Projektende nicht mehr zur Verfügung stehen werden. So bedauerlich das ist: Dieser Situation ließe sich nur abhelfen, wenn es eine institutionell angebundene, übergreifende Lernerkorpus-Infrastrukturund eine langfristige strukturelle Förderung gäbe.

Welche Ratschläge würden Sie Forschenden geben, die vorhaben, ein Korpus zu erstellen?

Matthias Schwendemann: Hier hilft es vielleicht, zunächst rückwärts und vor allem *FAIR* (*Findability*, *Accessibility*, *Interoperability*, *Reusability*) zu denken. Die nützlichsten Korpusdaten für Wissenschaftler:innen und Sprachdidaktiker:innen sind Daten, die öffentlich und mit möglichst geringen (oder gar keinen) Zugangsvoraussetzungen auffindbar, verfügbar und zugänglich sind. Die erste Frage, die man sich bei der Konzeption eines Lernerkorpus stellen könnte, ist also: Wie müsste meine Einverständniserklärung und meine Datenschutzerklärung aussehen, damit ich die Daten später möglichst ohne Einschränkungen veröffentlichen kann? Hier sind einige grundlegende rechtliche Fragen zu beachten, die in den letzten Jahren enorm an Bedeutung gewonnen haben. Sehr wichtig ist hier vor allem die Tatsache, dass sowohl eine Datenschutzerklärung erhoben werden muss, die vor allem klärt, wie die persönlichen Daten der Lernenden bzw. Teilnehmenden geschützt werden, als auch eine Einverständniserklärung, in der es darum geht, wie die erhobenen Daten denn tatsächlich verwendet und vor allem veröffentlicht werden. Hier sind in jedem einzelnen Fall unter Umständen eine Reihe von spezifischen Fragen und Kontexten zu bedenken und wir würden allen raten, sich hier schon früh und proaktiv mit den Rechtsabteilungen bzw. den Justiziaten der eigenen Institutionen zu besprechen und gemeinsame Lösungen zu erarbeiten. Einverständniserklärungen sollten im besten Fall so abgefasst sein, dass sie eine Veröffentlichung unter freien Creative Commons-Lizenzen (CC) oder die Veröffentlichung mit möglichst geringen Einschränkungen und Zugriffshürden erlauben. Sofern die dann erhobenen Einverständnis- und Datenschutzerklärungen dies zulassen, sollte ab diesem Zeitpunkt ein Open-Access- und ein Open-Science-Ansatz verfolgt werden. Daten könnten dann beispielsweise auf Repositorien, wie zum Beispiel dem IRIS-Repositorium oder auch dem DAKODA-Repositorium, gespeichert, veröffentlicht und verfügbar gemacht werden, was die Nachnutzung durch andere Wissenschaftler:innen begünstigen würde.

Gleichzeitig, und jetzt springe ich nochmal ganz an den Anfang, sollten sich Wissenschaftler:innen bei der Erstellung von Korpora natürlich auch fragen, wie die erhobenen Daten die konkrete Forschungsfrage, für welche das jeweilige Korpus erhoben werden soll, beantworten können. Aber darüber hinaus sollte die Nachnutzbarkeit der erhobenen Daten direkt mitgedacht werden. Bereits erhobene Daten können dann nachhaltig von anderen Wissenschaftler:innen und in Lehr- und Lernkontexten genutzt werden, wenn zum Beispiel umfangreiche Metadaten erhoben wurden, die eine Vielzahl von potenziellen Forschungsfragen beantworten können. Metadaten sind gewissermaßen Daten über die im Korpus verfügbaren Sprachdaten. Dies können Informationen über die Erhebung des Korpus, über die Hintergründe und Eigenschaften der Lernenden oder auch über die Art und Weise der verwendeten Elizitierungsverfahren sein. Auch wenn bei der Erhebung von Metadaten

«Bei der Erhebung von Korpusdaten sollte die potenzielle Nachnutzbarkeit dieser Daten immer schon mitgedacht werden.»

natürlich Fragen der Datensparsamkeit und auch potenziell ethische Fragen zunächst handlungsleitend sein sollten, gibt es im Bereich der Lernerkorpusforschung immer stärkere Bemühungen, sich auf eine Reihe von grundlegenden Metadaten zu einigen, die in jedem Fall erhoben werden sollten (vgl. das Core Metadata Schema for Learner Corpora von Paquot et al., 2023). Vorlagen wie das Schema von Paquot et al. (2023) stellen dabei sehr hilfreiche Vorlagen bei der Erstellung eigener Korpora dar. Neben den Metadaten wäre ein weiterer vielversprechender Punkt, die Vergleichbarkeit mit anderen Lernerkorpora des Deutschen als L2 direkt mitzudenken. Dies könnten Wissenschaftler:innen etwa dadurch erreichen, dass ähnliche Aufgabenformate wie in bereits bestehenden Korpora zugrunde gelegt werden, oder auch dadurch, dass ähnliche Gruppen von Lernenden untersucht werden. Nicht zuletzt sind erstsprachliche Vergleichsgruppen, die beispielsweise dieselben Aufgabenstellungen bearbeiten, eine sehr hilfreiche Ergänzung, um Lernerkorpusdaten weiter anzureichern. Ob es für den jeweiligen Forschungskontext sinnvoll ist, L1-Gruppen einzubeziehen, sollte bei der Konzeption neuer Korpora direkt mitbedacht werden. In der Regel werden solche Vergleiche die vorhandenen Daten und die daraus ableitbaren Schlussfolgerungen allerdings deutlich bereichern.

Und nicht zuletzt möchten wir (natürlich mit einem kleinen Augenzwinkern) auch mal dazu motivieren, ab und zu groß zu denken und sich nicht von der vielen, vielen Arbeit, die das Erheben und Erstellen von Lernerkorpora bedeutet, einschüchtern oder gar abschrecken zu lassen. Tatsächlich rechtfertigen die fertigen Lernerkorpora wirklich alle Mühen auf dem Weg dorthin und werden uns als Fachcommunity in den nächsten Jahren dabei helfen, gemeinsam viele Fragen des Fremd- und Zweitspracherwerbs nochmal ganz neu zu stellen und hoffentlich auch beantworten zu können.

Was soll ihrer Meinung nach zum Thema ergänzt werden?

Katrin Wisniewski & Matthias Schwendemann: Um die oben genannten (und weitere) Herausforderungen der Lernerkorpusarbeit anzugehen, sind eine gute Vernetzung und weiterhin kooperative und interdisziplinäre Projekte nötig. Vor allem in der Zusammenarbeit mit der Computerlinguistik sehen wir großes Potenzial.

Die Arbeit mit Lernerkorpora ist für Praktiker:innen oft mit Hindernissen verbunden. An viele Korpora kommen etwa Lehrende gar nicht heran, und meist ist eine einigermaßen aufwändige Einarbeitung in die jeweilige Korpus-Suchsprache erforderlich. Wir hoffen, dass in Zukunft mehr niedrigschwellige Angebote zur Lernerkorpusnutzung entwickelt werden (vielleicht vergleichbar dem ZuMult-Projekt, Fandrych et al. 2023). Vielleicht könnten KI-Anwendungen hier spannende Entwicklungen ermöglichen.

«Ein fertiges Lernerkorpus rechtfertigt alle Mühen auf dem Weg dorthin.»

DAKODA – Datenkompetenzen für DaF/DaZ

DAKODA ist ein interdisziplinäres Projekt mit dem übergeordneten Ziel, die Datenkompetenzen des wissenschaftlichen DaF/DaZ-Nachwuchses im Bereich Lernerkorpusforschung voranzutreiben. Im Projekt werden sprachtechnologische Ressourcen für erwerbsbezogene Fragestellungen entwickelt und auf Basis einer breiten Datengrundlage erprobt. DAKODA soll somit Möglichkeiten und Grenzen der Anwendung computerlinguistischer Verfahren für Lernersprachenanalysen explorativ ausloten. Dazu kooperiert das Projektteam am Herder-Institut der Universität Leipzig (Leitung: Prof. Dr. Katrin Wisniewski) mit dem Language Technology Lab der FernUniversität in Hagen (Leitung: Prof. Dr.-Ing. Torsten Zesch).

Das Projekt wird von 10.2022 – 9.2025 in der Förderline «Datenkompetenzen des wissenschaftlichen Nachwuchs» vom BMBF gefördert (https://www.bildung-forschung.digital/digitalezukunft/de/wissen/Datenkompetenzen/datenkompetenzen_wissenschaftlichen_nachwuchs/datenkompetenzen_wiss_nachwuchs.html).

www.dakoda.org

DAKODA

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Finanziert von der Europäischen Union
NextGenerationEU

Literatur

Beeh, Christoph / Drewnowska-Vargáné, Ewa / Kappel, Péter / Modrián-Horváth, Bernadett / Nolda, Andreas / Rauzs, Orsolya / Scheibl, György (2021): *Dulko-Handbuch*. Szeged, Hungary: Institut für Germanistik der Universität Szeged. https://doi.org/10.14232/dulko-handbuch-v1.0.

Clahsen, Harald / Meisel, Jürgen / Pienemann, Manfred (1983): *Deutsch als Zweitsprache. Der Spracherwerb ausländischer Arbeiter*. Tübingen: Narr. http://doi.org/10.25592/uhhfdm.1463.

Diehl, Erika / Christen, Helen / Leuenberger, Sandra / Pelvat, Isabelle / Studer, Thérèse (2000): *Grammatikunterricht: Alles für der Katz? Untersuchungen zum Zweitsprachenerwerb Deutsch*. Tübingen: Niemeyer Reihe germanistische Linguistik, 220. https://www.unige.ch/lettres/alman/de/recherche/abgeschlossene-projekte/digs/digs-korpus [12.03.2024].

Fandrych, Christian / Schmidt, Thomas / Wallner, Franziska / Wörner, Kai (2023): Zugänge zu mündlichen Korpora für DaF und DaZ: Das ZuMult-Projekt. In: *Korpora Deutsch als Fremdsprache*, 3(1).

Hirschmann, Hagen / Lüdeling, Anke / Shadrova, Anna / Bobeck, Dominique / Klotz, Martin / Akbari, Roodabeh / Schneider, Sarah / Wan, Shujun (2022): FALKO. Eine Familie vielseitig annotierter Lernerkorpora des Deutschen als Fremdsprache. In: *Korpora Deutsch als Fremdsprache* 2(2), 139-148, https://doi.org/10.48694/kordaf.3552.

Karges, Katharina / Studer, Thomas / Hicks, Nina (2022): Lernersprache, Aufgabe und Modalität: Beobachtungen zu Texten aus dem Schweizer Lernerkorpus SWIKO. In: *Zeitschrift für germanistische Linguistik* 50: 1, 104–130. https://doi.org/10.1515/zgl-2022-2050.

Pienemann, Manfred (Hg.) (2005): *Cross-Linguistic Aspects of Processability Theory*. Amsterdam: John Benjamins.

Pienemann, Manfred (1998): *Language processing and second language development*. Amsterdam: John Benjamins.

Schroeder, Christoph / Schellhardt, Christin (2015): Nominalphrasen in deutschen und türkischen Texten mehrsprachiger SchülerInnen. In: Ziegler, Arne / Köpcke, Klaus-Michael (Hrsg.): *Deutsche Grammatik in Kontakt*. Berlin: De Gruyter, 241–262. https://doi.org/10.1515/9783110367171-011.

Wisniewski, Katrin / Schöne, Karin / Nicolas, Lionel / Vettori, Chiara / Boyd, Adriane / Meurers, Detmnar / Abel, Andrea / Hana, Jirka (2013): MERLIN. An Online Trilingual Learner Corpus Empirically Grounding the European Reference Levels in Authentic Learner Data. In: *ICT for Language Learning 2013. Conference Proceedings*. https://conference.pixel-online.net/conferences/ICT4LL2013/common/download/Paper_pdf/322-CEF03-FP-Wisniewski-ICT2013.pdf [12.03.2024].

Wu, Zekun / Li, Yuan (2022): Zur syntaktischen Komplexität des Schriftdeutschen chinesischer Deutschlerner/-innen – Eine korpusbasierte Profilanalyse. In: *Deutsch als Fremdsprache* 4. https://doi.org/10.37307/j.2198-2430.2022.04.04.

Tema

LEARNER CORPUS RESEARCH MEETS FOREIGN LANGUAGE EDUCATION: EXAMPLES FROM THE SWISS LEARNER CORPUS SWIKO

Korpuslinguistische Anwendungen in pädagogischen Kontexten sind bisher eher Wunsch als Wirklichkeit. Dieser Beitrag stellt das mehrsprachige Schweizer Lernerkorpus SWIKO vor und zeigt anhand von zwei Szenarien, wie dieses Korpus beim Sprachenlehren und -lernen genutzt werden kann. SWIKO umfasst schriftliche und mündliche task-basierte Texte von SchülerInnen der Sekundarstufe I in den Schulfremdsprachen (Deutsch, Englisch und Französisch) und in der jeweiligen Unterrichtssprache. Szenario 1 argumentiert anhand von Analysen der fremdsprachlichen Texte, dass Lernerleistungen bezogen auf Aufgaben-Typen illustriert werden sollten. Szenario 2 nutzt u.a. das Subkorpus der Unterrichtssprache, um authentische Lernmaterialien zu entwickeln. Als Veranschaulichung dienen Arbeitsblätter zur Negation im Deutschen.

Nina Hicks
| Université de Fribourg
Thomas Studer
| Université de Fribourg

Introduction

According to Römer (2008; see also e.g., Flinz, 2021), there are two ways in which corpora – large collections of oral or written texts (Lemnitzer & Zinsmeister, 2015; cf. introduction of this edition) – can be applied in pedagogical settings: either indirectly via researchers and material writers; or directly via learners and teachers in the classroom.

In the indirect form, corpora and corpus linguistic findings contribute to the development of reference works, teaching materials, and teacher training. The authentic and typical examples of actual language use inform decisions on what should be taught and in what order (McEnery & Xiao, 2011). Learner corpora have increasingly contributed to this process, offering valuable insights on the mechanisms and challenges of language learning, for example based on analyses of errors or contrastive over- and underuse (for an overview see e.g., Meunier, 2020). More recently, the influence of other variables such as tasks on differences in language use has attracted researchers' attention (e.g., Alexopoulou et al., 2017)

In the direct form, referred to as *data-driven learning* (DDL), teachers and learners use corpus linguistic techniques and tools for pedagogical purposes. In the last decades, corpus linguists have highlighted the advantages of the DDL approach, particularly the exposure to rich, authentic language use as well as the autonomous and collaborative discovery approach (e.g., Gilquin & Granger, 2010; McEnery & Xiao, 2011). In line with the slogan "Every student a Sherlock Holmes" (Johns, 1997, p. 101), students are encouraged to observe enhanced corpus data, and, based on what they notice, hypothesize, generate rules, and check their insights into linguistic patterns. This active and autonomous involvement also allegedly increases motivation.

Despite these proclaimed advantages, the DDL approach has yet to find its way

into the everyday classroom (Meunier, 2020). Several factors contribute to this scarce uptake (e.g., Meunier, 2019; Römer, 2009; Tribble, 2008): Technical tools have to be available and, more importantly, mastered – a time-consuming endeavour for both teachers and learners, especially considering that the topic and approach have to be carefully selected and adapted to the specific context. Furthermore, while research results have shown positive effects of DDL on learning gains (e.g., Boulton & Cobb, 2017), most studies were conducted by corpus linguists (as opposed to teachers) with advanced and tertiary learners focusing on English as a foreign language (although see recent work by Crosthwaite, 2020 for younger learners; or Vyatkina, 2020 for German as a foreign language).

With these gaps in mind, the next section provides a brief overview of the SWIKO project and corpus, followed by two scenarios of how the corpus can be utilized in two educational settings: teacher training and the secondary school classroom.

SWIKO

SWIKO is a multilingual corpus currently being developed at the Institute of Multilingualism in Fribourg (CH). Following the trend towards communication and, more recently, action-orientation in foreign language education in Switzerland (e.g., Bertschy et al., 2015; Studer, 2023), the project investigates what vocabulary and grammar look like under these leading principles in learner texts at the end of mandatory schooling.

Between 2017 and 2022, data was collected among 14- to 17-year old students attending grades 10-12 (HarmoS) in both German- and French-speaking Switzerland as well as a German-English bilingual school in Eastern Switzerland. Students completed eight communicative tasks which were systematically varied by intended text type, topic, and task structure (Ellis et al., 2020). The corpus currently contains over 3'000 annotated written and spoken, paper- and computer-based productions by Swiss lower secondary school students in three languages (German, French and English), both as their language of schooling and foreign languages.

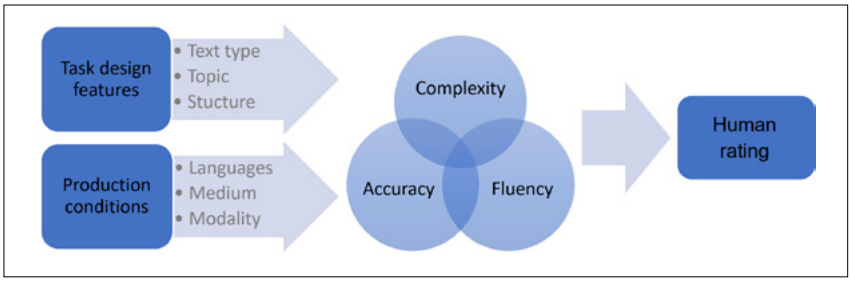


Figure 1
Overview of variables and analysis workflow in the SWIKO project.

The productions were then processed and analysed using corpus-linguistic methods. First, additional linguistic information was annotated (Schmid, 2013), such as the lemma (e.g., the tokens *gehe, gehst* und *ging* all stem from the lemma *gehen*) and part-of-speech (e.g., whether it's a verb or a noun). This allows us to analyse how long, diverse, dense, and sophisticated each learner production is (based on the CAF framework, Housen et al., 2012). Additionally, a *target hypothesis* – a grammatically and orthographically correct version (Lüdeling & Hirschmann, 2015) – was formulated for each learner text. By comparing the two versions, we can categorize the types of errors and, in extension, identify particularly challenging structures for learners. Finally, prospective foreign language teachers rated the productions based on the scales of the Common European Framework of Reference (Council of Europe, 2001, 2020). This enables us to examine which task design features and linguistic aspects correlate with human ratings of the learner texts (Figure 1).

In sum, grounded in Granger's (2015) Contrastive Interlanguage Analysis, we have been investigating how different tasks affect the linguistic properties of the resulting productions and how these relate to the human ratings (e.g., Karges et al., 2019, 2022; Studer & Hicks, 2022; Weiss et al., 2022). Based on our findings, the following chapters present two scenarios of how the SWIKO corpus can be used in educational settings. The first highlights some findings on how task design features affect the linguistic properties of the learner productions. The second discusses possibilities on how to use concordances from the German sub-corpus to develop teaching material for the secondary school classroom.

Nina Hicks works as a Research Assistant for the SWIKO/WETLAND project at the Institute of Multilingualism Fribourg. She holds an MA in Foreign Language Education from the University of Fribourg and has extensive experience teaching primary school students and German as a foreign language learners in Switzerland, California, and the UK.

Thomas Studer is a full Professor of German as a foreign and second language at the Department of Multilingualism and Foreign Language Education at the University of Fribourg and member of the board of directors of the Institute of Multilingualism in Fribourg. His research interests include corpus-based language acquisition, foreign language education, and language testing and assessment.

Showcasing task-based variation of learner language

To reflect the scope of language use that students encounter in their instructed foreign language classes, the SWIKO team developed eight different tasks (Ellis et al., 2020). As mentioned earlier, the task design features were systematically varied by intended text type, topic, and structure. As a result, students described and argued on more personal and more academic topics based on more and less restrictive input (e.g., answering short personal questions as opposed to writing a self-portrait without specific guidelines). Table 1 provides an overview of the tasks and variables.

ID	Description of task	Task type		Topic		Structure	
		des	arg	acad	pers	more	less
SWI01	Answer short personal questions	x		x		x	
SWI02	Describe graphs about pets in Switzerland	x			x	x	
SWI03	Discuss a list of vacation options		x	x		x	
SWI04	Discuss a list of the most important inventions		x		x	x	
SWI05	Create a self-portrait for a class exchange	x		x			x
SWI06	Present a topic (8 options)	x			x		x
SWI07	Discuss a later school start and finish		x	x			x
SWI08	Discuss replacement of foreign language classes with language exchange abroad		x		x		x

Table 1

Tasks and task variation in the SWIKO corpus. Task type was characterized as descriptive or argumentative, topic as academic or personal, and structure as more or less restrictive input given in the task prompt.

The systematic variation of task design features allows for insights on how the type of task affects the production. We analysed the linguistic aspects of the productions based on the CAF framework, which distinguishes three main components: "*complexity* is commonly characterized as the ability to use a wide and varied range of sophisticated structures and vocabulary in the L2, *accuracy* as the ability to produce target-like and error-free language, and *fluency* as the ability to produce the L2 with native-like[1] rapidity, pausing, hesitation, or reformulation" (Housen et al., 2012, p. 2, emphasis added).

In the next paragraphs, we present selected CAF differences based on two task design features – text type and topic – among 544 written German as a foreign language (DaF) productions in the SWIKO corpus. We believe that such insights can provide an illustrative data base for teacher training: On the one hand, it offers near-authentic examples of learner language, which can help trainees in developing a more realistic and nuanced understanding of students' output at the end of mandatory schooling. On the other hand, it can raise awareness of the importance of carefully selecting tasks for training and assessments as the type of task heavily affects the extent to which learners can expand on and demonstrate their linguistic competence.

Complexity

Among the written DaF productions in the SWIKO corpus, we observed that particularly the intended text type played an important role in regards to linguistic complexity: descriptive texts contained denser and more sophisticated vocabulary as well as more coordinated but fewer dependent clauses per sentence. In other words, students used more lexical words (particularly nouns) in their descriptions (mean noun ratio 0,32 vs. 0,18), which are often not found among the most frequent words of a language (mean logarithmic token frequency 4,63 vs. 4,31 based on SUBTLEX-DE, similar results based on Google Books 2000, Open Subtitles and DeReWo corpora). Furthermore, students were twice as inclined to link two or more main clauses with the co-ordinating conjunction *und* in their descriptive texts (mean coordinate phrases per clause 0,27 vs. 0,12). In contrast, when writing argumentative texts, they wrote longer sentences (mean sentence length in tokens 11,51 vs. 8,99) and used sub-ordinating conjunctions such as *weil*, *dass* or *wenn* three times as often to connect a main and a dependent clause (mean dependent clause ratio 0,32 vs. 0,11).

We exemplify these differences with two tasks which asked students to produce texts on an academic topic with more restrictive input, but which differ in terms of intended text type: a description of a graph about pets in Switzerland (SWI02) and a discussion of the allegedly most important inventions (SWI04). Both of the following texts were written by the same student (Figure 2).

The description of the graph is lexically denser and more sophisticated, as it contains more nouns (noun ratio 0,24 vs. 0,13) and fewer common words (mean SUBTLEX token log frequency 4,70 vs. 4,30). Conversely, the grammatical complexity is higher in the argumentation

1 Although the term *nativeness* has been debated, with scholars highlighting its vagueness (e.g., Cheng et al., 2021) and the ignorance of intra-individual variation among L1 groups (Shadrova et al., 2021). In SWIKO, we described our comparison group as *"language of schooling"*.

as sentences are longer (26 vs. 14 tokens per sentence), as well as more varied and complex. Specifically, the argumentative text contains five main clauses linked with the co-ordinating conjunctions *und* and *aber,* as well as four dependent clauses linked with the sub-ordinating conjunctions *dass* and *weil*. In comparison, the description consists of six main clauses exclusively linked with *und* and one dependent clause linked with *dass*.

Accuracy and Fluency

In order to investigate the accuracy of the written German productions of the SWIKO corpus in more details, a *target hypothesis* was formulated for each learner text. Comparing the two versions allows us to accurately trace the types of learner errors and develop a more refined picture of which structures proved particularly challenging for the learners. We measured accuracy in three aspects (Table 2): 1) the ratio of words written in a language other than German ("non-target" words); 2) the ratio of orthographic errors including capitalization, graphemes, and word boundary; and 3) the ratio of grammatical errors. This last aspect is again divided into four sub-categories: a) missing words, b) unnecessarily added words, and c) erroneously chosen or inflected words, all of which are measured at token level, as well as d) wrongly positioned constituents, which are measured as the ratio of (in)correct sentences. Table 2 provides an overview of the categories with examples.

In written productions, fluency is often equated with the length of the text, operationalized as the number of tokens[2] (Wolfe-Quintero et al., 1998).

Among written DaF productions in the SWIKO corpus, at the token level, the topic influenced both fluency and accuracy: in contrast to more academic topics, tasks on more personal topics generally resulted in longer (number of tokens 47 vs. 39) and more accurate texts (ratio of correct tokens 0,66 vs. 0,55). More specifically, they contained fewer errors both at the orthographic (error ratio 0,12 vs. 0,14) and grammatical level (error ratio 0,15 vs. 0,22), and more tokens in the target language (ratio of non-target tokens 0,06 vs. 0,10).

For example, when comparing two descriptive and less structured tasks,

SWI02: Graph about pets (descriptive) Student Ri211, 55 tokens		SWI04: Important inventions (argumentative) Student Ri211, 52 tokens
In diese graphique du hast 4 punkte[.] die ersten is über Wo hast ein tier und wir denken das eins out of zwei perssonen hast ein tier[.] punkte 2 wir haben 8.2 millionen tiere in der schweis[.] punkte 3 eine Katze catch zu viele other tiere und punkte 4 wir spend 800 millionen chf pro jahre fur die tiere in schweiss[.]		*ich denke das diese inventionen sind in die liste weil das sehr wichtig sind und die ordeung ist gut fur die fünf ersten aber icht fur die fünf lesten[.] ich denke das die car ist mehr wichtig als die brille und die ühr must gehen zent placen weil das sehr wichtig ist[.]*
0,24	noun ratio	0,13
4,30	mean SUBTLEX token log frequency	4,70
14	tokens per sentence	26
6 *(und)*	main clauses (*coordinating conjunctions*)	5 *(und, aber)*
1 *(dass)*	dependent clauses (*subordinating conjunctions*)	4 *(dass, weil)*

Figure 2

A descriptive and an argumentative text by the same student.

Category	Sub-category	Examples
Non-target	Language of schooling	*papillon (Schmetterling), ajutiere (hinzufügen)*
	Other languages	*glasses (Brille), o'clock (Uhr)*
Orthography	Capitalization	*computer, Acht, lieblingsTiere*
	Graphemes	*Personnen, Tire, bezucht, obwolh, helfbereit*
	Word boundary	*Lieblings Musik, Diewochenende, Liebst-du*
Grammar	Missing word	*In [der] Schweiz*
	Unnecessary word	*Es gibt 8,2 Mio. [die] Tiere*
	Wrongly inflected	*8,2 Million[en], du [hat], mit [deinem] Freunden*
	Wrongly chosen	*ich [habe] 14 Jahre alt, Ferien [im] Meer*
	Wrong position	*In der Schweiz [es gibt] viele Tiere.*

Table 2

Error categories with examples as analysed among the German productions in the SWIKO corpus.

2 Our token count excludes punctuation and numerals, and each entity is counted as one token irrespective of the number of words, e.g., *Titanic*, *Harry Potter and the Philosopher's Stone,* or *Nintendo Switch* are each considered a one entity-token.

students wrote longer and more accurate texts presenting themselves (SWI05, more personal) as opposed to presenting a topic such as languages or oceans (SWI06, more academic). Again, the following two texts were written by the same student (Figure 3). The more personal text is longer (67 vs. 47 tokens) and contains fewer non-target words (ratio 0,02 vs. 0,17), fewer orthographic errors (ratio 0,06 vs. 0,15) and fewer grammatical errors (ratio 0,12 vs. 0,28).

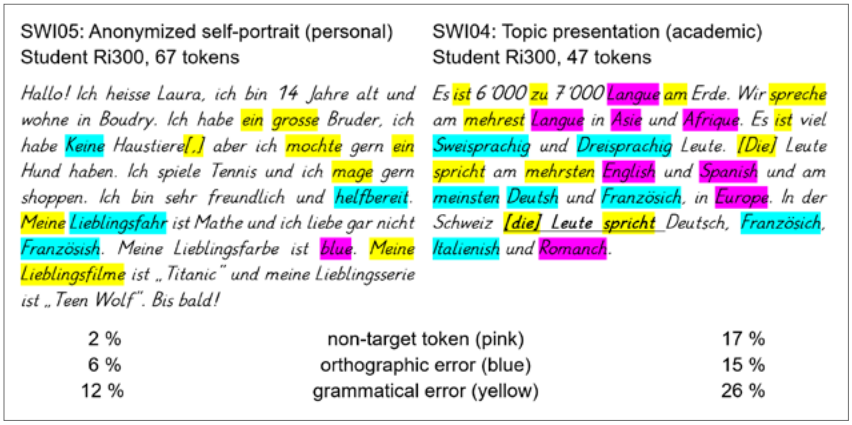


Figure 3

A text on a personal and a text on an academic topic by the same student.

Conversely, at the sentence level, the text type mattered most: sentences in descriptive texts were more often written in the correct order than sentences in argumentative texts (mean ratio of correct sentences 0,73 vs. 0,47; Figure 3 ratios 1,00 vs. 0,80). This could reflect the types of structures used: as reported in the complexity section, students wrote more subordinate phrases in argumentative productions, which require the more difficult verb final position and are therefore more prone to errors.

Negation in the DaF classroom

During our accuracy analysis (see section above), negation was revealed as a particularly challenging structure: German as a foreign language learners used *nicht* instead of *kein* in 54% of cases where *kein* was required, compared to just 2% among their German as the language of schooling peers. Moreover, it seems that learners apply the *"kein is followed by a noun"* rule correctly once they know it as it was never used erroneously; neither instead of *nicht* nor in the wrong position[3].

The following paragraphs therefore present ideas on how the German sub-corpora can be used in the German as a foreign language secondary school classroom to learn about negation. However, the underlying mechanisms of the task creation could be adapted to other topics such as capitalization or sub- and co-ordinating conjunctions (see Vyatkina, 2020 for examples on how to use the DWDS corpus in the DaF classroom).

DDL activities can be categorized on a continuum from teacher- to learner-led, and from relatively controlled such as gap exercises to open-ended, although more support might be beneficial for younger and less proficient learners (Gilquin & Granger, 2010). While inductive approaches are usually favoured in DDL, we aim to offer a wide spectrum of tasks and exercises as well as suggestions on alternatives for differentiation so that teachers can decide on whichever option best suits their class. Furthermore, our material (Übungsblatt Negation 1-5) is built around concordance lines extracted from SWIKOweb[4], i.e., lines of texts taken from the corpus and displayed with the highlighted key word in the middle. Other corpus data such as frequency lists or word clouds can be generated and used, for example to brainstorm vocabulary before a writing assignment (see Figure 4 for examples).

German as the language of schooling corpus

Concordances from the German as the language of schooling sub-corpus can be used in a variety of ways. As an introduction, these concordances can serve as an illustration from which learners derive rules – for example, *kein* is always followed by a noun, whereas *nicht* is often used in conjunction with adjectives or adverbs. We designed two difficulty levels, both of which follow an illustration – interaction – induction approach (Wicher, 2019) in which learners are encouraged to cooperate with their class peers. In an easier version, students derive rules from lists which were pre-sorted by the teacher according to the two rules (Übungsblatt Negation 1a). A more challenging version prompts students to group individual concordance lines by type and then derive the rules based on their observations (Übungsblatt

3 There were a few inflection errors (18%), e.g., *kein* instead of *keinen*. However, these only occurred slightly more frequently than among their peers with German as the language of schooling (8%).

4 Currently accessible via a personalized login.

Negation 1b). In either version, once the rules have been established, students can add further concordance lines from the corpus or their own examples.

In order to consolidate their knowledge, the students then create a mind map with collocations or chunks in which each of the two negation words commonly appear (Übungsblatt Negation 2). For scaffolding, the concordance lines from the introduction can be re-used (Übungsblatt Negation 1b), or additional concordance lines printed, or, if there is enough time, students could look up concordances themselves. Alternatively, a "class mind map" can be continuously and collectively expanded throughout the unit. In a following step, students complete gap exercises (Übungsblatt Negation 3), where they have to decide on which negation word is appropriate. This type of exercise can be created easily by omitting the highlighted *key word* of the concordance lines. A second exercise asks them to combine words from a list to common chunks, which was collected from corpus examples.

German as a foreign language corpus

Alternatively, concordances from the German as a foreign language sub-corpus can be used (Übungsblatt Negation 4): As a starting point, students reflect on how negation is formed in their language of schooling and any other languages they might know, before discussing their observations with their classmates and comparing them to the formation of negation in German. Then, they are asked to correct excerpts from the SWIKO corpus written by their German as a foreign language learning peers. In the exercise offered, students first have to decide whether a given concordance line actually does contain an error, and only correct it if necessary. Again, the level of difficulty can be adapted: In an easier version, a few concordance lines which all contain the same type of error can be offered. In a more challenging version, the students can additionally be asked to correct other types of errors in the concordance lines, such as orthographic or other types of grammatical errors. In either scenario, after investigating learner problems, it is recommended to highlight appropriate or correct use in follow-up exercises (Granger & Tribble, 1998).

und	112
ein	108
eine	84
Katze	79
ist	72
die	57
in	56
2	54
pro	54
für	49
1	47
gibt	39
400	39
hat	35
In	35
Hund	34

Figure 4

Excerpts from SWIKOweb based on SWI02 (graph about pets) in DaF texts: concordances (top), word cloud (bottom left), and frequency list (bottom right).

Application and transfer

Finally, in order to apply their knowledge, students can be prompted to write short texts. Out of the eight tasks in SWIKO, learners used negations most often when creating a self-portrait or discussing a list of the most important inventions, though other tasks might provide even more opportunities to use the target construction. Again, several scaffolding options are conceivable: The teacher can offer an exemplary text, or students can also be encouraged to use the mind map created earlier (Übungsblatt Negation 2). Furthermore, when creating the worksheets for the introduction and practice (Übungsblatt Negation 1-4), the teacher could select concordance lines only based on the same type of task, and these can then serve as additional templates.

Conclusion

Despite a large increase in corpus linguistic research studies, corpora have yet to find their way into the foreign language classroom. We aimed to bridge this gap by discussing two scenarios on how our corpus-linguistic research findings based on the rich and authentic Swiss Learner Corpus SWIKO can be used in foreign language education.

First, in teacher training, the productions can serve as an illustration of learners' abilities at the end of mandatory schooling, particularly in combination with our findings on task-based differences regarding the length, complexity, and accuracy of learner language. Second, our analysis of frequent errors in foreign language productions shed light on particularly challenging structures, while the language of schooling sub-corpus can serve as a peer-reference in the development of corresponding material. We exemplified this process focusing on negation in German, offering differentiated teaching material suitable for the secondary school classroom.

We hope that our contribution encourages readers take a leap and consider using learner corpora such as SWIKO in their classroom – whether as an authentic resource to illustrate task-based differences in learner productions or to introduce and consolidate a lexical or grammatical phenomenon through an autonomous and collaborative discovery approach.

References

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task Effects on Linguistic Complexity and Accuracy: A Large-Scale Learner Corpus Analysis Employing Natural Language Processing Techniques. *Language Learning*, *67*(S1), 180–208.

Bertschy, I., Cuenat, M. E., & Stotz, D. (2015). *Lehrplan Französisch und Englisch*. Passepartout - Fremdsprachen an der Volksschule.

Boulton, A., & Cobb, T. (2017). Corpus Use in Language Learning: A Meta-Analysis. *Language Learning*, *67*(2), 348–393.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment: companion volume*. Council of Europe Publishing.

Crosthwaite, P. (Ed.). (2020). *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners*. Routledge.

Ellis, R., Skehan, P., Li, S., Shintani, N., & Lambert, C. (2020). *Task-based language teaching: Theory and practice*. Cambridge University Press.

Flinz, Carolina (2021). Korpora in DaF und DaZ: Theorie und Praxis. *Zeitschrift für Interkulturellen Fremdsprachenunterricht*, 26(1), 1–43.

Gilquin, G., & Granger, S. (2010). How can data-driven learning be used in language teaching? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 359–370). Routledge.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, *1*(1), 7–24.

Granger, S. and C. Tribble (1998). Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning. In S. Granger (Ed.), *Learner English on Computer* (pp. 199-209). Routledge.

Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*. John Benjamins.

Johns, T. (1997). Contexts: The background, development and trialling of a concordance-based CALL program. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 100–115). Longman.

Karges, K., Studer, T., & Hicks, N. S. (2022). Lernersprache, Aufgabe und Modalität: Beobachtungen zu Texten aus dem Schweizer Lernerkorpus SWIKO. *Zeitschrift für germanistische Linguistik*, *50*(1), 104–130.

Karges, K., Studer, T., & Wiedenkeller, E. (2019). On the way to a new multilingual learner corpus of foreign language learning in school: Observations about task variation. In A. Abel, A. Glaznieks, V. Lyding, & L. Nicolas (Eds.), *Widening the Scope of Learner Corpus Research. Selected papers from the fourth Learner Corpus Research Conference* (pp. 137–165). Presses universitaires de Louvain.

Lemnitzer, L., & Zinsmeister, H. (2015). *Korpuslinguistik: Eine Einführung*. Narr Francke Attempto.

Lüdeling, A., & Hirschmann, H. (2015). Error annotation systems. In F. Meunier, G. Gilquin, & S. Granger (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 135–158). Cambridge University Press.

McEnery, T., & Xiao, R. (2011). What Corpora Can Offer in Language Teaching and Learning. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning* (pp. 364–380). Routledge.

Meunier, F. (2019). Data-Driven Learning: From Classroom Scaffolding to Sustainable Practices. *EL.LE*, *2*, 423–434.

Meunier, F. (2020). Introduction to learner corpus research. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge Handbook of Second Language Acquisition and Corpora* (pp. 23-36). Routledge.

Römer, U. (2008). Corpora and language teaching. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook* (pp. 112-130). Mouton de Gruyter.

Römer, U. (2009). Corpus research and practice: What help do teachers need and what can we offer? In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 83–98). John Benjamins.

Schmid, H. (2013). *TreeTagger—A Language Independent Part-of-speech Tagger* (3.2) [Computer software]. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger

Studer, T. (2023). Handlungsorientierung im Unterricht? Ja, klar. Aber was und wie genau eigentlich, warum und was kommt dabei heraus? In S. Demmig, S. Reitbrecht, B. Sorger, & H. Schweiger (Eds.), *IDT 2022: Beiträge zur Methodik und Didaktik von Deutsch als Fremd*Zweitsprache*. (pp. 19–29). Schmidt.

Studer, T., & Hicks, N. S. (2022). *The interplay of task variables, linguistic measures, and human ratings: Insights from the multilingual learner corpus SWIKO*. European Second Language Acquisition Conference, Fribourg.

Tribble, C. (2008). From Corpus to Classroom: Language Use and Language Teaching. *ELT Journal*, *62*(2), 213–216.

Vyatkina, N. (2020). *Incorporating Corpora: Using corpora to teach German to English-speaking learners* [Online instructional materials]. University of Kansas Open Language Resource Center. https://corpora.ku.edu/

Weiss, Z., Hicks, N. S., Meurers, D., & Studer, T. (2022). *Using linguistic complexity to probe into genre differences? Insights from the multilingual SWIKO learner corpus*. Learner Corpus Research Conference, Padua.

Wicher, O. (2019). Data-driven Learning in the Secondary Classroom: A Critical Evaluation from the Perspective of Foreign Language Didactics. In P. Crosthwaite (Ed.), *Data-driven Learning for the Next Generation: Corpora and DDL for Pre-tertiary Learners* (pp. 31–46). Routledge.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. University of Hawaii Press.

Übungsblätter

Download «Übungsblätter»:
https://ifm-swiko.unifr.ch/publications

Übungsblatt Negation 1a: Negation im Deutschen

Im Deutschen gibt es v.a. zwei Möglichkeiten, Sätze zu verneinen (Option 1 und 2). Wann wird welche Option gebraucht? Schau dir die Sätze an und versuche, eine Regel daraus abzuleiten. Tipp: Achte besonders auf das hervorgehobene Wort in der Mitte sowie das erste Wort nach dem hervorgehobenen Wort.

Option 1:

abe. Ich bin Vegetarier, dass heisst ich esse gar **kein** Fleisch. Meine Schwächen sind, manchmal nicht vie
en Grosseltern. Skiferien finde ich toll. Ich bin **kein** Fan von Jungendherbergen.Ja ich bin einverstanden
eren Spinnen hab ich angst, kleinere Spinnen sind **kein** Problem.Ich finde, dass die Elektrizität an erste
eine Mensch besitzt eine Katze der andere hat gar **kein** Tier. Es ist auch möglich das derjenige mit dem T
. Ich mag wenn ich die Frage zum Essen beantworte **keine** Pilze. Die Konsistänz und der Geschmack ist nicht
spiele ich Fussball oder Computerspiele. Ich habe **keine** Lieblingsmusik. Ich höre verschiedene Musikarten
t auf der Liste stehen, denn ohne den, hätten wir **keine** Bücher schreiben können. Das Fotoaparat ist auch

Option 2:

Aber was ich genau machen möchte, weiss ich noch **nicht** genau. Ich liebe Dessert, vorallem wenn das Desse
gefährlich sind und ekelhaft. Etwas was ich auch **nicht** gerne habe, ist wenn es in den Bergen sehr stark
anne. Mit dem Punkt Ausflüge in die Berge bin ich **nicht** einverstanden, weil ich mega gerne in die Berge f
n. Weil in den Bergen ist es immer sehr schön und **nicht** so viele Leute wie in Städten. Mit dem Punkt Städ
kt Städtereisen bin auch einverstanden, aber auch **nicht**, weil eine Städtereise zu machen ist auf einer Se
ndert haben. Denn die Welt ohne Elektrizität wäre **nicht** so cool und man hätte nicht so viele elektrische

Wie wird die Negation gebildet? Schreib die Regel und ein Beispiel dazu auf.

Regel 1: kein/e + _____ Beispiel: _____

Regel 2: nicht + _____ Beispiel: _____

Übungsblatt Negation 1b

Im Deutschen gibt es v.a. zwei Möglichkeiten, Sätze zu verneinen (Option 1 und 2). Wann wird welche Option gebraucht?

1) Sortiere die Streifen in zwei Gruppen und klebe sie auf.
2) Wie wird die Negation gebildet? Schreibe die Regeln und ein Beispiel dazu auf.

Regel 1: _____ + _____ Beispiel: _____

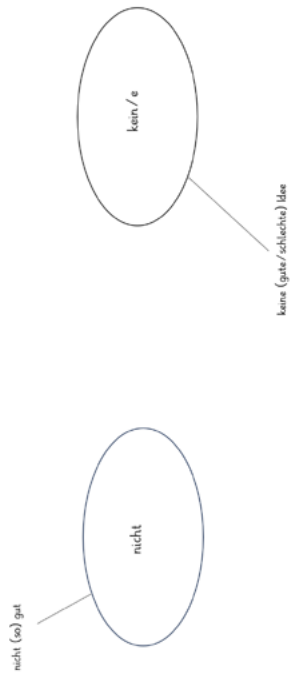
Regel 2: _____ + _____ Beispiel: _____

Material für Übungsblatt 1b ausdrucken und in Streifen schneiden

abe. Ich bin Vegetarier, dass heisst ich esse gar **kein** Fleisch. Meine Schwächen sind, manchmal nicht vie
en Grosseltern. Skiferien finde ich toll. Ich bin **kein** Fan von Jungendherbergen.Ja ich bin einverstanden
eren Spinnen hab ich angst, kleinere Spinnen sind **kein** Problem.Ich finde, dass die Elektrizität an erste
eine Mensch besitzt eine Katze der andere hat gar **kein** Tier. Es ist auch möglich das derjenige mit dem T
. Ich mag wenn ich die Frage zum Essen beantworte **keine** Pilze. Die Konsistänz und der Geschmack ist nicht
spiele ich Fussball oder Computerspiele. Ich habe **keine** Lieblingsmusik. Ich höre verschiedene Musikarten
t auf der Liste stehen, denn ohne den, hätten wir **keine** Bücher schreiben können. Das Fotoaparat ist auch
Aber was ich genau machen möchte, weiss ich noch **nicht** genau. Ich liebe Dessert, vorallem wenn das Desse
gefährlich sind und ekelhaft. Etwas was ich auch **nicht** gerne habe, ist wenn es in den Bergen sehr stark
anne. Mit dem Punkt Ausflüge in die Berge bin ich **nicht** einverstanden, weil ich mega gerne in die Berge f
n. Weil in den Bergen ist es immer sehr schön und **nicht** so viele Leute wie in Städten. Mit dem Punkt Städ
kt Städtereisen bin auch einverstanden, aber auch **nicht**, weil eine Städtereise zu machen ist auf einer Se
ndert haben. Denn die Welt ohne Elektrizität wäre **nicht** so cool und man hätte nicht so viele elektrische

Übungsblatt Negation 2: Kollokationen

Negationen kommen oft in typischen Wort-Verbindungen, so genannten *Kollokationen* vor.
Sammelt typische Kollokationen zu den zwei Begriffen *nicht* und *kein*:



Übungsblatt Negation 3: nicht oder kein/e?

1) Ergänze das fehlende Wort in der Lücke.

tan höre ich Klassik und Soundtracks. Ich bin mir _____ sicher, ob es wirklich Angst ist, aber es läuft m
eit. Das macht so viel Spass. Ich finde die Liste _____ umbedingt zutreffend, für mich. Ini Mit Reitferie
wohl man auch sagen könnte, dass es ohne Computer _____ Internet gäbe. Ausserdem ist die Brille viel zu n
st eine wichtige Erfindung, ohne das Rad würde es _____ Velos oder Autos und weitere Erfindungen geben. I
lte man meiner Meinung nach viel weniger oder gar _____ Fleisch mehr Essen, dass es gibt auch leckeres Es
n werden immer super. Reitferien sind sicher auch _____ schlecht. Ferien in der Schweiz finde ich auch go
h meistens einfach doppelt so viel Spass. Was ich _____ gerne mache, ist auf den Bauernhof die Ferien zu
Wanderferien fände ich ganz schrecklich. Ich bin _____ damit einverstanden, dass Städtereisen langweilig
ist ist mein Lieblingsdessert Fruchtsalat. Ich mag _____ Schlangen, da sie gefährlich sind und ekelhaft. K
en hingegen finde ich aber unangenehm, da sie oft _____ mehr Sauber sind. Mit dem Fusit Ferien mit Freude
Gleichgesinnten, man hat durch die Sprachbarriere _____ Freunde. Die Sprache wird zu einem Hassthema. Man
ann 2 Stunden Hausaufgaben machen, haben sie dann _____ Freilezit mehr. Es wäre aber gut, später in der Sc
rgang des Mittags ist gut. Aber so lange will je _____ Schüler in der Schule bleiben. Dann habe ich lieb

2) Bilde die Negation mit den vorgegebenen Wörtern und halte es in den Kasten fest.

die Idee - gut - sicher - die Zeit - mehr - die Lust - das Problem - genau
gerne - das Lieblingsessen - einverstanden - so toll - das Haustier

kein/e	nicht
keine (gute oder schlechte) Idee	nicht gut

Übungsblatt Negation 4: richtig oder falsch?

1) Wie wird die Negation in deiner Schulsprache gebildet? Und in anderen Sprachen, die du kennst?

Diskutiert eure Beobachtungen in der Klasse. Vergleicht sie dann mit der Negation im Deutschen.

2) Unten siehst du Textausschnitte von Sekundarschüler/innen, die Deutsch als Fremdsprache lernen. Korrigiere die Negationsfehler - aber Achtung! Nicht alle Textausschnitte enthalten Negations-Fehler. Entscheide zuerst, ob es richtig oder falsch ist und korrigiere nur falls nötig.

Beispiele:

Die Ferien in der Schweiz sind nicht gut weil es keine interessante Sachen hat. Die Ski ferien sind cool
keine
Schule. Und ich mag eat. Ich weiss es nicht. Ich nicht Lust zu ending die cours in 17 h 30.Ich wolte ni

neue Handy, weil mein Samsung ist kaput. Ich habe nicht ein Lieblingsbuch aber ich lese viele Buch und Ed
rgi. Ich mague volleyball am besten. Ich weiss es nicht. Mein Lieblings aktivité ist in die Shop gegangen
ot bauen oder mit Computer gut arbeiten. Ich habe kein Haustier aber an Sommer bekomme ich ein Rund. Ich
ant. 17h30 ist nicht gut für aktivitäten ich habe nicht Zeit. Die pause ist kurz. Ich habe nicht für wei
er Das ist mehr important als die Brille. Es gibt nicht Autos und es ist sehr important in die Planet. Ich
nein, weil wir lang aus us Hause sind. Wir haben nicht Freunden für spielen. diskutieren ect. Wir sind l
Bett weil man Haussaufgaben macht. Wir haben auch keine Lust für Aktivitäten mit freunden oder familie. D
, weil ist mit Freuds. Ich denke dass camping ist nicht cool, weil es is altmodish. Ich mag nicht pferdef
r ist super wir mussen hören Musik aber wir haben nicht elektricite wir massen nicht haben Computer nicht
ne zeit für essen Mittag. Die pose für mittag ist nicht eine gute Idee. Wir können Schlafen aber das ist

Tema

«Und dann isses aber trotzdem manchmal anders wie man spricht» – Verschmelzungsformen in der gesprochenen Wissenschaftssprache von Studierenden mit Deutsch als L1 und L2[1]

The article discusses the use and frequency of spoken language clitics in academic lectures among 121 students who speak German as their first language (L1) (n=25) or as a second or foreign language (L2) (n=96). The study reveals that students use a variety of clitics, and there are noticeable differences between students with German as L1 and L2, as well as between different academic contexts and data collection locations. The study's results also highlight the need for improvements in teaching German as an academic language, including those outlined in this text.

Matthias Schwendemann | Universität Leipzig
Franziska Wallner | Universität Leipzig

Einleitung

Während die Ratgeberliteratur in Bezug auf wissenschaftliche Vorträge und studentische Referate zwar grundsätzlich darauf hinweist, dass diese frei zu halten sind (vgl. u.a. Schäfer/Heinrich, 2010; Bayerlein, 2014), wird in Merkblättern zur Gestaltung von Vorträgen zumeist von der Verwendung umgangssprachlicher, eher mündlicher Formen abgeraten.[2] Allerdings zeigen empirische Analysen von deutschsprachigen wissenschaftlichen Vorträgen und studentischen Referaten, dass diese durch eine große Vielfalt an Mündlichkeitsphänomenen geprägt sind, die oft eher in der gesprochenen Alltagssprache verortet werden (vgl. Slavcheva/Meißner, 2014; v, 2017 und Schwendemann/Wallner, 2023). Vor diesem Hintergrund stellt sich die Frage, inwieweit sich Mündlichkeitsphänomene in studentischen Referaten von fortgeschrittenen Lernenden des Deutschen beobachten lassen. Anliegen der vorliegenden Studie war es, dies anhand des Gebrauchs von Klitisierungen zu überprüfen. Im Fokus standen dabei Verschmelzungen von mindestens zwei Wortformen (bspw. *gibts* [gibt es] oder *gehste* [gehst du]), die als noch nicht vollständig lexikalisiert gelten und als Besonderheit der gesprochenen Sprache angesehen werden (vgl. Duden, 2022, 551-555). Für die Analyse wurden studentische Vorträge aus dem GeWiss-Korpus (Gesprochene Wissenschaftssprache kontrastiv; Wallner, 2023) als Datengrundlage genutzt. GeWiss ist ein Vergleichskorpus der gesprochenen Wissenschaftssprache, das studentische Referate, Expertenvorträge und Prüfungsgespräche umfasst. Neben deutschsprachigen Daten sind im Korpus auch Daten auf Englisch, Polnisch und Italienisch enthalten. Die Daten stammen zum einen von Sprecher:innen, die die jeweiligen Sprachen als Erstsprache (L1) sprechen, zum anderen liegen für das Deutsche und das Englische auch L2-Daten vor (vgl. hierzu ausführlich Fandrych/Wallner, 2022). Die studentischen Referate von Sprecher:innen mit Deutsch als L2 wurden in verschiedenen akademischen Kontexten erhoben (darunter Bulgarien,

1 Wir möchten uns bei den beiden anonymen Gutachter:innen für ihre hilfreichen Kommentare bedanken, insbesondere hinsichtlich der didaktischen Perspektiven unserer Ergebnisse. Ihre Rückmeldungen haben wesentlich dazu beigetragen, den vorliegenden Beitrag zu schärfen und zu verbessern.

2 Diese Beobachtung basiert auf einer explorativen Analyse von insgesamt 20 Merkblättern und Webseiten mit Informationen zur Gestaltung von studentischen Referaten, die über die Homepages verschiedener germanistischer Institute deutschlandweit zugänglich sind.

Gruppe	Anzahl Sprecher:innen	Token
Deutsch als L1 in Deutschland	25	39.185
Deutsch als L2 in Deutschland	20	26.135
Deutsch als L2 in Bulgarien	19	29.768
Deutsch als L2 in Finnland	20	13.313
Deutsch als L2 in Polen	14	28.616
Deutsch als L2 in Großbritannien	23	33.814
Gesamt	**121**	**170.831**

Tabelle 1
Überblick über die studentischen Vorträge nach Erhebungsort

trans	(.)	und	dann	isses	aber	trotzdem	manchmal	anders	wie	man
norm		und	dann	ist es	aber	trotzdem	manchmal	anders	wie	man
lemma		und	dann	sein es	aber	trotzdem	manchmal	anders	wie	man
pos		KON	ADV	VAFIN PPER	PTKMA	ADV	ADV	ADV	PWAV	PIS

Abbildung 1
Anzeige der Annotationen im Transkriptbrowser ZuViel (Schmidt et al., 2023), GWSS_E_00001_SE_01

Deutschland, Finnland, Großbritannien und Polen). Aus diesem Grund eignet sich das GeWiss-Korpus als Ressource für eine vergleichende empirische Untersuchung von Mündlichkeitsphänomenen in der gesprochenen Wissenschaftssprache von fortgeschrittenen Lernenden.

Im Folgenden wird zunächst das methodische Vorgehen der Untersuchung erläutert. Im nächsten Schritt werden die Ergebnisse vorgestellt und anschließend diskutiert. Der Beitrag schließt mit einem didaktischen Ausblick und einem Fazit.

Methodisches Vorgehen

Die Grundlage für die Untersuchung bildeten 87 studentische Einzel- und Gruppenvorträge aus dem GeWiss-Korpus mit insgesamt 121 Sprecher:innen. Die Mehrheit der Sprecher:innen (insgesamt 96) sind fortgeschrittene Lernende des Deutschen. Für die übrigen 25 Sprecher:innen ist Deutsch die Erstsprache. Insgesamt umfasst die Datengrundlage 170.831 Token. Tabelle 1 gibt einen Überblick über die Anzahl der Sprecher:innen und Token sortiert nach Erhebungsort.

Die Daten im Gewiss-Korpus liegen als Audio und in Form einer aussprachenahen Transkription vor. Zudem erfolgten weitere korpuslinguistische Aufbereitungsschritte. Dazu zählt eine orthografische Normalisierung, bei der jedem transkribierten Token eine standardorthografische Entsprechung zugeordnet wird, sowie die Annotation von Wortarten (POS-Tagging) und die Lemmatisierung nach dem STTS 2.0.[3] Diese Aufbereitungsschritte ermöglichen eine systematische Erfassung von ausgewählten Mündlichkeitsphänomenen auf Tokenebene wie etwa Modalpartikeln, für die im STTS 2.0 das Kürzel PTKMA vergeben wird (vgl. Abbildung 1). Auch Klitisierungen lassen sich durch diese Aufbereitung leichter identifizieren, da in diesem Fall einem Token mehrere POS-Tags zugeordnet werden (vgl. *isses* in Abbildung 1 mit den POS-Tags VAFIN PPER (finites Hilfsverb + Personalpronomen)). Abbildung 1 zeigt anhand eines Ausschnittes aus einem Prüfungsgespräch die korpuslinguistische Aufbereitung der Sprachdaten. Die erste Zeile (trans) umfasst die aussprachenahe Transkription, die zweite (norm) die orthografisch normalisierte Fassung, die dritte (lemma) die Lemmatisierung (also die jeweilige Grundform der sprachlichen Einheiten) und die vierte (pos) die zugewiesenen Wortartkategorien gemäß dem STTS 2.0.

Die Ermittlung des Gebrauchs der Klitisierungen erfolgte mit Hilfe des Tools ZuRecht (Frick/Helmer/Wallner, 2023).[4] Dabei wurde eine sprecherbezogene Perspektive eingenommen. Das heißt, es wurde für alle Vortragenden

3 Beim STTS 2.0 handelt es sich um eine Erweiterung des ursprünglichen, Stuttgart-Tübingen-Tag-Set (STTS), zur Vergabe von Part-of-Speech-Tags (POS). Im STTS 2.0 werden im Vergleich zum STTS gesprochensprachliche Phänomene in das Tagging miteinbezogen (vgl. Westpfahl et al., 2017).

4 ZuRecht ist ein Werkzeug, das im ZuMult-Projekt (Fandrych et al., 2023) entwickelt wurde. Ziel des ZuMult-Projektes war es u.a., niedrigschwellige Zugangswege zu Korpora der gesprochenen Sprache für Didaktiker:innen zu schaffen, die geringere technische Vorkenntnisse voraussetzen als bisherige Zugriffsmöglichkeiten. Weitere Informationen finden sich unter: zumult.org [14.05.2024].

Matthias Schwendemann ist wissenschaftlicher Mitarbeiter in den Bereichen Linguistik und Angewandte Linguistik am Herder-Institut der Universität Leipzig. Seine Arbeitsschwerpunkte in Forschung und Lehre liegen in den Bereichen Lexikologie, Wissenschaftssprache und Erwerb und Entwicklung des Deutschen als Fremd- und Zweitsprache sowie der Analyse von Lernersprache. Derzeit ist er zudem Mitarbeiter im BMBF-geförderten Drittmittelprojekt DAKODA.

Franziska Wallner ist wissenschaftliche Mitarbeiterin am Herder-Institut der Universität Leipzig. Ihre Forschungsschwerpunkte sind unter anderen das Deutsche als fremde Bildungs- und Wissenschaftssprache, die korpusbasierte Erforschung der gesprochenen Sprache, Mündlichkeitsdidaktik sowie die Nutzung von Korpora im Kontext von Deutsch als Fremd- und Zweitsprache.

Klitisierung	Vorkommen als Klitisierung		Vorkommen standardnahe Realisierungen		Vorkommen weiterer analytischer Formen		Vorkommen gesamt
isses / is_s	35	(76 %)	4 *(ist es)*	9 %	7 *(is es / is s)*	15 %	46
gibts / gibs	35	64 %	19 *(gibt es)*	35 %	1 *(gibt s)*	1,5 %	55
son	24	71 %	0 *(so ein / einen)*	0 %	10 *(so n)*	29 %	34
gehts	14	70 %	5 *(geht es)*	25 %	1 *(geht s)*	5 %	20
mans	10	71 %	4 *(man es)*	29 %	0	0 %	14

Tabelle 2

Die fünf häufigsten Klitisierungen in den studentischen Vorträgen mit Deutsch als L1 (sprecherbezogen = nur die Vortragenden)

neben der Gesamtmenge an produzierten Token auch die Anzahl an jeweils realisierten Klitisierungen ermittelt. Als Klitisierung wurden dabei alle als miteinander verschmolzen transkribierten sprachlichen Einheiten gezählt. Neben Token, denen zwei POS-Tags zugeordnet wurden (wie im Fall von *isses*), fanden dabei auch sämtliche als assimiliert transkribierten sprachlichen Einheiten (bspw. *gibt_s* [gibt es], *so_n* [so ein]) Berücksichtigung. Bereits im ersten Schritt aus der Suchanfrage ausgeschlossen wurden sämtliche als Nichtwörter (XY, etwa nicht interpretierbare Einzelbuchstaben), als Wortabbrüche (AB, bspw. „[...] *wo die unterschiede in der* sch *(0.2) also in der aussprache* [...] *liegen*“ GWSS_E_00001_SE_01), als Häsitationen (NGHES, „[...] *und* ***ähm*** *°h also dieses modell bietet einfach (.) also s is sehr detailliert (.)* [...]“ GWSS_E_00001_SE_01) und als Eigennamen (NE) getaggte Einheiten. Alle 668 automatisch auf diese Weise identifizierten potenziellen Klitisierungen durchliefen dann eine doppelte manuelle Einstufung, um mögliche Fehltreffer auszuschliessen. Für die manuelle Überprüfung wurde mit Hilfe von Krippendorf's Alpha eine Interraterübereinstimmung berechnet. Diese lag in einem sehr hohen Bereich (Krippendorf's Alpha = 0,964). Auf diese Weise wurden insgesamt 660 Klitisierungen identifiziert, die im Folgenden in die Datenanalysen eingehen. Für alle Sprecher:innen wurden die produzierten Klitisierungen mit der Gesamtzahl der jeweils produzierten Token ins Verhältnis gesetzt, um so die Anzahl der Klitisierungen nach Sprecher:innen miteinander vergleichbar zu machen. Daraufhin wurden die Vortragenden nach Erhebungsort gruppiert. Für die in Deutschland erhobenen Daten wurde also zusätzlich noch zwischen L1- und L2-Sprecher:innen differenziert.

In einem ersten Schritt wurden nun die häufigsten Klitisierungen in den in Deutschland erhobenen L1-Vorträgen bestimmt und mit Vorkommen korrespondierender analytischer Formen verglichen (etwa *isses* vs. *ist es*). Zudem wurde geprüft, inwieweit die häufigsten Klitisierungen auch in den L2-Vorträgen nachgewiesen werden können.

In einem zweiten Schritt wurden dann die Sprecher:innengruppen der verschiedenen Erhebungsorte miteinander verglichen. Aufgrund der geringen Gruppengrößen wurde zu diesem Zweck der nichtparametrische Kruskal-Wallis-Tests eingesetzt. Im folgenden Abschnitt werden die Ergebnisse dieser Gegenüberstellung präsentiert.

Ergebnisse

Tabelle 2 zeigt die fünf häufigsten Verschmelzungsformen, die für die in Deutschland erhobenen Vorträge von Studierenden mit Deutsch als Erstsprache ermittelt werden konnten.[5] Neben der Anzahl der Vorkommen als Klitisierung werden in Tabelle 2 auch die Vorkommenshäufigkeit für die jeweilige standardnahe Form sowie für weitere analytischen Formen angegeben. Dabei lässt sich beobachten, dass die klitisierten Formen deutlich häufiger auftreten als die standardnahen und/oder weitere analytische Formen. Studierende verwenden also eher *isses* als *ist es* oder *son* als *so ein / einen* in ihren wissenschaftlichen Vorträgen.

Die absoluten Frequenzen bezüglich der fünf häufigsten Verschmelzungsformen bei Studierenden mit Deutsch als L1 in Deutschland im Vergleich zu den übrigen Erhebungsgruppen und -orten verdeutlichen zudem weitere Tendenzen (vgl. Tabelle 3). Besonders auffällig ist, dass bestimmte Formen fast ausschließlich von Studierenden mit Deutsch als L1 produziert werden: *son* und *mans*. Ausserdem wird deutlich, dass neben den in Deutschland erhobenen L2-Daten nur die Daten aus Großbritannien etwas mehr Verschmelzungsformen aufweisen, diese sich aber auf *isses* und *gibts* beschränken.

Die deskriptive Statistik (siehe Tabelle 4)

5 Die Abfragecodes, die in ZuRecht zur Ermittlung der Klitisierungen verwendet wurden, können auf Anfrage gern zur Verfügung gestellt und dann für eigene Suchanfragen angepasst werden.

und der Violinplot (siehe Abbildung 2) zeigen allerdings, dass sich die Verteilung über die unterschiedlichen Erhebungskontexte und vor dem Hintergrund, ob Deutsch als L1 oder L2 gesprochen wird, deutlich unterscheidet.

Sowohl die deskriptiven Daten als auch die Violinplots deuten darauf hin, dass Studierende mit Deutsch als L1 insgesamt frequenter Klitisierungen verwenden. Gleichzeitig scheinen aber erhebliche Unterschiede zwischen den einzelnen Sprecher:innen zu bestehen, was anhand der hohen Standardabweichung erkennbar ist. Eine ähnlich hohe Streuung der Daten lässt sich sonst nur noch in den L2-Daten aus Deutschland und den Daten aus Großbritannien beobachten, die ebenfalls vergleichsweise hohe Standardabweichungen aufweisen. Die Vorträge aus den übrigen Erhebungskontexten (Bulgarien, Finnland und Polen) scheinen sich stärker untereinander zu ähneln und enthalten insgesamt deutlich weniger realisierte Klitisierungen.

Der nichtparametrische Kruskal-Wallis-Test, der durchgeführt wurde, um Unterschiede zwischen den Gruppen zu berechnen, ist mit einer moderaten Effektstärke ($\eta 2 = 0{,}47$) signifikant ($\chi 2\,(5) = 59{,}036$; $p < 0{,}001$) und zeigt, dass zwischen den einzelnen Erhebungsorten Unterschiede in Bezug auf die Anzahl an realisierten Klitisierungen bestehen. Für mehrfache Vergleiche angepasste Dunn-Bonferroni-Post-hoc-Analysen zeigen zudem, dass sich die Vorträge von Sprechenden mit Deutsch als L1 hinsichtlich der Vorkommen von Klitisierungen signifikant von allen anderen Sprecher:innengruppen in den übrigen Erhebungskontexten unterscheiden (vgl. Tabelle 5). Die bereits oben angesprochenen Tendenzen bezüglich der in Deutschland erhobenen L2-Daten sowie der Daten aus Großbritannien erreichen allerdings nicht das Signifikanzniveau ($\alpha < 0.05$). In Tabelle 5 werden außerdem die Effektstärken der einzelnen Vergleiche mit dem Korrelationskoeffizienten *r* angegeben. Diese liefern zumindest eine Bestätigung der herausgearbeiteten Tendenzen, wobei die Vergleiche zwischen den Vorträgen von Studierenden mit Deutsch als L2 in Deutschland und Großbritannien zu den Vorträgen in Bulgarien, Finnland und Polen jeweils moderate Effektstärken zeigen.

Klitisierung	L1 Deutschland	L2 Deutschland	L2 Bulgarien	L2 Finnland	L2 Polen	L2 Großbritannien
isses / is_s [ist es]	35	8	1	-	4	12
gibts / gibs [gibt es]	35	15	3	3	2	16
son / so_n [so ein / einen]	24	1	-	-	1	-
gehts [geht es]	14	3	2	-	-	-
mans [man es]	10	-	-	-	-	-

Tabelle 3

Überblick über die Frequenz der fünf häufigsten Formen in Vorträgen von Studierenden mit Deutsch als L1 im Vergleich zu den anderen Erhebungskontexten

Deskriptive Statistik						
	Anteil Klitisierungen pro 100 Token					
	Deutsch L1 Deutschland	Deutsch L2 Deutschland	Deutsch L2 Bulgarien	Deutsch L2 Finnland	Deutsch L2 Polen	Deutsch L2 Großbritannien
Sprecher:innen	25	20	19	20	14	23
Median	0.767	0.207	0.000	0.000	0.000	0.124
Mean	0.811	0.285	0.046	0.068	0.070	0.271
Standardabweichung	0.393	0.303	0.089	0.130	0.102	0.330
Varianz	0.154	0.092	0.008	0.017	0.010	0.109

Tabelle 4

Deskriptive Statistik zum Anteil der Klitisierungen in allen Erhebungskontexten

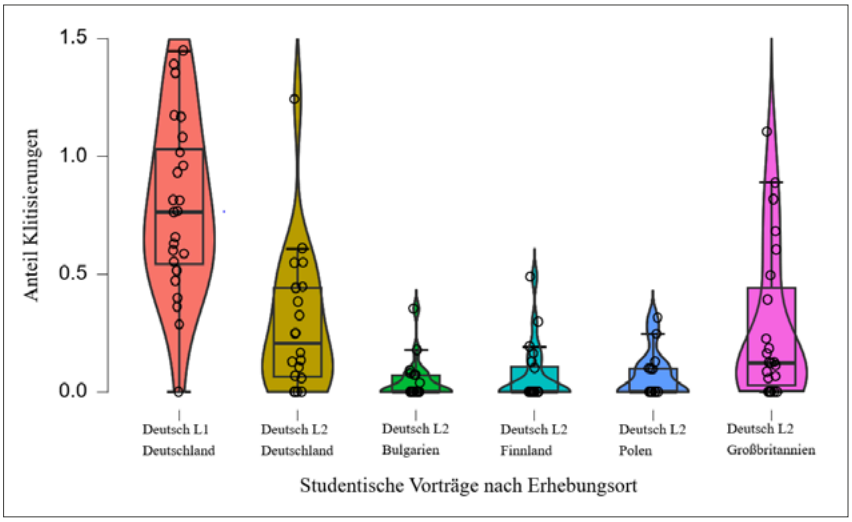


Abbildung 2

Violinplot mit der Verteilung der Klitisierungen in den verschiedenen Erhebungskontexten

Diskussion

Die Ergebnisse zeigen, dass sich in studentischen Vorträgen eine beachtliche Anzahl an Klitisierungen nachweisen lässt, die üblicherweise eher in alltagssprachlichen Kommunikationssituationen verortet werden. Gleichzeitig wird deutlich, dass studentische Vorträge von Sprecher:innen mit Deutsch als L1 durch eine signifikant höhere Zahl an Klitisierungen geprägt sind als Vorträge von Studierenden mit Deutsch als L2. Allerdings zeigen sowohl die Vorträge von Studierenden mit Deutsch als L2, die in Deutschland erhoben wurden, als auch die Vorträge der Studierenden aus Großbritannien eine deutliche Tendenz zur frequenteren Realisierung von Verschmelzungsformen. Im Falle der in Deutschland erhobenen Daten könnten hier etwa Einflüsse aus der Umgebungssprache angenommen werden. Die Studierenden mit Deutsch als L1 geben Input, der direkt aufgenommen werden kann. Im Falle der Daten aus Großbritannien könnte vielleicht die Sozialisation in der anglophonen Wissenschaftskultur eine Rolle spielen, die ebenfalls durch eine größere Nähe zu alltagssprachlichen und gesprochensprachlichen Registern und durch große sprachliche Variation innerhalb der gesprochenen Wissenschaftssprache gekennzeichnet ist (vgl. Biber et al., 2002). Einen besonders eindrücklichen Hinweis auf Einflüsse aus der alltagssprachlichen Mündlichkeit auf die gesprochene Wissenschaftssprache liefert das Auftreten des sich seit einiger Zeit im Sprachgebrauch etablierenden Artikels *son/sone* (‚so einen/e') (vgl. Duden, 2022:554f.; vgl. Fandrych, 2022: 64-66). Dieser neue Artikel kann dabei funktional sowohl deiktische Funktionen übernehmen als auch als Ersatz für andere indefinite Artikelformen dienen. In der vorliegenden Studie konnte der Artikel *son/sone* fast ausschließlich in den Daten nachgewiesen werden, die in Deutschland erhoben wurden und hier noch einmal deutlich frequenter in den Daten der erstsprachlichen Sprecher:innen. Das könnte ein Hinweis darauf sein, dass es sich um eine Struktur handelt, die relativ direkt aus der gesprochenen Alltagssprache in die gesprochene Wissenschaftssprache übertragen wird. Hier sind allerdings zwei Einschränkungen der vorliegenden Studie zu nennen: Die Datengrundlage der Analysen ist begrenzt. Dies betrifft einerseits die geringen Gruppengrößen, andererseits aber auch die Anzahl der von den einzelnen Sprecher:innen realisierten Token. Zudem ist zu berücksichtigen, dass auch die Inhalte der Vorträge und das gesamte Setting Einfluss auf die Auswahl der sprachlichen Mittel haben kann. Wenn Sprecher:innen keine oder nur wenige Klitisierungen produzieren, heißt dies folglich nicht unbedingt, dass Klitisierungen nicht zu ihrem sprachlichen Repertoire zählen. Die Ergebnisse können daher nicht ohne weiteres verallgemeinert werden. Überdies liegen die Erhebungen der Vorträge schon etwas mehr als zehn Jahre zurück. Gerade die Tendenz zur Verwendung von Formen wie *son/sone* könnte sich seitdem deutlich verstärkt haben.

Post hoc-Dunn-Bonferroni-Vergleich anhand von Erhebungsort und Deutsch als L1/L2	z	Wi	Wj	p	p_{bonf}	r
DeuL1_Deutschland - DeuL2_Deutschland	3.233	101.080	67.950	< 0.001 ***	0.018 **	0.482
DeuL1_Deutschland - DeuL2_Bulgarien	6.239	101.080	36.211	< 0.001 ***	< 0.001 ***	0.941
DeuL1_Deutschland - DeuL2_Finnland	6.148	101.080	38.075	< 0.001 ***	< 0.001 ***	0.917
DeuL1_Deutschland - DeuL2_Polen	5.237	101.080	41.357	< 0.001 ***	< 0.001 ***	0.839
DeuL1_Deutschland - DeuL2_Großbritannien	3.781	101.080	63.761	< 0.001 ***	0.002 **	0.546
DeuL2_Deutschland - DeuL2_Bulgarien	2.900	67.950	36.211	0.004 **	0.056	0.464
DeuL2_Deutschland - DeuL2_Finnland	2.766	67.950	38.075	0.006 **	0.085	0.437
DeuL2_Deutschland - DeuL2_Polen	2.234	67.950	41.357	0.025 *	0.382	0.383
DeuL2_Deutschland - DeuL2_Großbritannien	0.401	67.950	63.761	0.688	1.000	0.061
DeuL2_Bulgarien - DeuL2_Finnland	-0.170	36.211	38.075	0.865	1.000	-0.027
DeuL2_Bulgarien - DeuL2_Polen	-0.428	36.211	41.357	0.669	1.000	-0.075
DeuL2_Bulgarien - DeuL2_Großbritannien	-2.602	36.211	63.761	0.009 **	0.139	-0.402
DeuL2_Finnland - DeuL2_Polen	-0.276	38.075	41.357	0.783	1.000	-0.047
DeuL2_Finnland - DeuL2_Großbritannien	-2.459	38.075	63.761	0.014 *	0.209	-0.375
DeuL2_Polen - DeuL2_Großbritannien	-1.935	41.357	63.761	0.053	0.795	-0.318

Tabelle 5

Übersicht über die einzelnen Vergleiche der Post hoc-Dunn-Bonferroni-Analyse * p < .05, ** p < .01, *** p < .001

Didaktischer Ausblick und Fazit

Die Ergebnisse verdeutlichen, dass Studierende mit Deutsch als L1 und als L2 zunächst dafür sensibilisiert werden sollten, dass bestimmte gesprochensprachliche Phänomene der Alltagssprache durchaus als Input in der gesprochenen Wissenschaftssprache erwartbar sind. In einem nächsten Schritt müssten Studierenden darauf aufbauend potenzielle Verschmelzungsformen in der gesprochenen Wissenschaftssprache und vor allem funktionelle Kontexte für diese vermittelt werden, in denen diese üblicherweise auftreten könnten. Gemeint ist damit vor allem die rezeptive Perspektive: Da Referate häufig auch Bestandteil der Wissensvermittlung sind, sollten Studierende auf zu erwartende Formen vorbereitet werden. Auch im Hinblick auf gegenseitiges Feedback sollten Studierende wissen, dass diese Formen

durchaus üblich sind und nicht sanktioniert werden müssen. Daneben wären in diesem Zusammenhang aber auch die Funktionen von Verschmelzungsformen näher zu beleuchten. Weitere Untersuchungen der studentischen Vorträge aus dem GeWiss-Korpus haben etwa gezeigt, dass Mündlichkeitsphänomene wie bspw. Verbapokope und Modalwörter häufig der epistemischen Abschwächung oder der Markierung der Vorläufigkeit der eigenen Aussagen dienen (Fandrych/Wallner, angenommen). Für die Vermittlung sind diese Befunde hoch relevant, da gerade diese Art von Funktionalisierungen sprachlicher Mittel für einen adäquaten Sprachgebrauch entscheidend sind. Inwieweit dieser modalisierende Effekt auch durch den Gebrauch von Verschmelzungsformen unterstützt wird, müsste jedoch noch systematisch überprüft werden.

Nicht zuletzt verdeutlichen Studien wie diese weiterhin bestehende Desiderata in der Lernerkorpuslandschaft: So existieren, trotz der sehr dynamischen Situation im Bereich gesprochensprachlicher L2-Korpora des Deutschen (vgl. Wisniewski, 2022), neben dem GeWiss-Korpus nach wie vor kaum Korpora, die (erst- und lernersprachliche) Daten zur (gesprochenen) Wissenschaftssprache zur Verfügung stellen (vgl. aber das vietnamesische Deutschlernerkorpus VIELKO; Ho, 2023). Solche Daten sind aber enorm relevant, um genauer verstehen zu können, wie Studierende mit Deutsch als L1 oder als L2 sich wissenschaftssprachliche Strukturen aneignen und sich damit sprachlich im Laufe ihres Studiums professionalisieren. Hier ist zudem bis jetzt sehr wenig darüber bekannt, inwiefern sich diese beiden Gruppen potenziell in ihren sprachlichen Entwicklungswegen voneinander unterscheiden. Ähnlich stellt sich aber nicht zuletzt die Situation für die Erforschung der gesprochenen deutschen Wissenschaftssprache für Sprecher:innen mit Deutsch als L1 dar. Während in der vorliegenden Studie Daten von L1-Sprecher:innen untersucht wurden, die in Deutschland erhoben wurden und daher u.U. Klitisierungsformen als besonders präsent herausgearbeitet wurden, die gerade im ‚deutschländischen' Kontext besonders frequent sind, ist hier im DACHL-Raum sicherlich von einer großen und bedeutsamen Variationsbreite unter L1-Sprecher:innen auszugehen, deren Erforschung ebenfalls ein dringendes Desiderat darstellt.

Literatur

Bayerlein, Oliver (2014): *Campus Deutsch*. Ismaning: Hueber.

Biber, D. & Conrad, S. & Reppen, R. & Byrd, P. & Helt, M. (2002). Speaking and Writing in the University: A Multidimensional Comparison. *TESOL Quarterly*, 36, 9-48. DOI: 10.2307/3588359.

Duden Grammatik = Wöllstein, A. (2022). *Duden: Die Grammatik*(10., völlig neu verfasste Auflage.). Berlin: Dudenverlag.

Fandrych, C. (2022). Konzepte der Grammatikvermittlung (auch) im Kontext mündlicher und digitaler Kommunikationsformate. In: Demmig, S. & Reitbrecht, S. & Sorger, B. & Schweiger, H. (Hrsg.), *idt 2022 mit.sprache.teil.haben. Band: Beiträge zur Methodih und Didaktik Deutsch als Fremd*Zweitsprache*. Berlin: Erich Schmidt.

Fandrych, C. & Wallner, F. (2022). Funktionale und stilistische Merkmale gesprochener fortgeschrittener Lerner:innensprache: Methodische und konzeptionelle Überlegungen am Beispiel von GeWiss. *Zeitschrift für Germanistische Linguistik*, 50(1), 202–239.

Fandrych, C. & Wallner, F. (2023). Das GeWiss-Korpus: Neue Forschungs- und Vermittlungsperspektiven zur mündlichen Hochschulkommunikation. In: Deppermann, A. / Fandrych, C. / Kupietz, M. / Schmidt, T. (Hrsg): *Korpora in der germanistischen Sprachwissenschaft: Mündlich, schriftlich, multimedial*. Berlin / Boston: De Gruyter, 129-160.

Fandrych, C. & Wallner, F. (angenommen). Positionierungen in studentischen Vorträgen. Eine korpuslinguistische Analyse. Erscheint in *Deutsch als Fremdsprache*, 61(4).

Fandrych, C. & Schmidt. T. & Wallner, F. & Wörner, K. (2023) (Hrsg.). Zugänge zu mündlichen Korpora für DaF und DaZ: Das ZuMult-Projekt. *Korpora Deutsch als Fremdsprache*, 3(1).

Frick, E. & Helmer, H. & Wallner, F. (2023). ZuRecht: Neue Recherchemöglichkeiten in Korpora gesprochener Sprache für Gesprächsanalyse und Deutsch als Fremd- und Zweitsprache. *Korpora Deutsch als Fremdsprache,* 3(1), 44-71. DOI: 10.48694/kordaf.3730.

Ho, T.B.V. (2023). VIELKO - Vietnamesisches Lernerkorpus. *Korpora Deutsch als Fremdsprache,* 3(1), 152–158. DOI: 10.48694/kordaf.3739.

Schäfer, Susanne & Heinrich, Dietmar (2010): *Wissenschaftliches Arbeiten an deutschen Universitäten: Eine Arbeitshilfe für ausländische Studierende im geistes- und gesellschaftswissenschaftlichen Bereich*. München: Iudicium.

Schmidt, T. & Schwendemann, M. & Wallner, F. (2023). Transkriptvisualisierung und Arbeiten mit Transkripten. *Korpora Deutsch als Fremdsprache*, 3/1, 72-91. DOI: 10.48694/kordaf.3723.

Schwendemann, M. & Wallner, F. (2023). Mündlichkeitsphänomene in der gesprochenen Wissenschaftssprache: Korpuslinguistische Befunde und didaktische Perspektiven. *Informationen Deutsch als Fremdsprache*, 50/5, 505-524. DOI: 10.1515/infodaf-2023-0083.

Slavcheva, A. & Meißner, C. (2014). Also und so in wissenschaftlichen Vorträgen. In: Fandrych, Christian & Meißner, Cordula & Slavcheva, Adriana (Hrsg.), *Gesprochene Wissenschaftssprache: Korpusmethodische Fragen und empirische Analysen*. Heidelberg: Synchron, 113-132.

Wallner, F. (2017). Diskursmarker funktional. Eine quantitativ-qualitative Beschreibung annotierter Diskursmarker im GeWiss-Korpus. In: Fandrych, C. & Meißner, C. & Wallner, F. (Hrsg.), *Gesprochene Wissenschaftssprache – digital. Verfahren zur Annotation und Analyse mündlicher Korpora*. Tübingen: Stauffenburg, 107-122.

Wallner, F. (2023). GeWiss Ein Korpus der gesprochenen Wissenschaftssprache. In: *Korpora Deutsch als Fremdsprache*, 3(1), 159-165. DOI: 10.48694/kordaf.3738.

Westpfahl, S. & Schmidt, T. & Jonietz, J. & Borlinghaus, A. (2017). *STTS 2.0: Guidelines für die Annotation von POS-Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS). Arbeitspapier*. IDS Mannheim. Online: https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/6063/file/Westpfahl_Schmidt_Jonietz_Borlinghaus_STTS_2_0_2017.pdf [12.12.2023].

Wisniewski, K. (2022). Gesprochene Lernerkorpora des Deutschen: Eine Bestandsaufnahme. *Zeitschrift für germanistische Linguistik*, 50(1), 1-35. DOI: 10.1515/zgl-2022-2047.

Tema

GÉNÉRER DES PRINCIPES DIDACTIQUES À PARTIR D'UN CORPUS

Kann ein Korpus an Beobachtungen und Aufzeichnungen dazu dienen, um die Wirksamkeit bestimmter Handlungen von Lehrpersonen auf den Kompetenzzuwachs von Lernenden aufzuzeigen? Die Antwort lautet « ja », wie die Dissertation der Autorin dieses Artikels zu mündlichen Interaktionen zwischen Schülerinnen und Schülern und den Einführungen durch verschiedene Lehrpersonen zu Sprechaufgaben zeigt. Aus der Korpusanalyse werden didaktische Prinzipien abgeleitet, mit denen die mündlichen Interaktionen der Lernenden durch einen verbesserten Input der Lehrperson optimiert werden können. Die Untersuchung zeigt, dass die Art und Weise, wie eine Lehrperson eine Aufgabe zum interaktiven Sprechen einführt, die Bearbeitung dieser Aufgabe durch die Lernenden massgeblich beeinflusst.

Gwendoline Lovey
| PH FHNW

Gwendoline Lovey est chercheuse et chargée d'enseignement à l'institut du primaire de la Haute École Pédagogique FHNW. Elle a soutenu sa thèse « Interaktives Sprechen im lehrwerkbasierten Fremdsprachenunterricht der Grundschule » en 2023 à l'Université d'Augsbourg en didactique des langues et littératures romanes.

Recherche sur la production orale

La production orale est souvent considérée comme l'objectif central de l'enseignement des langues étrangères (Lütge 2014, 147) car la capacité à s'exprimer verbalement jouit d'une valeur particulière dans la société. Or, contrairement aux compétences écrites, les études empiriques sur les compétences orales restent rares et sont souhaitées par la communauté scientifique (Burwitz-Melzer 2014, 25 ; Martinez 2014, 160). Dans un enseignement ancré dans l'instructivisme, l'interaction orale en classe se limite à celle entre l'enseignant.e et un.e ou plusieurs élèves. Avec le passage à un enseignement basé sur une approche plus constructiviste, les interactions orales entre les pairs deviennent plus importantes (Schramm/Aguado 2010, 191-192). Cette approche part en effet du principe que l'échange oral entre les élèves, par exemple dans le cadre d'un travail en binôme, favorise davantage l'apprentissage que les activités en grand groupe (Wolff 2002). La thèse « Interaktives Sprechen im lehrwerkbasierten Fremdsprachenunterricht der Grundschule » (Lovey 2024) prend en considération ce changement de paradigme et examine, d'une part, comment un.e enseignant.e peut encourager la production orale des élèves en classe de langue et, d'autre part, si une interaction orale entre pairs est possible dès l'école primaire alors que les élèves sont encore au stade de débutants. Le présent article expose quelques analyses choisies de l'étude, qui débouchent sur des conclusions pour la pratique.

Étude menée

L'étude menée porte sur les interactions orales entre pairs à l'école primaire en cours de langue étrangère. La question de recherche générale est la suivante : « Comment est abordée la compétence 'parler en interaction' dans l'enseignement du français (L2) basé sur un moyen d'enseignement au niveau primaire ? ». Cet article ne présente qu'une partie des

Enseignante	Madame Müller	Madame Huber	Madame Schmid	Madame Gerber
Nombre d'élèves	13	19	16	16
Milieu de l'école	rural	urbain	urbain	urbain
Nombre de langues dans la classe	1	11	9	9
L1 des élèves focus	suisse allemand	(suisse) allemand, arabe, italien, serbe, somalien, turc	(suisse) allemand, anglais, italien, turc	(suisse) allemand, anglais, italien

Tab. 1 Caractéristiques des classes participant à l'étude

	Madame Müller		Madame Huber	Madame Schmid	Madame Gerber
Activité A : *Quiz* Répondre à des questions de quiz	réalisée en groupes (3)		réalisée en groupes (3)	réalisée en groupes (3)	réalisée en groupes (2)
Activité B : *Questionnaire* Vérifier les hypothèses sur les réponses possibles à un questionnaire	réalisée en groupes (3)		réalisée en groupes (1)	réalisée en groupes (6)	supprimée
Activité C : *Questions* Échanger les réponses à des questions	C1 réalisée en groupes (2)	C2 réalisée en groupes (3)	réalisée en plénière	réalisée en plénière	réalisée en plénière
Activité D : *Trucs à savoir* Répondre à des questions de connaissances	supprimée		réalisée en groupes (2)	supprimée	réalisée en groupes (1)
Activité E : *Métiers* Échanger sur son métier de rêve	E1 réalisée en plénière	E2 réalisée en groupes (3)	supprimée	réalisée en plénière	supprimée

Tab. 2 Réalisation et formes sociales des activités menées dans les classes

résultats du projet de recherche, à savoir les deux aspects suivants :

A. Comment les enseignant.e.s introduisent-elles/ils les activités d'interaction orale proposées par le moyen d'enseignement ?
B. Quelles sont les caractéristiques de l'interaction orale des élèves lors du travail avec ces activités ?

Pour répondre à la question de recherche, un corpus a été constitué de ce qui se dit en classe au moment de travailler des activités d'interaction orale, moyennant des observations. On a donc enregistré, d'une part, les introductions réalisées par les enseignant.e.s pour les activités d'interaction orale et, d'autre part, les interactions orales des élèves lors du travail avec ces activités.

L'étude s'est déroulée dans quatre classes de 8e Harmos, dans le canton de Soleure. Le français y est enseigné comme première langue étrangère avant l'anglais et selon les mêmes modalités (nombre de leçons hebdomadaires, les objectifs d'apprentissage et le moyen d'enseignement). Les caractéristiques des classes sont présentées dans le tableau 1. Les noms des enseignantes et des élèves ont été anonymisés. Les quatre classes représentent les trois centres urbains et une région rurale du canton. Tous les élèves ont entre 12 et 13 ans. Pour chaque classe, l'enseignante choisit 6 élèves focus, à savoir 2 élèves plutôt fort.e.s, 2 élèves moyen.ne.s et 2 élèves plutôt faibles. L'étude porte donc sur 24 élèves répartis dans ces quatre classes soleuroises de 8e Harmos et leurs 4 enseignantes. Le point de départ de l'enquête est fixé en 2016 alors que tous les élèves du canton de Soleure apprennent le français à partir de la 5e Harmos (8 ans) à l'aide de la 1re édition du moyen d'enseignement *Mille feuilles* (Ganguillet et al. 2014).

Dans les activités proposées par *Mille feuilles* (ibid.) visant à développer les compétences d'interaction orale, les élèves sont encouragés à échanger de nouvelles informations ensemble. Les élèves abordent ainsi des sujets de discussion réels en petits groupes, ce qui requiert une double concentration sur le contenu et sur la langue, susceptible d'entraîner une « surcharge cognitive liée à la double mobilisation du fond et de la forme » (Manoïlov 2019, 25). Pour atténuer cette surcharge cognitive, des idées de réponse et des débuts de phrase sont systématiquement fournis pour chaque activité d'interaction orale.

Les classes participant à l'étude suivent 2 périodes de français par semaine et peuvent être observées lors du travail avec les 5 activités d'interaction orale du parcours d'apprentissage présélectionné ; à savoir *Mille feuilles 6.2, Quelle question !* (Ganguillet et al. 2014). Cependant, aucune enseignante ne prévoit de travailler les cinq activités d'interaction orale avec sa classe (cf. Tab. 2).

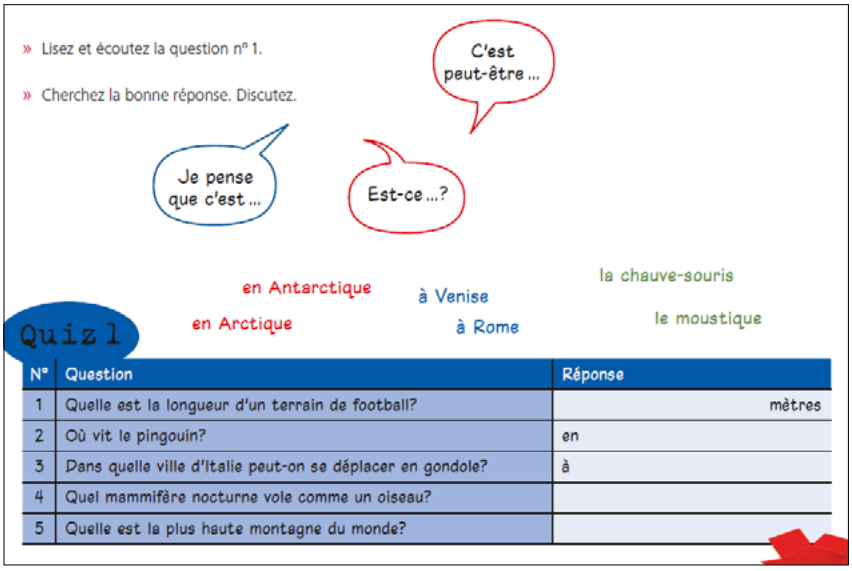


Fig. 1
Extrait de l'activité A *(Quiz)*, *Mille feuilles* (Ganguillet et al. 2014)[2]

Données présentées

Dans le présent article, est exposé un échantillon du travail effectué autour de l'activité A *(Quiz)*. Cette activité a été choisie pour illustrer le travail des enseignantes et des élèves observés puisqu'elle a été travaillée par toutes les classes et que toutes les enseignantes ont choisi de travailler l'interaction orale en petits groupes (cf. Tab. 2). Un extrait de l'activité A illustre la manière dont les interactions orales sont conçues dans ce moyen d'enseignement (voir Fig. 1). Dans l'activité A *(Quiz)*, les élèves répondent ensemble aux questions d'un quiz. Pour créer de plus longues séquences et pour s'engager dans une interaction orale du type question-réponse-validation, les élèves ajoutent des expressions formulaïques à leurs réponses. Les débuts de phrases tels que « Je pense que c'est... » ou « C'est peut-être... » invitent leur interlocuteur/interlocutrice à une réaction du type « Oui, c'est juste » ou « Non, je ne pense pas ». Les expressions formulaïques, appelées *chunks*, sont généralement présentées dans des bulles à côté de l'activité.

L'analyse se concentre ici sur 4 critères :

1. Durée (de l'introduction et du travail en groupes)
2. Forme (avec ou sans participation des élèves / en plénière ou en petits groupes)
3. Utilisation de la langue cible (par l'enseignante et les élèves focus)
4. Gestion des *chunks* (expressions formulaïques dans les bulles)

Le tableau 3 montre une vue d'ensemble du travail réalisé pour l'activité A *(Quiz)* par les quatre enseignantes avec leurs classes respectives pour les quatre critères.

L'objectif ambitieux d'une introduction à une activité d'interaction orale consiste à donner aux élèves tous les éléments pour qu'ils puissent travailler le plus longtemps possible de manière autonome, tout en parlant le plus et le plus correctement possible en français.

Pour le critère de la durée, on distingue la durée de l'introduction à l'activité d'interaction orale de celle du travail effectué en autonomie par les élèves focus par la suite. On constate que le pourcentage du temps de parole des élèves varie entre

Mme Müller, Mme Huber et Mme Schmid réalisent quatre activités d'interaction orale chacune, Mme Gerber en réalise trois. Lors de la réunion préparatoire avec la chercheuse, les enseignantes expliquent qu'elles suppriment certaines activités parce qu'elles veulent passer plus rapidement au parcours d'apprentissage suivant, parce qu'elles ne correspondent pas à leur tâche finale ou parce qu'une activité ne les convainc pas. En revanche, Mme Müller travaille les activités C et E de deux manières différentes (C1/C2, E1/E2) et Mme Schmid multiplie les interactions de l'activité B en l'organisant comme un speed-dating avec des changements réguliers de partenaires. Certaines activités d'interaction orale sont réalisées en petits groupes de 2 ou 3 élèves comme indiqué dans le moyen d'enseignement, d'autres sont travaillées en plénière pour différentes raisons (gestion de la classe, du temps, des devoirs etc.). Parfois, le/les groupe/s plus faibles ne parviennent pas à réaliser l'activité. Pour la classe de Mme Gerber par exemple, le groupe d'élèves focus faibles reste muet lors de l'activité A. En revanche, pour l'activité D, il s'agit d'une mesure de différenciation de la part de l'enseignante et l'activité n'est proposée qu'aux élèves focus fortes. Finalement, le corpus comprend 17 introductions réalisées par les enseignantes (transcriptions orthographiques) et 37 interactions entre les élèves (transcriptions phonétiques en API). Les introductions des enseignantes durent entre 0'59'' et 11'54''. Les interactions entre élèves durent entre 0'0'' et 11'54''.[1]

1 Toutes les transcriptions de la thèse peuvent être consultées dans l'annexe disponible en ligne (Lovey 2024).
2 Depuis 2020 il existe une nouvelle édition de ce magazine : Cavelti et al. 2020.

Activité A (Quiz)

Critères		Madame Müller et sa classe		Madame Huber et sa classe		Madame Schmid et sa classe		Madame Gerber et sa classe	
1. Durée	Durée de l'introduction (enseignante)	06'13''	37 %	11'54''	45 %	00'59''	14 %	07'02''	27 %
	Durée du travail en groupes (élèves)	10'40''	63 %	14'30''	55 %	06'04''	86 %	19'20''	73 %
2. Forme	Forme de l'introduction	exclusive		participative		exclusive		participative	
	Forme de l'interaction	en groupes		en groupes		en groupes		en groupes	
3. Utilisation de la langue cible	Parties en français durant l'introduction (enseignante)	31 %		91 %		79 %		73 %	
	Parties en français durant l'activité (élèves focus : fort.e.s / moyen.ne.s / faibles)	52 %	57 %	63 %	56 %	84 %	62 %	93 %	84 %	80 %	47 %	29 %	0 %
4. Gestion des chunks	Introduction des chunks (enseignante)	oui		oui		non		oui	
	Exemple(s) en plénière avec des chunks (enseignante)	oui		non		non		oui	
	Utilisation des chunks durant l'activité (élèves focus : fort.e.s / moyen.ne.s / faibles)	oui	non	oui	non	non	non	non	non	non	oui	oui	non

Tab. 3
Vue d'ensemble du travail autour de l'activité A (Quiz)[3]

55 % (classe de Mme Huber) et 86 % (classe de Mme Schmid).

Le critère de la forme détermine la manière dont a été introduite puis réalisée l'activité d'interaction orale. Les quatre enseignantes mettent en place un travail en groupes. Tandis que Mme Müller et Mme Schmid expliquent elles-mêmes les consignes et donnent quelques exemples, Mme Huber et Mme Gerber font participer la classe à l'explication des consignes.

Le critère de l'utilisation de la langue cible indique le pourcentage des énoncés en français par rapport à l'utilisation de l'allemand, du suisse allemand ou d'autres langues. Le pourcentage est calculé sur la base du nombre de mots (nombre de mots en français par rapport au nombre total), et ceci pour les enseignantes au moment de l'introduction et pour les élèves au moment de travailler en autonomie. Les pourcentages varient entre 31 % et 91 % pour les enseignantes et entre 0 % et 93 % pour les élèves focus.

Le critère de la gestion des *chunks* s'applique également aux deux groupes (enseignantes et élèves focus) : il est d'abord indiqué si les *chunks* sont introduits et/ou utilisés dans des exemples par l'enseignante, et ensuite si ces derniers sont utilisés dans les travaux de groupes par les élèves focus.

Dans les paragraphes qui suivent, les interactions de chaque enseignante avec sa classe sont décrites afin de déterminer dans quelle mesure ses actions lors de l'introduction influencent les interactions orales de ses élèves.

Madame Müller et sa classe

Dans son introduction à l'activité A *(Quiz)*, Mme Müller explique la consigne, introduit les *chunks* et fait des exemples :

Madame Müller 04:31
Lest immer zuerst die Frage laut, *par exemple « Qui a crié Eurêka ? »* und nachher die Antwort dazu *« Je pense que c'est Achimedes (sic). »* oder *« C'est peut-être Archimedes ? »* ou *« Est-ce Archimedes ? »*. Dann *« oui »* ou *« non »*, *« j'ai quelque chose d'autre »* oder so. Dass ihr das auf Französisch macht. *D'accord ?* Auftrag klar ?

3 L'analyse des activités B, C, D, E corrobore les résultats concernant les quatre critères présentés pour l'activité A *(Quiz)*.

L'introduction de Mme Müller dure 6 minutes et 13 secondes, dont 1 minute et 37 secondes sont employées pour des explications grammaticales, ce qui équivaut à plus d'un quart de l'introduction. Ensuite, les élèves disposent de 10 minutes et 40 secondes pour travailler en groupes ce qui correspond à une répartition de 37 % pour l'introduction et 63 % pour le travail effectué en autonomie.

L'introduction est en majeure partie exclusive : Mme Müller explique et les élèves écoutent. À certains moments, ils répètent les *chunks* en chœur et ils répondent aux questions de Mme Müller qui leur demande de traduire les *chunks* mot à mot.

Lors de l'introduction, Mme Müller a plus souvent recours à l'allemand (69 %) qu'au français (31 %). Ceci est principalement dû aux traductions qu'elle fournit et aux explications de grammaire pour lesquelles Mme Müller utilise également l'allemand standard. Pour introduire l'activité A *(Quiz)*, elle procède à une analyse formelle des *chunks*, en analysant d'abord l'ordre des mots en français et en allemand dans la phrase subordonnée et en expliquant ensuite de quels mots se compose l'expression « c'est » :

Madame Müller 02:43
genau oder « das ist », hä ? Also das ist die Abkürzung eigentlich *c* Apostroph von *ce* (elle note *« ce »* et *« c' »* au tableau) (2) « es » oder *« est »* « ist » das, *« est »*. Und jetzt haben wir da das Umgekehrte : Jetzt kommt zuerst *« est »* und nachher das *« ce »* (3). Also wenn das hier « das » ist, heisst das wohl (dit le nom d'un élève)

Elève Mü-1 03:17
« ist es »

Madame Müller 03:18
« ist es » oder « ist das », hä ?

Le pourcentage de l'utilisation de la langue cible est plus élevé pour les élèves focus que pour leur enseignante. Ils utilisent le français entre 52 % à 63 % pour leurs énoncés durant les 10 minutes et 40 secondes d'activité. On constate cependant que les questions et les réponses sont certes dites en français mais que les échanges autour de ces phrases se font principalement en allemand ou en suisse allemand :

Elève focus Mü-1 00:10
Liesisch du zerst vor ?
Liest du zuerst vor ?

Elève focus Mü-2 00:12
[kɛl ɛ la kɛl ɛ lɔ̃ ɛ la lɔ̃ŋʃəʁ dy tɛʁɛ̃ də futbal.]
Quelle est la longueur d'un terrain de football ?

Elève focus Mü-1 00:19
eh [sɑ̃ vɛ̃t mɛtʁ]
cent vingt mètres

Elève focus Mü-2 00:23
Du muesch e Satz nä. Du muesch irgendwie e Satz vo do nä.
Du musst einen Satz nehmen. Du musst irgendwie einen Satz von hier nehmen.

Elève focus Mü-1 00:29
[ʒə pɑ̃s kə] (2) [sɑ̃ vɛ̃t mɛtʁ.]
Je pense que cent vingt mètres (sic).

Elève focus Mü-2 00:34
[wi se ʒyst.]
Oui, c'est juste.

Le recours à l'allemand voire au suisse allemand par les élèves reflète le comportement linguistique de Mme Müller. Ce constat coïncide avec les résultats d'autres études empiriques, comme par exemple celle de Tesch (2010), qui montre que l'utilisation de la langue cible par les élèves dépend fortement du langage employé par l'enseignant.e (cf. ibid. : 195). En ce qui concerne la gestion des *chunks*, nous avons constaté que Mme Müller fournissait des explications grammaticales lors de son introduction dans le but d'assurer un emploi correct de ces expressions. Cependant, l'analyse du corpus révèle que ses explications ont peu de succès, et peuvent même avoir un effet négatif sur les énoncés des élèves. Le *chunk* « Je pense que c'est… » est utilisé de manière erronée par la plupart des élèves lors du travail en autonomie. Dans l'extrait ci-dessus, l'élève focus Mü-1 dit « Je pense que cent vingt mètres », en omettant le « c'est » pourtant si longuement expliqué par Mme Müller. En tout, l'élève focus Mü-1 répète cette erreur cinq fois durant le travail de l'activité A *(Quiz)*, en disant aussi « Je pense que à Rome », « Je pense que Mount Everest », « Je pense que Neil Armstrong » et « Je pense que 'au revoir' ». Les élèves focus Mü-5 et Mü-6 omettent le « c' » chaque fois qu'ils utilisent ce même *chunk* durant l'activité A *(Quiz)* :

Elève focus Mü-5 06:32
[ʒə pɔ̃s (.) kə ɛ sɑ̃ vɛ̃t mɛtʁ]
Je pense que est 120 mètres (sic).

Elève focus Mü-6 06:40
[ʃə pɑ̃s kə ɛ ɑ̃n ɑ̃taʁktik]
Je pense que est en Antarctique (sic).

Elève focus Mü-6 07:19
[ʃə pɑ̃s kə ɛ la ʃuvɛsuʁi.]
Je pense que est la chauve-souris (sic).

Elève focus Mü-5 07:41
[ʒə pɔs kə ɛ munt ɛːvɛːrɛːst]
Je pense que est Mount Everest (sic).

Elève focus Mü-5 08:18
[ʒpɔs kə ɛ tʃapon]
Je pense que est Japon (sic).

Tandis que l'erreur persiste pour les élèves focus dits faibles, elle est corrigée dans le groupe des élèves fortes. Lorsque l'élève focus Mü-1 répond à la question « Comment dit-on 'auf Wiedersehen' en français ? » par « Je pense que 'au revoir' », l'élève focus Mü-2 la corrige en disant correctement la phrase complète « Je pense que c'est 'au revoir' ». Lors d'une utilisation ultérieure, l'élève focus Mü-1 finit par utiliser correctement ce *chunk* en disant « Je pense que c'est Pierre ». La correction est donc atteinte par le biais de l'imitation de la forme correcte entre pairs.

Madame Huber et sa classe

Dans son introduction à l'activité A *(Quiz)*, Mme Huber explique à ses élèves qu'ils doivent lire les questions et les réponses à haute voix en utilisant les *chunks* donnés par le moyen d'enseignement. L'introduction dure 11 minutes et 54 secondes, dont 8 minutes et 4 secondes sont employées pour faire des exemples en plénière, ce qui équivaut à plus de deux tiers de l'introduction. Ensuite, les élèves disposent de 14 minutes et 30 secondes pour travailler en groupes ce qui correspond à une répartition de 45 % pour l'introduction et 55 % pour le travail effectué en autonomie. Les élèves de la classe de Mme Huber disposent proportionnellement du moins de temps pour travailler l'activité de manière autonome si on la compare aux trois autres classes. Cependant, dans la classe de Mme Huber, certains élèves participent déjà à hauteur de 34 % à l'oral pendant l'introduction, car Mme Huber organise cette phase de manière participative. Ainsi, les élèves ont l'opportunité de s'exercer à la prise de parole interactive en plénière dès le début de l'activité :

Madame Huber 30:49
Alors qu'est-ce qu'on doit faire ? Sur cette page ? (3)

Elève Hu-7 30:57
3) Wir müssen die Fragen hier beantworten.

Madame Huber 31:03
Oui. Et tu essaies en français ? Toi aussi ?

Elève Hu-7 31:08
Kästchen [remplir]

Madame Huber 31:10
remplir

Elève Hu-7 31:11
[ʁɑ̃pliʁ] [lə] Kästchen (il rigole)
remplir le « Kästchen »

Madame Huber 31:16
remplir le questionnaire, d'accord. Et ici, sur cette page ? On doit faire quoi ? C'est la même chose ?

Lors de l'introduction, Mme Huber utilise la langue cible à 91 %. Elle gère la classe en français, en utilisant des mots parallèles, en montrant ce qu'il faut faire par des gestes ou en répétant plusieurs fois la même question en variant le vocabulaire et les structures. Le pourcentage de l'utilisation de la langue cible est aussi assez élevé auprès des élèves focus. Ils utilisent le français entre 56 % à 84 % pour leurs énoncés durant les 14 minutes et 30 secondes d'activité. On constate que même les élèves focus dits faibles essaient de parler français ensemble pour organiser le travail ou pour clarifier des questions qui s'éloignent de l'activité d'interaction orale à proprement parler, donc des questions et des réponses du quiz. L'élève focus Hu-5 se renseigne par exemple en français auprès de ses camarades sur le numéro de page à laquelle se trouve l'activité à faire :

Elève focus Hu-5 00:11
[a la paʃ?]
à la page ?

Elève focus Hu-6 00:13
[a la paʒ] eh [disyiet]
à la page, eh, dix-hiéte

Elève Hu-8 00:16
[nɔ̃ nɔ̃ nɔ̃]
non non non

Elève focus Hu-5 00:18
[a la paʒ disyit]
à la page dix-huit

Elève focus Hu-6 00:20
[wi]
oui

Elève focus Hu-5 00:21
[e disnoef]
et dix-neuf

Lors de l'introduction, Mme Huber explique la signification des *chunks* et les fait répéter à ses élèves en chœur. Or, ils ne sont pas utilisés dans les exemples qui occupent pourtant une part importante de son introduction à l'activité : les élèves donnent certes les réponses en français mais sans les compléter par un des *chunks* proposés par le moyen d'enseignement, ce qui leur permettrait de faire des phrases plus longues. Ce qui est le cas durant le travail en plénière l'est aussi durant les travaux en groupes puisqu'aucun.e élève n'utilise ces *chunks* durant le travail en autonomie.

Madame Schmid et sa classe

Dans son introduction à l'activité A *(Quiz)*, Mme Schmid explique la consigne en insistant sur le fonctionnement de l'interaction orale :

Madame Schmid 00:01
et la deuxième personne, elle donne la réponse, ensuite vous changez les rôles (indique le changement avec les mains), *c'est cette, la personne qui a donné la réponse qui pose maintenant la question et l'autre personne donne la réponse, et vous continuez jusqu'à la, à, au quiz numéro six* (montre avec les mains que cela continue), *question numéro 10. Alors vous commencez avec le quiz numéro cinq, première question, et vous finissez au quiz numéro six, dixième question. D'accord ? Bon.*

La forme de l'introduction est exclusive puisque seule Mme Schmid s'exprime. Les élèves n'interagissent donc pas encore oralement au moment de l'introduction. Or, ils disposent de 6 minutes et 4 secondes pour travailler en groupes par la suite ce qui correspond à une répartition de 14 % pour l'introduction et 86 % pour le travail effectué en autonomie. En effet, l'introduction de Mme Schmid ne dure que 59 secondes. Grâce à son introduction concise, Mme Schmid parvient à optimiser le temps pendant lequel les élèves peuvent s'exprimer. La courte durée de l'introduction de Mme Schmid s'explique, entre autres, par le fait qu'elle ne mentionne pas explicitement les *chunks* disponibles pour les élèves.

Lors de l'introduction, Mme Schmid s'exprime à hauteur de 79 % en français. Elle se permet une parenthèse en allemand au moment d'expliquer la présence des « Diktiergeräte ». Le pourcentage de l'utilisation de la langue cible est aussi assez élevé auprès des élèves focus. Ils utilisent le français entre 80 % à 93 % pour leurs énoncés durant les 6 minutes et 4 secondes d'activité. Les élèves de la classe de Mme Schmid utilisent aussi des expressions en français pour maintenir la communication entre eux au moment de faire l'activité. Ces dernières ne sont pas notées dans des bulles (p.ex. « c'est ton tour ») :

Elève focus Sc-1 01:17
[komɑ̃ sapel lə ʃĩɛ̃ də bul]
Comment s'appelle le chien de Boule ?

Elève focus Sc-2 01:21
[il sapel bilə]
Il s'appelle Bill.

Elève focus Sc-1 01:24
ehm, [wui :]
Oui.

Elève focus Sc-1 01:26
(3.5) [sɛ tɔ̃ tuʁ]
C'est ton tour.

Elève focus Sc-2 01:31
ehm [ki sɔ̃ le soeʁ də mɔ̃ pɛʁ]
Qui sont les sœurs de mon père ?

Elève focus Sc-1 01:36
[me tɑ̃t]
Mes tantes.

Pendant les interviews, Mme Schmid et ses élèves confirment que la répétition de ces expressions se fait régulièrement en classe de français afin d'entrainer les élèves à parler français ensemble. On observe en revanche dans cet extrait que les élèves de la classe de Mme Schmid donnent les réponses sans ajouter de début de phrase

et qu'ils n'utilisent donc pas les *chunks* proposés par le moyen d'enseignement.

Madame Gerber et sa classe

Dans son introduction à l'activité A *(Quiz)*, Mme Gerber explique la consigne, introduit les *chunks* et fait des exemples. Elle répète plusieurs fois que l'activité ne consiste pas seulement à trouver les réponses aux questions mais qu'il faut aussi échanger en langue cible :

Madame Gerber 04:02
Also nicht nur eure Idee sagen, sondern ihr sollt auch immer sagen *« Je pense que c'est »* oder *« Est-ce »*, « Ist das » oder *« C'est peut-être »*, « Das ist vielleicht ».

La forme de l'introduction est plutôt participative puisque Mme Gerber fait lire les consignes à ses élèves. Elle leur pose ensuite des questions par rapport au déroulement de l'activité et leur fait dire les *chunks* en allemand.

Mme Gerber utilise le français à hauteur de 73 % pour son introduction à l'activité de l'interaction orale. Dans les moments où elle s'exprime en langue cible, elle s'appuie sur des informations notées au préalable au tableau telles que le matériel à utiliser, les numéros de page, la forme sociale et le temps à disposition pour cette activité. Elle utilise assez régulièrement l'allemand pour traduire des mots ou des expressions, notamment pour gérer la classe ou pour s'assurer que les élèves comprennent ses questions ou ses consignes. Elle répète alors en allemand ce qu'elle dit en français, créant une sorte d'écho :

Madame Gerber 03:12
Vous discutez. Also das heisst ? Was müsst ihr machen ?
Qu'est-ce que vous devez faire ?
[…]

Madame Gerber 05:27
Alors utilisez, verwendet, *utilisez ces bulles pour discuter.*
[…]

Madame Gerber 05:46
Die Lösungen, *regardez les solutions à la page 89.*

Les traductions rallongent d'une part la durée de l'introduction et, d'autre part, donnent lieu à un discours en langue cible ponctué d'interruptions. Au moment du travail en groupes, on observe une alternance des langues similaire auprès des élèves focus puisqu'elles communiquent aussi les informations importantes non seulement en français, mais aussi en allemand :

Critères	Madame Gerber et sa classe	
Utilisation de la langue cible dans l'activité A (Quiz)	Partie 1 (00:00-05:00)	Partie 2 (05:01-19:20)
Elèves focus Ge-1 et Ge-2	75 %	36 %
Elèves focus Ge-3 et Ge-4	50 %	22 %
Elèves focus Ge-5 et Ge-6	1 %	0 %

Tab. 4
Utilisation de la langue cible par rapport à la durée de l'activité

Elève focus Ge-1 00:55
[kel ɛ ply lɔ̃ nom də vil də fʁɑ̃s]
Quel est le plus long nom de ville de France ?

Elève focus Ge-4 00:59
[sɛʁə]
Saint-Ré…

Elève focus Ge-1 00:59
Welcher ist der grösste Name in Französisch ?

Elève focus Ge-1 01:03
(3) [sɑ̃][sɑ̃][sɑ̃]
Saint, Saint, Saint…

Dans l'exemple ci-dessus, une traduction de la question par l'élève focus Ge-1 ne semble pas nécessaire puisque l'élève focus Ge-4 commence à donner la réponse avant cette dernière.

Le pourcentage de l'utilisation de la langue cible varie entre 0 % à 47 % pour les énoncés des élèves focus durant les 19 minutes et 20 secondes d'activité. C'est le pourcentage le plus bas des quatre classes mais il est à mettre en relation avec la durée la plus longue pour cette même activité. Si on ne prend en considération que les cinq premières minutes du travail en groupes, les pourcentages de l'utilisation de la langue cible sont alors nettement plus élevés que durant le reste du temps (cf. tabl. 4).

Les élèves focus ont bien compris qu'il fallait utiliser les *chunks* que Mme Gerber a introduits et les utilisent pour formuler leurs réponses ce qui mène à des séquences plus longues. Les deux élèves focus dites fortes de la classe de Mme Gerber les emploient si bien qu'elles

donnent la même réponse quatre fois, avec différentes tournures :

Elève focus Ge-1 01:52
[kesk u sə tʁuv lə ply gʁɑ̃ ɛʁ (.) aʁepoʁ doeʁop]
Qu'est-ce que (sic) *Où se trouve le plus grand aéroport d'Europe ?*

Elève focus Ge-2 02:00
[aloʁ ?]
Alors ?

Elève focus Ge-2 02:02
London.

Elève focus Ge-2 02:04
[ʒə pɑ̃s kə se lɔ̃dʁ]
Je pense que c'est Londres.

Elève focus Ge-1 02:08
[se poetɛtʁ lɔ̃dʁ]
C'est peut-être Londres.

Elève focus Ge-2 02:08
Aso London.

La première fois, l'élève focus Ge-2 donne la bonne réponse en allemand et sans l'utilisation d'un *chunk*, la deuxième fois cette même élève ajoute un début de phrase à sa réponse et la formule en français. L'élève focus Ge-1 complète la réponse par une troisième version avec un autre début de phrase et l'élève focus Ge-2 clôt la séquence par une réponse en suisse allemand.

Conclusions et conséquences pour la didactique

Les quatre enseignantes ont consacré plus au moins de temps à l'activité d'introduction orale. Pour optimiser le temps de parole des élèves, il faut essayer de faire en sorte que l'introduction soit la plus courte possible, tout en veillant à donner aux élèves les informations nécessaires de manière à ce qu'ils puissent effectuer le travail attendu en groupes. La technique de Mme Gerber, qui consiste à noter les informations les plus importantes au tableau, s'est révélée utile. En revanche, ce qui prolonge la durée de l'introduction sans montrer d'effet positif, est son utilisation systématique de traductions. Ces interruptions ne semblent pas nécessaires si la langue des instructions est construite de manière cohérente. Mme Schmid montre que cela est possible si on répète régulièrement ces expressions. Les explications grammaticales de Mme Müller sont aussi peu, voire pas du tout, utiles pour lancer une activité d'interaction orale dont l'objectif principal est de travailler sur la communication. L'exemple des élèves focus de Mme Müller montre que les explications métalinguistiques n'ont pas l'effet escompté, mais qu'en revanche les élèves sont tout à fait capables de se corriger mutuellement pendant le travail en groupes. Cette constatation soutient les recommandations didactiques selon lesquelles les enseignant.e.s devraient privilégier une plus grande tolérance aux erreurs pendant la phase de communication (cf. Thaler 2014). Faire des exemples en plénière est également utile. Cependant, comme on peut le voir dans la classe de Mme Huber, de (trop) nombreux exemples prolongent la durée de l'introduction et ne sont peut-être pas indispensables pour tous les élèves de la classe. Dès lors, une mesure de différenciation pourrait consister à demander aux élèves de la classe qui, après une première série d'exemples, souhaite en faire d'autres en plénière avant de poursuivre l'activité en plus petits groupes. Celles et ceux qui ont compris peuvent alors travailler de manière autonome et augmenter ainsi leur temps de parole. Même s'il est évidemment important d'accorder plus de temps au travail de groupes qu'à l'introduction, l'analyse du corpus montre qu'il existe une limite temporelle pour le travail en autonomie pour des élèves de cet âge : dans la classe de Mme Gerber, qui dispose de la plus grande partie du temps pour réaliser l'activité, l'interaction orale entre les apprenants dans la langue cible diminue massivement après cinq minutes. Cette observation d'une baisse assez nette après cinq à sept minutes de travail en autonomie se confirme dans toutes les classes et pour toutes les activités observées.

Quant à la forme de l'introduction, le choix de présenter l'introduction à l'activité d'interaction orale sous une forme participative, permet aux élèves de s'exprimer dans la langue cible dès l'introduction. Dans les classes de Mme Huber et de Mme Gerber, dans lesquelles les élèves lisent ou (co)expliquent les consignes et apportent leurs propres exemples, leur temps de parole augmente. Mais si l'on veut faire participer oralement tous les élèves à l'introduction,

celle-ci se prolonge, comme le montre la durée des introductions de ces deux enseignantes.

Par rapport à l'utilisation de la langue cible, on observe dans ce corpus une relation causale tendancielle entre le comportement linguistique de l'enseignante et celui des élèves. Plus l'enseignante utilise la langue cible, plus les élèves le font également. Cette observation confirme Tesch (2010), qui affirme que si la langue cible n'est que rarement utilisée par l'enseignante comme langue de travail, les élèves ont l'impression qu'il est trop compliqué de communiquer en français. L'effet sur les élèves est alors proportionnel et ils ne réagissent qu'à de petites doses de français. En revanche, si l'enseignant.e recourt systématiquement à la langue cible, elle ou il donne l'impression que c'est quelque chose de tout à fait faisable, même à un niveau débutant. Les exemples de Mme Huber et de Mme Schmid permettent d'illustrer ce point. Pour que les élèves s'expriment en langue cible, l'enseignant.e doit donc jouer un rôle de modèle et parler le plus souvent possible en français. Dans la mesure du possible, il ou elle évite les traductions systématiques, car les élèves reproduisent ce qu'ils entendent. Il est important d'utiliser la langue cible en dehors des moments directement liés à l'interaction orale pour introduire ce langage de classe (organisation du travail, consignes, clarifications, etc.) qui permet aux élèves de maintenir une conversation en français même s'ils s'éloignent de la production langagière minimale exigée par l'activité.

Les différentes manières dont les quatre enseignantes utilisent les *chunks* montrent qu'il est indispensable d'attirer l'attention des élèves sur ce support linguistique. Même si les bulles sont bien visibles sur la page d'activité, les élèves focus n'en font manifestement pas usage à moins que leur enseignante ne les y invite explicitement. Dans les classes de Mme Huber et de Mme Schmid, les *chunks* n'étaient pas présentés en plénière. Par conséquent, les séquences des élèves focus de ces deux classes se résument généralement au schéma question-réponse, contrairement à celles des élèves focus des classes de Mme Müller et de Mme Gerber qui produisent des séquences question-réponse-validation. Ainsi, il ne suffit pas d'inclure les chunks dans le moyen d'enseignement, mais il faut aussi que l'enseignant.e les introduise explicitement avec des exemples concrets. Néanmoins, il est important que le contenu reste au centre des activités communicatives et que les élèves ne les considèrent pas comme des exercices ayant pour seul objectif de s'entraîner à des structures sans signification. Dans l'exemple de la classe de Mme Gerber, l'interaction orale semble assez scolaire et s'éloigne quelque peu de l'authenticité recherchée par le moyen d'enseignement. L'analyse complète du corpus, qui prend en compte toutes les transcriptions des quatre classes pour toutes les activités d'interaction orale observées, y ajoute une analyse thématique et tient également compte des (auto-)évaluations et des entretiens menés avec les enseignantes et leurs élèves, confirme les observations présentées dans cet article. L'analyse surtout qualitative du corpus a en effet montré que les interactions orales des élèves variaient en fonction de la manière dont un.e enseignant.e introduisait une activité d'interaction orale. Dans l'esprit d'une recherche appliquée, des approches didactiques ont ainsi pu être générées et/ou confirmées afin de fournir aux enseignant.e.s intéressé.e.s des principes didactiques empiriquement validés qu'elles/qu'ils pourront expérimenter dans leurs propres cours de langue.

Ouvrages cités

Burwitz-Melzer, Eva (2014): «Die Sprachkompetenz im Fremdsprachenunterricht angemessen üben und beurteilen». In: Burwitz-Melzer, Eva; Königs; Frank G.; Riemer, Claudia (éd.): *Perspektiven der Mündlichkeit. Arbeitspapiere der 34. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts.* Tübingen: Narr Francke Attempto Verlag, 17-28.

Cavelti, Stephanie; Derron, Véronique; Robertini, Claudia (2020): *Mille feuilles 6.3. Mais pourquoi ?* Berne: Schulverlag plus AG.

Ganguillet, Simone; Grossenbacher, Barbara; Lovey, Gwendoline; Sauer, Esther; Thommen, Andi; Trommer, Bernadette (2014): *Mille feuilles 6. 6.1: Eurêka – j'ai trouvé ! 6.2: Quelle question!* 1re édition. Berne : Schulverlag plus AG.

Lovey, Gwendoline (2024): *Interaktives Sprechen im lehrwerkbasierten Fremdsprachenunterricht der Grundschule. Reihe Romanistische Fremdsprachenforschung und Unterrichtsentwicklung*. Tübingen: Verlag Gunter Narr. Les annexes en ligne : https://files.narr.digital/9783381120314/Zusatzmaterial.pdf

Lütge, Christiane (2014): «Perspektiven für die Mündlichkeit im Fremdsprachenunterricht». In: Burwitz-Melzer et al. (éd.), 147-153.

Manoïlov, Pascale (2019): Repenser l'organisation des tâches pour favoriser le développement des interactions orales entre pairs. *Les langues modernes* 3/2019, 18-27.

Martinez, Hélène (2014): «Mündlichkeit – Zur Frage der Umsetzung wissenschaftlicher Er-kenntnisse und didaktischer Umsetzungskonzepte in die fremdsprachliche Unterrichtspraxis». In: Burwitz-Melzer, Eva; Königs; Frank G.; Riemer, Claudia (éd.): *Perspektiven der Mündlichkeit. Arbeitspapiere der 34. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts.* Tübingen: Narr Francke Attempto Verlag, 154-164.

Schramm, Karen; Aguado, Karin (2010): «Videographie in den Fremdsprachendidaktiken – Ein Überblick». In: Aguado et al. (Hg.), 185-214.

Tesch, Bernd (2010): *Kompetenzorientierte Lernaufgaben im Fremdsprachenunterricht. Konzeptionelle Grundlagen und eine rekonstruktive Fallstudie zur Unterrich*tspraxis (Französisch). Frankfurt a.M.: Peter Lang.

Thaler, Engelbert (2014): «Lehrersprache». *Praxis Fremdsprachenunterricht*, 01/2014, 15-16.

Wolff, Dieter (2002): *Fremdsprachenlernen als Konstruktion. Grundlagen für eine konstruktivistische Fremdsprachendidaktik*. Frankfurt a.M. et al.: Peter Lang.

Tema

« JE NE SAIS PAS », « JE N'SAIS PAS », « CH'SAIS PAS », « CHAIS PAS » … QUELLE PLACE POUR LA VARIATION PHONIQUE DANS L'ENSEIGNEMENT/APPRENTISSAGE DU FRANÇAIS LANGUE ÉTRANGÈRE ?

Obwohl Fachleute seit langem den Französischunterricht befürworten, welcher auf authentischem Sprachgebrauch basiert, ist die Rolle der Variation im Unterricht von Französisch als Fremdsprache immer eine Herausforderung. Für die Lernenden ergeben sich daraus Schwierigkeiten in der Wahrnehmung, aber auch in der Produktion. Diese veranschaulichen wir in diesem Beitrag anhand der Aneignung der Form "je ne sais pas" ("ich weiss nicht") und zeigen, inwiefern die Daten aus den Projekten "Phonologie du Français Contemporain" (PFC) und "InterPhonologie du Français Contemporain" (IPFC) einen interessanten Beitrag zum besseren Verständnis der Art und Weise leisten, wie Muttersprachler/innen die phonetische Variation in der gesprochenen Sprache erfassen (PFC-Daten) und wie sich die Lernenden diese Phänomene aneignen (IPFC-Daten).

Isabelle Racine
| Université de Genève

Isabelle Racine est professeure à l'Université de Genève, où elle dirige l'École de langue et de civilisation françaises (ELCF). Elle co-dirige également les projets « Phonologie du Français Contemporain » (PFC) et « InterPhonologie du Français Contemporain » (IPFC).

Dans un article paru il y a plus d'un demi-siècle et intitulé « Quel français enseigner ? », Coste (1969) soulignait déjà la nécessité, pour l'enseignement, de prioriser le français « contemporain », qu'il définissait comme un français devant prendre en considération non seulement « la langue et ses règles » mais également « les variations contrôlées ou attendues du discours », soit « les usages ». Si le caractère fondamentalement dynamique et hétérogène des pratiques langagières est maintenant bien attesté et décrit dans les travaux menés en linguistique, sa place dans l'enseignement du français reste un défi de taille (voir p. ex. Detey, 2017). Si cette problématique se pose évidemment pour l'écrit, c'est toutefois lorsque l'on s'intéresse à la langue orale qu'elle devient aigüe. Ainsi, on ne compte plus les publications, aux titres évocateurs – « If This is French, Then What Did I Learn in School ? » (Durán & McCool, 2003), « Pourquoi les Français ne parlent-ils pas comme je l'ai appris ? » (Weber, 2006) ou, plus récemment, « "Le français parlé ? eh ben j'savais pas ce que c'était !" : production et compréhension de la variation diaphasique en français parlé en FLE » (Surcouf & Ausoni, 2022). Celles-ci pointent le décalage entre l'oral modèle des cours de langue, dépourvu d'authenticité sociolinguistique et s'apparentant à un « no speaker's land » (Waugh & Fonseca-Greber, 2002), et la réalité bien plus diversifiée des formes attestées.

Si la prise en compte dans l'enseignement/apprentissage du français de la variation n'est certes pas aisée, comme le soulignent de nombreux chercheur-es, deux obstacles majeurs font consensus : d'une part, le manque de formation des enseignant-es – Saudan & Gajo (2022 : 39) relèvent en effet que le traitement linguistique de la variation est pointé comme une difficulté majeure dans la formation des futur-es enseignant-es de/en langue(s) (voir également à ce propos Lafine & Thomas, 2020, concernant les aspects lexicaux). D'autre part, l'attachement à la conception d'un standard dominant, que l'on retrouve tant chez les enseignant-es que chez les apprenant-es.

Saudan & Gajo (2022 : 39) relèvent en effet que le traitement linguistique de la variation est pointé comme une difficulté majeure dans la formation des futur-es enseignant-es de/en langue(s).

Celui-ci renvoie, selon Coste (2019 : 18), aux cultures éducatives et linguistiques associées historiquement et idéologiquement au français dans lesquelles prévalent « une vision normative focalisée sur les régularités systémiques bien plus que sur la variation », et ce dans les représentations comme dans les pratiques. Or, si, comme le rappelle Valdman (2000 : 648), « l'objectif déclaré de l'approche communicative est l'acquisition [...] d'une maitrise quasi native de la langue-cible », celle-ci ne peut pas faire l'impasse sur la dynamique et l'hétérogénéité des usages et des répertoires[1].

Fort-es de ces considérations pédagogiques, on doit toutefois se demander quelle est la réalité sur le terrain, soit quelles sont, concrètement, les difficultés que rencontrent les apprenant-es de FLE ? Surcouf & Ausoni (2022) apportent un début de réponse à cette question, à travers une expérience menée auprès de 19 apprenant-es de FLE, de niveau B2, en contexte universitaire homoglotte à Lausanne. La séquence « il devait y avoir je sais pas dix personnes un truc comme ça », produite [idvejavwaʁʃepadipɛʁsɔnɛ̃tʁykkɔmsa] et présentée oralement en classe (4 écoutes), n'a été transcrite correctement que par 2 apprenant-es sur 19. Or, les participant-es à cette étude, de 10 L1s différentes, comptabilisaient en moyenne 7 ans d'études du français et vivaient depuis plus de trois ans en milieu francophone. Les séquences [idvejavwaʁ] et [ʃepa], qui n'ont été identifiées correctement que par 7 et 4 apprenant-es respectivement, sont les plus problématiques. Ce ne sont donc pas tant le lexique ou la morphosyntaxe qui créent ce que les auteurs nomment des « trous d'air » dans la compréhension (Surcouf & Ausoni, 2022 : 146), mais bien des éléments phonético-phonologiques. En effet, dans la première séquence, on observe la chute de la consonne /l/ dans le pronom « il » devant un verbe à initiale consonantique (« devoir ») et la chute du schwa – ou E muet – dans « devait », [dve]. On retrouve également cette chute du schwa dans le « je » de la deuxième séquence. Le contact entre /ʒ/ et /s/ engendré par la chute du schwa induit un autre phénomène phonique, l'assimilation, soit l'adaptation de l'une des consonnes à l'autre – ici le /s/, produit sans vibrations des cordes vocales va influencer la prononciation du /ʒ/, produit avec vibrations et le transformer en sa contrepartie sourde, /ʃ/. La suite /ʃs/ se simplifie ensuite – ou s'assimile encore – pour ne garder que le premier élément /ʃ/. A cela s'ajoute encore un phénomène préalable, bien attesté dans les corpus oraux de français, la très fréquente chute du « ne » de négation. Au final, on constate donc que la forme écrite « je ne sais pas », à laquelle correspond une séquence qui peut être prononcée en 4 syllabes [ʒə-nə-sE-pa][2], est ici produite en 2, [ʃE-pa], moyennant des ajustements complexes, mais qu'elle peut aussi avoir d'autres réalisations, [ʒən-sE-pa], [ʒə-sE-pa] ou [ʃsE-pa].

Afin d'évaluer la nécessité – ou non – de familiariser les apprenant-es avec ce type de formes dans leur parcours d'apprentissage, il est nécessaire de s'intéresser à la fréquence à laquelle apparaissent ces différentes formes chez les natifs/ves. Si Surcouf et Ausoni (2022) ont déjà examiné cela dans deux corpus oraux de français, nous y ajoutons ici les données tirées de trois points d'enquêtes réalisés en Suisse romande – soit dans un contexte très proche de celui de leur étude –, dans le cadre du projet « Phonologie du Français Contemporain » (ci-après PFC, Detey et al., 2016), considéré comme une référence

1 Si le volume complémentaire du CECRL (Conseil de l'Europe, 2018) s'est affranchi de la notion de maitrise (quasi) native, cela ne change en rien notre propos ici, le but étant de pouvoir comprendre le français parlé au quotidien.

2 La notation [E] indique que le timbre de la voyelle de « sais » peut varier, en termes de degré d'aperture, sur un continuum dont les extrémités sont [e] et [ɛ], avec une préférence, pour une grande partie de la Suisse romande, pour [ɛ] (Racine, 2016). Cet élément ne constitue toutefois pas une variation centrale par rapport aux autres considérées ici.

Ce ne sont donc pas tant le lexique ou la morphosyntaxe qui créent ce que les auteurs nomment des « trous d'air » dans la compréhension (Surcouf & Ausoni, 2022 : 146), mais bien des éléments phonético-phonologiques.

en matière de variation phonique. Nous y avons examiné les réalisations de la séquence « je ne sais pas » – dont le rôle dans l'interaction a notamment été décrit par Pekarek Doehler (2016) – dans un sous-corpus constitué des données conversationnelles de 38 francophones suisses romand-es : 13 Neuchâtelois-es, 12 Vaudois-es et 13 Genevois-es. Sur les 273 occurrences que totalise ce sous-corpus, seules trois sont produites avec la présence du « ne » de négation, une fois sous la forme [ʒənəsEpa] et deux fois [ʒənsEpa]. Le « ne » est donc élidé dans 98.90 % des cas. Par ailleurs, la forme la plus fréquente est de loin [ʃEpa], produite dans 81.32 % des occurrences (n = 222), contre 11.36 % pour [ʒəsEpa] (n = 31) et 6.23 % pour [ʃsEpa] (n = 17).

Ainsi, au vu de la prépondérance de la forme [ʃEpa] dans le français parlé au quotidien, si les apprenant-es ne sont pas préparé-es à rencontrer cette forme, les « trous d'air » observés dans leur compréhension par Surcouf et Ausoni (2022) semblent inéluctables. Or, les données de Surcouf et Giroud (2016 : 9), concernant plus largement la chute du schwa, montrent que les ressources audio tirées de manuels ne permettent pas de sensibiliser les apprenant-es à ce type de réalisations. En effet, en se basant sur un corpus constitué d'enregistrements tirés de dix manuels généralistes couramment utilisés en FLE, les auteurs relèvent un taux de chute du schwa de 31 % seulement. Ces données contrastent drastiquement avec celles de Lyche (2016 : 9), qui observe un taux moyen de chute du schwa de 65 % en se basant sur les données PFC de 143 Français-es (5 enquêtes).

Un autre indice permettant d'examiner l'appropriation de la variation phonique consiste à s'intéresser aux productions des apprenant-es. En effet, même si perception et production impliquent des processus cognitifs fondamentalement différents, une analyse des réalisations peut aider à en savoir plus sur l'*input* auquel sont exposé-es les apprenant-es ainsi que sur son rôle dans ce processus d'appropriation. À ce titre, le projet « InterPhonologie du Français Contemporain » (ci-après IPFC, Detey et al., 2016), dont l'objectif est de constituer et d'analyser une large base de données d'apprenant-es de différentes L1s, collectées avec un protocole identique de recueil de données, fournit des pistes intéressantes.

En nous basant sur un sous-corpus d'IPFC-Suisse[3], nous avons examiné les réalisations de la séquence « je ne sais pas » chez 8 apprenant-es issu-es de deux populations distinctes : 4 italophones tessinoises et 4 tigrinyophones érythréens. Hormis l'un des tigrinyophones, un peu plus âgé – 32 ans –, tous/tes ont entre 21 et 23 ans. Leur profil d'apprentissage du français est en revanche sensiblement différent : les 4 Tessinoises ont commencé le français à l'école primaire et l'ont étudié jusqu'à la maturité, soit sur une durée allant de 10 à 13 ans, selon les cursus suivis, à côté de l'allemand et de l'anglais. Elles ont ensuite entamé des études d'orthophonie en français à l'Université de Neuchâtel, cadre dans lequel elles ont été enregistrées, après 18 mois dans ce cursus. Les 4 Erythréens ont commencé le français à leur arrivée à Genève, soit entre 2.5 et 3.5 ans avant l'enregistrement, en suivant des cours à raison de 3 heures hebdomadaires environ dans une institution étatique, une école privée ou un cadre associatif. En Erythrée, leur scolarité (primaire et secondaire) s'est déroulée en tigrinya et en anglais. Une étudiante en Master FLE

3 La collecte et l'analyse des données du sous-corpus IPFC-Suisse a bénéficié, de 2016 à 2020, du soutien du Fonds national de la recherche scientifique (projet no 169707). Le sous-corpus utilisé ici sera à terme rendu public via la base de données IPFC, actuellement en construction.

Au final, on constate donc que la forme écrite « je ne sais pas », à laquelle correspond une séquence qui peut être prononcée en 4 syllabes [ʒə-nə-sE-pa], est ici produite en 2, [ʃE-pa], moyennant des ajustements complexes, mais qu'elle peut aussi avoir d'autres réalisations, [ʒən-sE-pa], [ʒə-sE-pa] ou [ʃsE-pa].

de l'Université de Genève, elle-même d'origine érythréenne, les a recrutés et enregistrés.

Si le nombre d'occurrences de la séquence « je ne sais pas » est réduit dans ce sous-corpus – 28 au total, 20 chez les italophones et 8 chez les tigrinyophones –, la répartition n'est néanmoins pas dénuée d'intérêt, comme l'illustrent les tableaux ci-dessous :

	TI_1	TI_2	TI_3	TI_4
[ʒənsEpa]	5	-	-	-
[ʒəsEpa]	2	4	2	5
[ʃsEpa]	1	-	1	-

Tableau 1: Nombre d'occurrences et répartition des différentes réalisations de la séquence « je ne sais pas » (n = 20) chez les 4 apprenantes tessinoises (TI).

	ERY_1	ERY_2	ERY_3	ERY_4
[ʒəsEpa]	2	1	1	-
[ʃEpa]	-	3	-	1

Tableau 2: Nombre d'occurrences et répartition des différentes réalisations de la séquence « je ne sais pas » (n = 8) chez les 4 apprenants érythréens (ERY).

On constate tout d'abord que la forme la plus proche de l'écrit, [ʒənəsEpa], n'a jamais été réalisée. Aucune forme avec le « ne » de négation n'est d'ailleurs produite par les tigrinyophones, alors qu'il s'agit de la forme dominante chez TI_1, qui produit à 5 reprises [ʒənsEpa], réalisation qui n'apparaissait qu'à 2 reprises sur les 273 occurrences de notre corpus de natifs/ves suisses romands. À l'inverse, la forme la plus élidée, [ʃEpa], également la plus courante chez les natifs/ves avec 81.32 % des occurrences, n'est présente que chez 2 tigrinyophones, notamment chez ERY_2, qui la réalise à trois reprises. Deux apprenantes tessinoises, TI_1 et TI_3, produisent en revanche la forme [ʃsEpa], présente chez les natifs/ves dans 6.23 % des cas, réalisation qui n'apparait pas chez les tigrinyophones. Enfin, la forme la plus souvent produite par les 8 apprenant-es est [ʒəsEpa], qui totalise 17 des 28 réalisations, mais qui ne représentait que 11.36 % des occurrences natives.

Ce début d'analyse est révélateur à différents niveaux. On peut premièrement constater que, malgré une répartition des formes produites par les 8 apprenant-es encore éloignée de celle des natifs/ves, un début d'appropriation des formes variées semble présent dans nos données, mais à des degrés divers selon les phénomènes. En effet, si la variation morpho-syntaxique, représentée par l'absence du « ne » de négation, semble acquise chez 7 des 8 apprenant-es, la chute du schwa ne l'est en revanche pas, et la double assimilation présente dans la forme [ʃEpa], largement préférée par les natifs/ves, l'est encore moins.

Deuxièmement, le comportement différencié des deux populations testées, qui diffèrent à la fois par leur parcours – étudiantes universitaires suisses vs appprenants érythréens issus de la migration – et donc aussi par la manière dont ils/elles sont entré-es dans le français, est intéressant. On observe en effet que, même si les deux populations se rejoignent sur la réalisation [ʒəsEpa], la répartition des autres formes n'est pas identique, avec une prépondérance de formes plus normées (avec le [n] de négation et/ou sans chute du schwa) chez

Ainsi, au vu de la prépondérance de la forme [ʃ]pa] dans le français parlé au quotidien, si les apprenant-es ne sont pas préparé-es à rencontrer cette forme, les « trous d'air » observés dans leur compréhension par Surcouf et Ausoni (2022) semblent inéluctables.

les Tessinoises que chez les Erythréens, qui semblent s'approcher davantage d'une langue orale authentique.

Ces constatations ne sont pas surprenantes, les spécialistes ayant pointé du doigt le contexte de classe comme actuellement insuffisant pour l'appropriation des phénomènes variables du français, alors qu'un séjour en milieu francophone constitue un élément déterminant dans l'amélioration de la maitrise de la gestion de ceux-ci. Nos données de production et les données de Surcouf et Ausoni (2022) en perception permettent toutefois de constater qu'être dans un « bain » linguistique, même de longue durée – plus de 3 ans chez Surcouf & Ausoni (2022) et entre 18 mois et 3.5 ans pour les italophones et les tigrinyophones – ne suffit pas. Les chercheur-es se sont récemment intéressé-es de plus près à ce qui se passe pendant le séjour, en observant notamment la socialisation des apprenant-es durant leur séjour, avec l'idée que leur progression est fortement liée à leur degré d'implication socio-langagière avec des natifs/ves (pour des études sur le français, voir entre autres Chamot et al., 2021 ; Gautier & Chevrot, 2017 ; Kennedy Terry, 2017 ; Thomas & Mitchell, 2022).

Les données présentées dans cette contribution plaident ainsi pour un apprentissage davantage basé sur les usages réels, tel que l'envisageait déjà Coste (1969), y compris en contexte scolaire. À ce titre, on voit ici tout l'intérêt de l'apport des corpus oraux, que ce soit pour cerner finement les difficultés des apprenant-es ou pour mieux documenter et comprendre les usages réels des francophones dans toute leur diversité. Les corpus de natifs/ves peuvent aussi se convertir en ressources proposant à la fois un *input* authentique et des activités propices à une approche du français parlé au quotidien (voir, à ce sujet, tout le travail sur la didactisation des corpus oraux mené notamment par Virginie André – Fleuron –, Carole Etienne – CLAPI-FLE–, Marie Skrovec – ESLO-FLEU – et l'équipe de PFC-EF)[4]. Ce travail contribue ainsi à apporter des éléments de réponse significatifs à la question posée par Coste il y a plus d'un demi-siècle, « Quel français enseigner ? », qui reste toujours encore aujourd'hui un défi de taille.

4 Pour Fleuron, voir https://fleuron.atilf.fr/ ; pour CLAPI-FLE et CORAIL, voir http://clapi.icar.cnrs.fr/FLE/ et http://clapi.icar.cnrs.fr/Corail/ ; pour PFC-EF, voir https://www.projet-pfc.net/le-projet-pfc-ef/ et pour ESLO-FLEU, voir https://www.ortolang.fr/market/corpora/eslo-fleu.

Références

Chamot, M., Racine, I., Regan, V. & Detey, S. (2021). Une ou des immersion(s) ? Regard sur l'acquisition de la compétence sociolinguistique par des apprenants anglophones irlandais de FLE. In : E. Puskta (éd.), *La prononciation du français langue étrangère. Perspectives linguistiques et didactiques*. Tübingen : Narr Francke Attempto Verlag, pp. 133-161.

Conseil de l'Europe (2018). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer. Volume complémentaire avec de nouveaux descripteurs*. Disponible en ligne : https://rm.coe.int/cecr-volume-complementaire-avec-de-nouveaux-descripteurs/16807875d5.

Coste, D. (1969). Quel français enseigner ? *Le français dans le monde*, 65, 12-18.

Coste, Daniel (2019). Le plurilinguisme entre variation et évaluation. In : L. Gajo, J.-M. Luscher, I. Racine & F. Zay (éds), *Variation, plurilinguisme et évaluation en français langue étrangère*. Berne : Peter Lang, pp. 15-24.

Detey, Sylvain (2017). La variation dans l'enseignement du français parlé en FLE : des recherches linguistiques sur la francophonie aux questionnements didactiques sur l'authenticité. In : J. An-Chyun, B. Montoneri & M.-J. Maître (éds), *Échanges culturels aujourd'hui : langue et littérature*. New Taipei City : Tamkang University Press, pp. 93-114.

Detey, S., Durand, J., Laks, B. & Lyche, C. (eds) (2016). *Varieties of Spoken French*. Oxford: Oxford University Press.

Detey, S., Racine, I., Kawaguchi, Y. & Zay, F. (2016). Variation among non-native speakers: The Interphonology of Contemporary French. In: S. Detey, J. Durand, B. Laks & C. Lyche (eds.), *Varieties of spoken French*. Oxford: Oxford University Press, pp. 491-502.

Durán, R. and McCool, G. (2003). If this is French, then what did I learn in school?, *The French Review*, 77 (2), 288-299.

Gautier, R. & Chevrot, J.-P. (2015). Social networkds and acquisition of sociolinguistic variation in a study abroad context: A preliminary study. In: R. Mitchell, N. Tracy-Ventura & K. McManus (eds), *Social interaction, identity and language learning during residence abroad*. Amsterdam: The European Second Language Association, pp. 169-184.

Kennedy Terry, K. (2017). Contact, context, and collocation: The emergence of sociostylistic variation in L2 French learners during study abroad, *Studies in Second Language Acquisition*, 39 (3), 553-578, https://doi.org/10.1017/S0272263116000061.

Lafine, J. et Thomas, A. (2020). La Suisse romande dans les manuels de FLE à l'école obligatoire en Suisse alémanique, *Babylonia 1-2020*, 34-43.

Lyche, C. (2016). Approaching variation in PFC: The schwa level. In: S. Detey, J. Durand, B. Laks & C. Lyche (eds), *Varieties of Spoken French*. Oxford: Oxford University Press, pp. 352-362.

Pekarek Doehler, S. (2016). More than an epistemic hedge: French je ne sais pas 'I don't know' as a resource for the sequential organization of turns and actions, *Journal of Pragmatics*, 106, 148-162.

Racine, I. (2016). Le français en Suisse. In: S. Detey, I. Racine, Y. Kawaguchi & J. Eychenne (éds), *La prononciation du français dans le monde : du natif à l'apprenant*. Paris : CLE International, pp. 44-48.

Saudan, V. et Gajo, L. (2022). Les francophonies – un concept en émergence pour la formation des enseignant-es et les contextes de la migration. Quelques remarques préliminaires concernant le projet d'élaboration d'une didactique de la francophonie pluricentrique, plurilingue et plurielle. In : M. Causa et S. Richard (éds), *Pour une francophonie plurielle, plurilingue et pluricentrique*. Paris : L'Harmattan, pp. 23-46.

Surcouf, C. & Ausoni, A. (2022). « Le français parlé ? eh ben j'savais pas ce que c'était ! » : production et compréhension de la variation diaphasique en français parlé en FLE, *Mélanges Crapel*, 43 (1), 130-156.

Surcouf, C. & Giroud, A. (2016). À quelle langue accède l'apprenant ? Examen critique du traitement de l'oral dans les premières leçons de manuels de français langue étrangère, *Linguistik online*, 78 (4), https://doi.org/10.13092/lo.78.2947.

Thomas, A. and Mitchell, R. (2022). Can variation in input explain variation in typical spoken target-language features during study abroad? *Journal of the European Second Language Association*, 6 (1), 60-77, https://doi.org/10.22599/jesla.91.

Valdman, A. (2000). Comment gérer la variation dans l'enseignement du français langue étrangère aux États-Unis, *The French Review*, 73 (4), 648-666.

Waugh, L. R. et Fonseca-Greber, B. (2002). Authentic Materials for Everyday Spoken French: Corpus Linguistics vs. French Textbooks, *Arizona Working Papers in SLAT 9*, 114-127.

Weber, C. (2006). Pourquoi les Français ne parlent-ils pas comme je l'ai appris ? *Le Français dans le monde*, 345, 31-33.

Tema

LES CORPUS COMME INPUT ET COMME OUTPUT : L'EXEMPLE DES MARQUEURS BON ET BIEN

This article presents some results of the DiCoi project, which aims to create teaching material from oral corpora of French and to study the development of the interactional competence of learners of L2 French in vocational training for manual professions. We describe the main lines of a teaching intervention on the discourse markers *bon* and *bien* and the production of these markers in the longitudinal (two years) L2 French learner corpus DiCoi, including free interactions between peers. The results show that the use of *bien* is more productive than that of *bon*, as well as an increase in the frequency and diversity of the functions of the two markers among learners. The discussion comes back to the use of corpora as input and output.

Anita Thomas
| Université de Fribourg
France Rousset
| Université de Genève

Anita Thomas est professeure au département de plurilinguisme et didactique des langues étrangères à l'Université de Fribourg.

France Rousset est doctorante à l'Université de Genève et a été collaboratrice scientifique dans le projet DiCoi.

1. Introduction

L'objectif de cet article est double. Premièrement, il présentera une démarche didactique sur les marqueurs discursifs (MD) *bon* et *bien* développée à partir de corpus de français parlé puis testée dans dix classes de jeunes en formation professionnelle de métiers manuels en Suisse romande. Deuxièmement, il présentera la production de ces deux marqueurs par seize apprenant·e·s ayant le français langue seconde (L2) à partir d'une analyse des données du corpus longitudinal d'interactions libres récolté lors des interventions en classe. Ces deux types de corpus, comme input et comme output, sont au cœur du projet de recherche appliquée DiCoi, dans le cadre duquel l'étude a été menée.

Dans cet article, nous traiterons les questions suivantes : 1) comment enseigner les marqueurs discursifs – *bon* et *bien* – polysémiques et polyfonctionnels avec des corpus oraux de français ? 2) quel est le développement longitudinal de la production de ces deux MD par les apprenant·e·s L2 dans des interactions libres ? Pour ce faire, nous aborderons l'utilisation des corpus en didactique du FLE avant de présenter les grandes lignes du projet DiCoi. Nous nous concentrerons ensuite sur la didactisation des MD *bon* et *bien* dont nous présenterons les principaux emplois. Nous montrerons ensuite comment les apprenant·e·s utilisent ces deux MD au cours de deux années puis nous terminerons sur une discussion.

2. Utilisation des corpus comme ressource didactique

L'utilisation des corpus de langue parlée en didactique des langues étrangères s'est intensifiée au cours de ces dernières années en raison de leur mise à disposition grâce à internet. Ces bases de données comprenant à la fois des enregistrements audios et leur transcription permettent d'exposer les apprenant·e·s à la langue telle qu'elle est réellement utilisée (Boulton & Tyne, 2014). Parmi

les corpus oraux de français existants, nous pouvons citer le corpus FLEURON (https://fleuron.atilf.fr) qui comprend des interactions principalement administratives en milieu universitaire, mais aussi quotidiennes et le corpus CLAPI (http://clapi.icar.cnrs.fr) qui est composé de différents types d'interactions allant du privé au professionnel. Ce corpus a donné lieu à deux plateformes didactisées – CLAPI-FLE (http://clapi.ish-lyon.cnrs.fr/FLE/accueil.php) et CORAIL (http://clapi.icar.cnrs.fr/Corail/index.html) – toutes deux proposent des activités ciblant plusieurs phénomènes interactionnels fréquents. Finalement, le corpus OFROM (https://ofrom.unine.ch) comprend différents types d'interactions faisant intervenir des locuteur·trice·s de Suisse romande. Nous avons utilisé ces corpus et ressources didactisées pour construire notre propre matériel didactique.

3. Le projet de recherche DiCoi

Le projet de recherche appliquée DiCoi (digitalisation – corpus – interaction, https://centre-plurilinguisme.ch/fr/recherche/Dicoi) a été financé par le Centre de compétence sur le plurilinguisme (2021–2024). Il a pour objectif de tester l'apport des corpus de français parlé comme ressource didactique dans l'enseignement de la compétence d'interaction en français L2 et de décrire le développement longitudinal sur deux ans d'apprenant·e·s en formation professionnelle ayant un niveau intermédiaire (B1 – B2) de français (Thomas & Rousset, 2023). Le projet a été mené dans dix classes mixtes d'élèves ayant le français depuis la naissance (37 élèves francophones) ou comme L2 (30 élèves apprenant·e·s). Ces élèves suivent un apprentissage de base (attestation de formation professionnelle, AFP) pour des métiers manuels dans les secteurs de la restauration, la cuisine, la construction, la confection, la menuiserie, l'automobile ou encore du sanitaire.

3.1 Matériel didactique

Les interventions didactiques à partir des corpus ont été réparties sur deux ans. Nous avons réalisé huit interventions et onze exercices sous format numérique, que nous avons testés de manière empirique dans les dix classes. Les exercices ont été construits à l'aide du logiciel H5P (https://h5p.org) qui permet de présenter les exercices sous forme de livre numérique. Nous avons ainsi construit des vidéos comprenant des inputs théoriques, des activités de compréhension orale et des analyses accompagnées d'exercices interactifs de différents types (vrai/faux, choix multiples, *drag and drop*, etc.). Les exercices ont été construits de sorte que les élèves puissent aller à leur rythme, réécouter certains extraits, afficher ou non les transcriptions ou encore refaire les activités. Tous les extraits proviennent de corpus, pour la plupart tirés des bases de données susmentionnées, et ont été intégrés soigneusement en tenant compte des difficultés des élèves, en particulier des apprenant·e·s L2 (vocabulaire et syntaxe accessible, élocution, input oral et écrit en parallèle, etc.). Le matériel didactique complet (y compris le déroulement des interventions en classe) est disponible en libre accès sur la plateforme https://dicoi.ch.

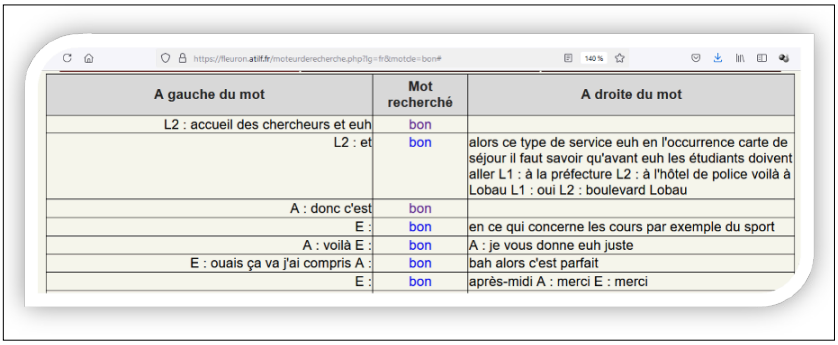
Les thématiques abordées dans le matériel didactique visent principalement la compétence d'interaction, par exemple les marqueurs discursifs *genre, trop, juste, bon* et *bien,* mais aussi des situations langagières professionnelles comme les interactions au travail ou l'entretien d'embauche.

Les séquences didactiques ont eu pour but de familiariser les élèves à l'utilisation des corpus et de leur permettre de gagner en indépendance à ce niveau-là en développant des compétences transversales notamment en informatique.

3.2 Corpus longitudinal d'interactions libres

Dans le but de documenter le développement longitudinal des apprenant·e·s du français L2, nous avons réalisé un enregistrement audio des élèves au moyen d'un dictaphone au début de chacune des huit interventions. Nous leur avons demandé de discuter librement deux par deux durant environ dix minutes sur les sujets de leur choix, tout en leur suggérant quelques thèmes, par exemple leur travail ou l'actualité[1]. Les données ont ensuite été transcrites à l'aide du logiciel CLAN (https://dali.talkbank.org/clan). Ce corpus sera disponible sur le corpus SWIKO (https://ifm-swiko.unifr.ch, accès sur demande) qui rassemble les données d'apprenant·e·s de nombreux projets qui ont été menés à l'institut de plurilinguisme (voir par exemple Karges et al., 2020).

1 Dans la mesure du possible, nous avons essayé de garder les mêmes paires d'apprenant·e·s, mais selon les classes nous avons été obligés de changer les binômes (absences, nombre impair, pas d'autre apprenant·e, etc.).



A gauche du mot	Mot recherché	A droite du mot
L2 : accueil des chercheurs et euh	bon	
L2 : et	bon	alors ce type de service euh en l'occurrence carte de séjour il faut savoir qu'avant euh les étudiants doivent aller L1 : à la préfecture L2 : à l'hôtel de police voilà à Lobau L1 : oui L2 : boulevard Lobau
A : donc c'est	bon	
E :	bon	en ce qui concerne les cours par exemple du sport
A : voilà E :	bon	A : je vous donne euh juste
E : ouais ça va j'ai compris A :	bon	bah alors c'est parfait
E :	bon	après-midi A : merci E : merci

Figure 1

Extrait du concordancier de FLEURON pour *bon*

Etape 4

Complétez au fur et à mesure le tableau ci-dessous selon le modèle (= extrait 1).

	voisins de *bon*	fonction (= à quoi ça sert ?)
1	et euh bon	gagner du temps quand on réfléchit à ce qu'on va dire ensuite, souvent avec *euh*
2		
3		

Figure 2

Tableau à remplir par les élèves à partir du concordancier

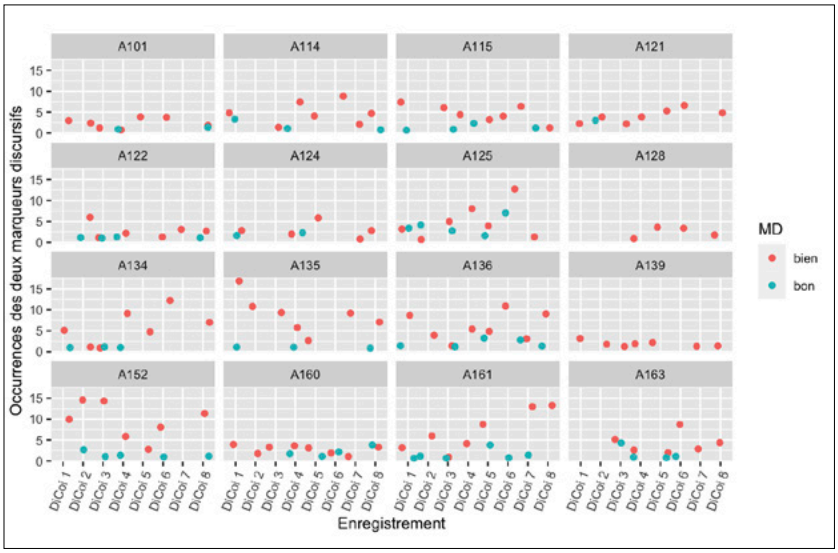


Figure 3

Fréquence d'utilisation de *bon* et *bien* par les apprenant·e·s

4. *bon* et *bien* : deux marqueurs polysémiques et polyfonctionnels

Plusieurs études se sont intéressées aux fonctions de ces deux MD tant à l'écrit qu'à l'oral en interaction. Peltier et Ranson (2020) proposent un état des lieux des analyses antérieures sur *bon* en établissant une nouvelle catégorisation des fonctions. Cette classification recense neuf fonctions textuelles (par exemple introduction d'un nouveau thème, résultat, formulation, ...) et deux fonctions touchant à l'attitude (contraste et résignation). Il en ressort aussi que *bon* est souvent accompagné d'autres MD. On retrouve notamment des collocations comme *mais bon* connotant souvent un contraste, *euh bon* dans des situations de reformulation / hésitation, *bon alors* pour introduire un nouveau thème, ou encore *ah bon* dans des acceptations. Finalement, *bon* permet à la fois d'initier et de clore une séquence (Lefeuvre, 2011).

Quant à *bien,* de nombreuses études se sont penchées sur ses emplois (voir par exemple Moline, 2012 pour une vue d'ensemble). En effet, *bien* peut avoir une valeur d'intensité, de (demande de) confirmation, d'approximation ou encore d'atténuation (Mosegaard Hansen, 1998 Parmi les fonctions recensées de *bien* en tant que MD dans les interactions, on relève celles d'introduire un nouveau thème, d'introduire une réponse, souvent avec *eh bien*, de valider le tour précédent ou de clore une séquence, souvent avec *très bien*.

5. *bon* et *bien* : séquence didactique

A partir des études antérieures, plusieurs choix didactiques ont été faits. Nous avons mis en place une séquence didactique d'une durée d'une heure réalisée lors de la troisième intervention en classe. Ainsi, les élèves possédaient déjà quelques connaissances sur les corpus, puisqu'ils avaient travaillé dessus durant les deux premières interventions. Dans cette optique, nous avons décidé de suivre l'approche du *data-driven learning* (André, 2019 ; Johns, 1991) et de faire travailler les élèves directement sur les données du corpus FLEURON (Figure 1). Nous avons porté notre choix sur ce corpus car il propose un concordancier permettant d'accéder directement aux interactions et

il était déjà connu des élèves lors d'inputs précédents.

Après avoir rappelé son utilisation, nous avons entamé l'analyse du MD *bien* en plénum afin d'activer la démarche d'analyse des interactions. En étant semi-guidés, les élèves ont ensuite travaillé en autonomie afin de relever et d'identifier les différentes fonctions de *bon* dans le corpus FLEURON en les reportant sur une feuille (Figure 2). Nous avons ensuite procédé à une mise en commun en plénum ouvrant une discussion sur leurs pratiques langagières quotidiennes.

A la fin de l'intervention, nous avons récolté les feedbacks des élèves sur l'activité. Il en est ressorti que la majorité des élèves a apprécié de travailler sur ce corpus qu'ils ont jugé facile d'utilisation et utile. Néanmoins, la tâche d'identification des fonctions a été perçue comme étant difficile tout en ayant permis une prise de conscience de la polysémie et de la polyfonctionnalité de ces deux MD utilisés et entendus au quotidien.

6. Développement longitudinal de *bon* et *bien* par les apprenant·e·s L2

Rappelons que les enregistrements ont eu lieu au début de chaque intervention pour une durée de 10 minutes sous la forme d'interactions libres entre paires. Ces interactions ont été transcrites dans CLAN, permettant de réaliser différents types d'analyses. Dans cet article, nous nous focalisons sur les données longitudinales de 16 apprenant·e·s. Ces apprenant·e·s ont été sélectionnés parce qu'ils ont participé à tous les enregistrements (ou en ont manqué maximum un) et que les enregistrements étaient transcrits au moment de l'analyse. Comme le montre le tableau 1, les apprenant·e·s ont des L1 non-européennes et la plupart sont des hommes.

Tout d'abord, nous avons effectué une recherche concernant la fréquence d'utilisation des deux MD par les seize apprenant·e-s. Les résultats sont présentés dans la Figure 3. L'axe horizontal correspond aux huit enregistrements et l'axe vertical indique le nombre d'occurrences pour chaque MD. A noter que A121 était absent à DiCoi7 et A124 à DiCoi6, A125 à DiCoi8.

Code	Classe	L1	Genre	Années en Suisse (2021)	Code	Classe	L1	Genre	Années en Suisse (2021)
A101	01	Tigrinya	F	7	A134	05	Tigrinya	H	4
A114	02	Tigrinya	H	4	A135	05	Tigrinya	H	3
A115	02	Tigrinya	H	6	A136	05	Tigrinya	H	7
A121	03	Tigrinya	H	7	A139	05	Ouzbek	H	6
A122	03	Tigrinya	H	5	A152	06	Tibétain	F	11
A124	04	Tigrinya	H	6	A160	07	Tigrinya	H	8
A125	04	Tigrinya	H	5	A161	07	Tigrinya	H	8
A128	04	Tigrinya	H	4	A163	07	Tigrinya	H	8

Tableau 1
Apprenant·e·s de l'étude

Au total, nous avons relevé 89 occurrences pour *bon* et 511 pour *bien*. Dans FLEURON *bien* était également plus fréquent que *bon* (226 *bon* et 425 *bien*). Ainsi la plus faible fréquence de *bon* dans les données correspond à ce que l'on trouve en L1. On note tout de même que A128 et A139 n'ont par exemple jamais utilisé *bon* dans leurs discussions sur la période des deux années. Toutes les productions des deux marqueurs sont correctes.
Rappelons que l'intervention sur *bon* et *bien* a eu lieu lors de l'intervention 3. Nous observons l'utilisation de ces MD par certain·e·s apprenant·e·s déjà avant l'input. C'est par exemple le cas pour A114, A121, A124 ou A135. Nous n'observons pas de forte augmentation de l'utilisation des MD lors de l'enregistrement 4, après l'input. Bien qu'il semble y avoir un développement pour A125 et A161 au fil du temps, tendant vers une utilisation plus accrue de ces deux MD, force est de constater que la fréquence ne permet pas de discerner de réels effets de l'input, ni d'observer un développement dans le temps. En outre, la variation au niveau de la fréquence d'utilisation peut être influencée par divers facteurs, en particulier le sujet de discussion et l'engagement des participant·e·s.

6.1 Développement des fonctions de *bon*

Ceci nous a amené à procéder à l'annotation des fonctions et structures dans les transcriptions comme dans les exemples (1) et (2). Dans l'extrait (1), A125 clôt la séquence en cours. Il utilise *voilà* pour indiquer qu'ils se dirigent vers la fin d'une séquence, suivi d'une pause et de *bon merci beaucoup* venant la clore. Dans l'extrait (2), une séquence sur la danse se clôt par

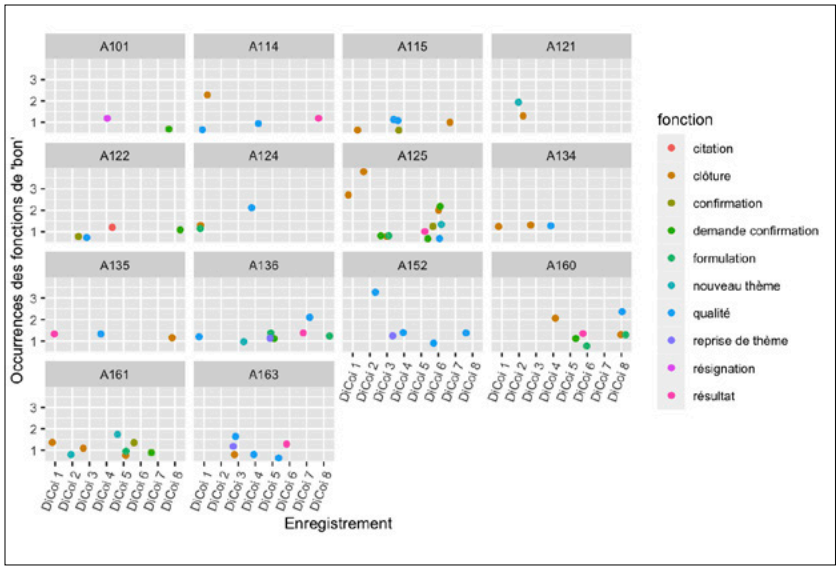


Figure 4
Fonctions de *bon* produites par les apprenant·e·s

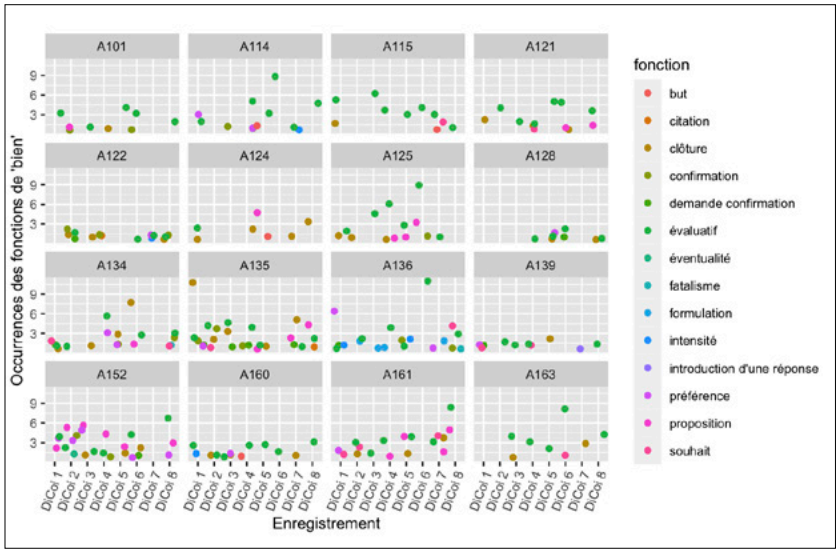


Figure 5
Fonctions de *bien* produites par les apprenant·e·s

(1) Enregistrement 1 : A124 et 125
*A125 : aussi sur internet (..) xxx voilà (..) **bon** (..) merci beaucoup .
%com : MD_bon_clôture
*A124 : merci au'voir .

(2) Enregistrement 5 : A160 et 161
*A161 : mais tu danses quand même à la maison ou pas .
*A160 : nan &=rires je connais pas danse .
*A161 : &=rires ok tu danses pas à la maison .
*A160 : +< hm hm .
*A161 : ok ok (.) bon t' as fait quoi d'autre pendant le week-end .
%com : MD_bon_nouveau_thème
*A160 : euh je rien sauf sortir discuter avec des amis et euh (..) hm : un peu discuté avec mon frère .

ok ok produit par A161. Après une courte pause, A161 propose un nouveau thème de discussion et l'introduit grâce à *bon*.

Le résultat de l'analyse des fonctions de *bon* est présenté dans la Figure 4. Pour mémoire, les apprenants A128 et A139 n'ont pas produit d'occurrences de *bon*. On constate que les fonctions de clôture et de qualité sont présentes dès le premier enregistrement chez plusieurs locuteurs comme A114, A125 ou encore A161. Néanmoins, on observe un développement de l'utilisation de *bon* à partir de l'enregistrement 4 pour A160. De plus, A125, A136, A160 et A161 tendent à l'utiliser dans un nombre croissant de fonctions au cours du temps (fonctions de clôture, formulation, résultat, reprise de thème, nouveau thème ou encore de confirmation / demande de confirmation).

6.2 Développement des fonctions de bien

Nous avons procédé de manière similaire pour l'analyse de la production de *bien*. Deux exemples du codage des fonctions de *bien* sont présentés en (3) et (4).

Dans l'extrait (3), A134 demande confirmation avec *français*, validée par A139 avec *ouais*. A134 valide ensuite le tour de A139 et se dirige vers la fin de la séquence sur l'importance du français en utilisant *c'est bien*, ce que nous avons codé comme une fonction de clôture. Dans l'extrait (4), il s'agit de l'expression fixe *ou bien* en fin de tour de parole.

Le résultat de l'analyse des fonctions de *bien* est présenté dans la Figure 5. On observe que *bien* est utilisé dans plus de fonctions différentes comparé à *bon* et ce par l'ensemble des participant·e·s. Les fonctions présentes dès les premiers enregistrements sont surtout celles d'évaluation et de clôture. Bien que certain·e·s apprenant·e·s produisent *bien* dans une variété de fonctions (évaluation, souhait, reformulation, confirmation/demande de confirmation, préférence) déjà avant notre intervention 3, en particulier A135, A136, A152 et A161, on voit que la diversification des fonctions se fait dès le quatrième enregistrement. Même si ce développement n'est pas forcément lié à notre intervention didactique, on ne peut exclure qu'elle a contribué à un développement de la production de *bien* en rendant saillant les diverses utilisations.

(3) Enregistrement 1 : A134 et A139
*A139 : quelle langue que (.) j'aimerais parler ?
*A134 : tu veux : tu aimerais (.) t' &-aim- t' aimerais bien apprendre .
*A139 : ouais j' aimerais bien apprendre (.) ouais .
*A134 : +< hm (.) .
*A139 : ben celui que le français pour l' instant ouais c' est l'important (.) ici c' est le français .
*A134 : français ouais .
*A139 : +< ouais (.) .
*A139 : pour mon formation aussi (.) ouais (.) hm (.) &-qu- ouais .
*A134 : français xxx (.) .
*A139 : ouais .
*A134 : ouais c'est **bien** .
%com : MD_bien_clôture

(4) Enregistrement 5 : A124 et A125
*A124 : nan nan nan jamais jamais moi j' apprends toujours j' apprends j'apprends .
*A125 : jusqu' à mort ?
*A124 : jusqu' à mourir ouais .
*A125 : ça va ou bien .
%com : MD_bien_expression

Les graphiques de la Figure 5 visualisent en outre un développement dans le temps pour plusieurs apprenant·e·s, A124, A134, A163 et plus faiblement A128. Les apprenant·e·s A135, A136 et A161 utilisent beaucoup *bien* et dans différentes fonctions.

7. Discussion

Dans cette contribution nous avons présenté un exemple d'utilisation des corpus comme ressource pour un public d'apprenant·e·s L2 du français qui se forme à des métiers manuels. Bien que l'utilisation des corpus dans une approche de type *data-driven learning* exige une certaine réflexion métalinguistique, il semble que ce genre de tâche soit faisable avec un public non académique. Ce résultat confirme les études antérieures menées par André (par exemple André, 2019).

Quant à l'effet de notre intervention, les données récoltées ne nous permettent pas d'établir un lien direct entre notre travail sur *bon* et *bien* dans les classes et la production de ces deux MD par les seize apprenant·e·s de cette étude. Néanmoins, les données longitudinales montrent un certain développement au niveau de la diversité des fonctions utilisées et ce même pour *bon* qui est plus rare dans les données. En ce sens l'étude suggère qu'une approche sur corpus pour un phénomène interactionnel fréquent en français permet d'augmenter les processus attentionnels.

Le recours à des enregistrements d'interaction libres comme méthode pour l'observation du développement longitudinal d'apprenant·e·s comprend à la fois des forces et des faiblesses. La force d'un corpus d'interactions libres réside dans l'authenticité des données récoltées. Les apprenant·e·s parlent de leur quotidien et de ce qui les intéresse. La faiblesse de la méthode réside dans l'absence de contrôle des connaissances, comme on l'aurait par le biais de tests d'élicitation ciblés.

Contrairement à des apprenant·e·s de classes de langues à l'école obligatoire ou post-obligatoire, les apprenant·e·s du projet DiCoi évoluent dans un environnement riche au niveau de l'input du français parlé. Une entrée dans la langue par l'oral, sans passer par les manuels traditionnels fortement influencés par les règles de l'écrit, pourrait expliquer l'utilisation relativement productive de MD dès les premiers enregistrements. Les données de ce corpus décrivent donc principalement un développement informel du français L2.

Remerciements

Le projet DiCoi a été financé par le Centre scientifique de compétence sur le plurilinguisme (Fribourg, Suisse). Il a été réalisé en partenariat didactique avec Centres de Formation Professionnelle de l'État de Fribourg (CD-CFP). Nous remercions les enseignant·e·s et les élèves ayant accepté de participer au projet ainsi que les assistant·e·s de recherche pour leur soutien lors de la récolte des données.

Références

André, V. (2019). Pourquoi faire de la sociolinguistique des interactions verbales avec des enseignants et des apprenants de Français Langue Étrangère ? *Linx. Revue des linguistes de l'université Paris X Nanterre, 79*. https://doi.org/10.4000/linx.3694

Boulton, A., & Tyne, H. (2014). *Des documents authentiques aux corpus : Démarches pour l'apprentissage des langues*. Didier.

Johns, T. (1991). *Should you be persuaded: Two samples of data-driven learning materials. English Language Research Journal, 4, 1–16.*

Karges, K., Studer, T., & Wiedenkeller, E. (2020). Textmerkmale als Indikatoren von Schreibkompetenz. *Bulletin Suisse de Linguistique Appliquée*, 117-140.

Lefeuvre, F. (2011). Bon et quoi à l'oral : Marqueurs d'ouverture et de fermeture d'unités syntaxiques à l'oral. *Linx. Revue des linguistes de l'université Paris X Nanterre, 64-65*, 223-240. https://doi.org/10.4000/linx.1417

Moline, E. (2012). Aperçu des emplois de bien en français contemporain. *Travaux de linguistique, 65*(2), 7-26. https://doi.org/10.3917/tl.065.0007

Mosegaard Hansen, M.-B. (1998). *The Function of Discourse Particles : A Study with Special Reference to Spoken Standard French*. John Benjamins Publishing.

Peltier, J. P. G., & Ranson, D. L. (2020). Le marqueur discursif bon : Ses fonctions et sa position dans le français parlé. *SHS Web of Conferences, 78*, 01006. https://doi.org/10.1051/shsconf/20207801006

Thomas, A., & Rousset, F. (2023). Utilisation des corpus pour l'enseignement de l'interaction en formation professionnelle de métiers manuels : Exemple d'un exercice numérique sur « genre ». *Corpus, 24*. https://doi.org/10.4000/corpus.7899

Tema

‘FRAPPER’ OU ‘JETER SUR’, COMMENT CHOISIR ? APPORT DE L’ANALYSE DE CORPUS EXPÉRIMENTAUX [1]

Teachers and researchers have shown the heterogeneity of the lexical competence. Recent studies have highlighted didactic proposals involving explicit learning around a single thematic domain, based on linguistic models, notably referring to movement verbs (Garcia-Debanc & Aurnague 2020). We offer a corpus method to study the distribution of verbs of another thematic domain -- namely collision -- and to study the semantic acquisition of L2 verbs. This study opens potentially new ways to study and teach vocabulary in an additional language.

Mireille Copin
| Université Toulouse Jean Jaurès
Inès Saddour
| Université Toulouse Jean Jaurès

Introduction

Lorsqu’un apprenant de français L2 produit l’énoncé suivant « y a une [dame] qui est frappé l’autre par sa balle aussi » lors d’une description de vidéo (Image 1), l’enseignant pourrait se demander s’il doit simplement corriger l’erreur de préposition (*avec* plutôt que *par*) ou s’il doit plutôt suggérer d’employer un autre verbe, comme *jeter sur*. Sur quels critères l’enseignant va-t-il baser sa rétroaction et que choisir pour ce type de situation ? *frapper* ou *jeter (sur)* ?

Pour répondre à ces questions, il faut disposer de données linguistiques suffisantes pour déterminer les contextes d’emploi de ces deux verbes de contact. Les corpus – en particulier expérimentaux — se révèlent très utiles pour vérifier en contexte les hypothèses sur les préférences lexicales de locuteurs L1 et d’apprenants. En s’appuyant sur des corpus en français L1 et L2, cette étude a pour objectif de fournir des éléments de réponse sur la distribution des verbes comme *jeter sur* et *frapper* en français L1, lors de la narration de scènes telles que celles présentées en Image 1 ; et de comparer cette distribution avec les productions d’apprenants arabophones syriens du français L2. Il s’agit à la fois d’identifier les contraintes d’usages régissant l’emploi de ces verbes en L1, et de vérifier si ces contraintes sont acquises par les apprenants de L2, afin de fournir quelques perspectives didactiques sur l’enseignement du vocabulaire. La section 1 présentera un bref état des lieux des travaux récents en didactique du lexique. En section 2, nous détaillerons les données de nos corpus et nos analyses. La dernière section discutera des résultats en lien avec les perspectives didactiques envisagées.

La recherche en didactique du vocabulaire

Les pratiques des enseignants et les recherches sur l’apprentissage du vocabulaire montrent que l’apprentissage du

1 Ce travail a bénéficié d’une aide de l’État gérée par l’Agence Nationale de la Recherche dans le cadre de l’appel à projets intitulé « Appel à projets générique 2020 pour le projet ANR JCJC CLASS » (Référence ANR-20-CE28-0019-01, PI Inès Saddour).

sens d'un mot ne garantit pas son réemploi approprié lors d'activités de production (cf. Grossman 2011 et Sardier & Roubaud 2020 pour des états des lieux sur la didactique et le réemploi du lexique). Ce constat peut s'observer autant en L1 qu'en L2, ce qui invite la recherche en didactique du lexique à s'appuyer autant sur les apports de travaux en français L1 (FL1), qu'en langue étrangère (FLE) (David et al. 2022). Il faut noter toutefois une différence entre l'acquisition L1 et L2 : en classe de FLE, se rajoute une hétérogénéité d'exposition à la langue cible (input qualitatif et quantitatif) ainsi que de langue d'origine, en plus de l'hétérogénéité habituellement trouvée en classe (origine sociale, dispositions cognitives, etc).

Par ailleurs, l'enseignement du lexique a pendant longtemps été mis de côté par rapport à d'autres domaines linguistiques, tant en L1 qu'en L2. Souvent pensé comme difficile à didactiser (Leeman 2000), le lexique et la façon de l'enseigner bénéficient d'un regain d'intérêt, grâce à l'essor de travaux récents dans différents domaines (didactique mais aussi acquisition, linguistique et psychologie) ou encore l'apport des ressources numériques (Sardier & Roubaud 2020). Cependant, la question de la ré-approriation des résultats des recherches par les enseignants se pose (Tremblay & Ronveaux 2018), et l'on peut souligner que l'utilisation de listes de mots reste encore une pratique courante. David et al. (2022) présentent des travaux récents qui mettent en lumière des pratiques pédagogiques intéressantes, insistant sur la nécessité de travailler les liens entre sémantique et syntaxe, notamment pour le domaine verbal. Au-delà d'un enseignement quantitatif, l'aspect qualitatif est mis en avant, en favorisant le développement de compétences métalinguistiques autant chez les enfants que les adultes. Enfin, David et al. (2022) estiment que la contextualisation du lexique est au cœur de l'apprentissage, d'autant plus avec les méthodes communicatives et actionnelles. Leur analyse de pratiques récentes montre également que les démarches inductives et explicites sont largement plébiscitées dans les manuels et pratiques des enseignants pour favoriser la mémorisation et le réemploi du vocabulaire.

Par ailleurs, les travaux de Garcia-Debanc & Aunargue (2020) en FL1 soulignent l'intérêt de s'appuyer sur les recherches menées en psycholinguistique pour y puiser autant des principes qui favorisent la mémorisation, que des modèles linguistiques nécessaires pour proposer un enseignement systématique du vocabulaire. Les auteurs proposent de se focaliser sur un domaine sémantique – en lien avec les besoins des apprenants – pour proposer des activités thématiquement cohérentes, et aider à l'élaboration des matériaux d'enseignement. Par la suite, le modèle peut également être employé pour analyser et évaluer les productions des élèves. Cependant, cette démarche nécessite l'existence ou l'élaboration d'un tel modèle pour un champ notionnel donné, reliant les unités de sens avec leur mode d'expression dans la langue.

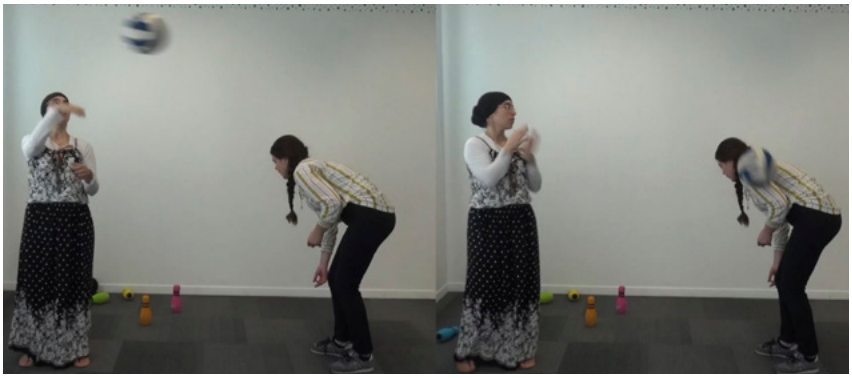


Image 1
Captures d'écran d'une des 16 vidéos (nº 9)

Ainsi, les activités didactiques de Garcia-Debanc & Aunargue (2020) s'organisent autour du champ notionnel du déplacement et s'appuient sur les travaux en sémantique cognitive de Talmy (2000). Ce modèle classe les verbes selon les composantes sémantiques exprimées : la direction (*monter, sortir*) ou encore la manière (*glisser, ramper*). Les auteurs proposent un déroulement didactique partant d'une liste de verbes de déplacement pour arriver à l'étude de ces mêmes verbes en contexte (textes littéraires, productions écrites), en passant notamment par une activité de classification selon leur sémantisme. Le modèle de Talmy (2000) est ainsi un point de référence pour l'enseignant pour guider et accompagner les apprenants dans leurs classements. La démarche est inductive, et l'apprenant est pleinement mobilisé dans l'accès au sens et la construction des savoirs.

Mireille Copin est doctorante à l'université Toulouse Jean Jaurès. Sa thèse porte sur la causalité en français L2 par des arabophones Syriens.

Inès Saddour est maîtresse de conférences à l'université Toulouse Jean Jaurès et étudie l'acquisition langagière et la socialisation des arabophones syriens en France.

« Les deux situations sont conceptualisées différemment et conduisent donc à l'emploi de verbes différents. »

A l'instar de ces recherches qui ont étudié le déplacement de manière approfondie, les recherches en psycholinguistique devraient se pencher sur d'autres domaines sémantiques et soutenir ainsi les enseignants via la création de matériel didactique basé sur les modèles linguistiques. Le présent article propose d'explorer un autre champ notionnel : la collision.

Analyse contrastive des verbes de collision employés en FRL1 et FRL2

Contiguïté spatio-temporelle

Nous souhaitons étudier les caractéristiques des scènes de collision qui peuvent influencer les choix verbaux. Ce que nous appelons scènes de collision sont, ici, des événements durant lesquels un personnage A touche un autre personnage B avec un objet O. Nous avons gardé constants les paramètres de la situation (*intentions, médiation par O, effet*, etc.) mais fait varier la contiguïté spatio-temporelle. Tout comme les études de Bellingham et al. (2020) nous nous intéressons particulièrement à l'effet du caractère direct de la situation sur la verbalisation. En effet, il est postulé qu'à travers les langues, une situation plus directe ne sera pas décrite par les mêmes expressions qu'une situation moins directe, selon le principe d'Iconicité (Haiman 1983). Une des mesures du caractère direct d'une situation peut être sa contiguïté spatio-temporelle (Bellingham et al. 2020). Ainsi nous postulons que la contiguïté entre l'action de A avec son objet et le moment de collision entre B et O aura une incidence sur la réalisation linguistique en français : les deux situations sont conceptualisées différemment et conduisent donc à l'emploi de verbes différents. Ce type de scène partage quelques caractéristiques avec les scènes de déplacement, puisque les locuteurs peuvent choisir d'exprimer le déplacement de l'objet, qu'il soit causé par A (*A jette O sur B*) ou présenté comme spontané, sans intervention de A (*O tombe sur B*). Mais il est également possible de se concentrer sur le contact, en employant des verbes exprimant l'action de A sur B (*frapper, heurter B*), avec ou sans précision de l'instrument (*avec O*). Les deux descriptions *frapper B* et *frapper B avec O* se distinguent en termes de *médiation par O* (±).

Données

Les corpus analysés ont été obtenus auprès de 29 locuteurs arabophones syriens L1 apprenants du français L2 (FRL2) en France, de niveau débutant à intermédiaire, et 22 locuteurs francophones L1 (FRL1), pour un total de 568 enregistrements[2]. Les participants ont raconté le contenu de 16 vidéos de situations de collision entre personnes avec divers objets. Dans la moitié des vidéos, une personne A envoie un objet O en l'air qui touche une autre personne B *(- contiguïté)*. Dans l'autre moitié, A touche directement B avec O *(+ contiguïté)*. La collision est toujours accidentelle et surprend les deux personnages *(- intention* pour toutes les vidéos). Le tableau 1 récapitule la liste des vidéos.

Les participants ont produit entre 1 et 3 verbes par verbalisation pour parler spécifiquement de la collision, l'emploi d'un seul verbe étant toutefois majoritaire (82 % en FRL1 et 91 % en FRL2).

Les verbes relevés dans les corpus FRL1 et FRL2 sont présentés dans le tableau 2.

Pour nos analyses, nous nous sommes inspirées de travaux en psychologie sur le classement de verbes comme *casser* et *couper* (Majid et al. 2008). A l'instar de ces travaux, nous avons relevé les verbes employés pour décrire chaque vidéo, et nous avons fait des analyses en clusters. En d'autres termes, nous avons utilisé un algorithme de classification, qui a constitué des groupes parmi les vidéos selon les verbes employés pour les décrire. Les groupes se basent sur une distance calculée à partir du décompte verbal. Ainsi, toutes les vidéos classées ensemble par l'algorithme sont globalement décrites par les mêmes verbes. Inversement, des verbes différents sont utilisés pour décrire les vidéos classées dans des groupes distincts.

2 Trois participants (FRL2) ont été exclus pour avoir parlé en anglais. 200 enregistrements sur 768 restant ont été écartés car ils ne décrivaient pas la collision.

+ Contiguïté		- contiguïté	
1	Une femme joue au cerf-volant et touche l'autre femme.	9	Une femme lance un ballon sur l'autre femme.
2	Une femme secoue une serviette et touche l'autre.	10	Une femme lance une chaussure sur l'autre.
3	Une femme passe le balai et touche l'autre.	11	Une femme lance une boule de bowling sur l'autre.
4	Une femme joue avec un élastique et touche l'autre.	12	Une femme lance un papier sur l'autre.
5	Une femme ouvre son parapluie et touche l'autre.	13	Une femme lance une bouteille d'eau sur l'autre.
6	Une femme enroule une carte et touche l'autre.	14	Une femme lance un gant sur l'autre.
7	Une femme remet son sac à dos et touche l'autre.	15	Une femme lance une bille sur l'autre.
8	Une femme met son écharpe et touche l'autre.	16	Une femme jongle et lance une pomme sur l'autre.

Tableau 1 Description et numérotation des items

Verbes présents dans les corpus	FRL1	FRL2
toucher	23 %	22 %
lancer	11%	9%
envoyer	11%	1%
jeter	7%	13%
heurter	7%	<1%
cogner	5%	
atterrir	4%	
taper	3%	14%
percuter	3%	
frapper	2%	17%
surprendre, déranger	2%	1%
échapper, lâcher, s'apercevoir	2%	
balancer	1%	2%
tomber	1%	3%
mettre, ouvrir sur, rater	1%	1%
arriver, balayer sur, donner un coup, effleurer, perdre, réagir, recevoir, secouer sur, se prendre	1%	
passer	< 1%	1%
se cogner	< 1%	< 1%
atteindre, bousculer, frôler, glisser, partir, perturber, pousser, retomber, se cogner, s'effleurer, s'énerver, s'ouvrir, tirer, viser	< 1%	
faire mal		4%
voler		2%
venir, bouger		2%
relancer		1%
(accident), bouger, choquer, énerver, envoler, être, faire un lancer, ramasser, s'embêter, s'enlever, trier sur		< 1%

Tableau 2 Fréquence des différents verbes employés en FRL1 et FRL2

Conceptualisation de la contiguïté et apprentissage des verbes de collision : quelques résultats

Les premières analyses à partir du relevé de tous les verbes employés montrent la diversité verbale en FRL1, par comparaison aux productions en FRL2. On compte 43 verbes différents en FRL1, contre 32 en FRL2, mais seulement 16 verbes sont communs aux deux corpus. Ce résultat n'a rien de surprenant, et ce d'autant plus entre des locuteurs L1 et L2, et chez des apprenants de niveaux différents (David et al. 2022). Il existe tout de même des similarités entre les deux groupes. Le verbe *toucher* est ainsi le plus fréquent dans les deux corpus, représentant presque un quart des verbes utilisés, respectivement 23 % des occurrences en FRL1 et 22 % en FRL2. Si le verbe *lancer* est présent dans des proportions similaires en FRL1 (11 %) et FRL2 (9 %), ce n'est pas le cas du verbe *jeter (+ préposition)*, presque moitié moins présent en FRL1 (7 %) qu'en FRL2 (13 %). A l'inverse, le verbe *envoyer (+ préposition)* est relativement fréquent en FRL1 (11 %), mais peu employé en FRL2 (1 % soit 3 occurrences). Inversement, les verbes *taper* et *frapper* sont aussi utilisés par des locuteurs des deux groupes mais plus fréquemment en FRL2 (14 % et 17 % respectivement) qu'en FRL1 où ils ne représentent que 3 % et 2 % des verbes. A la place, de nombreux synonymes comme *heurter, cogner* ou *percuter* – presque absents en FRL2 – sont utilisés. Les autres verbes les plus fréquents sont *atterrir* (4 % en FRL1, 0 % en FRL2) et *faire mal* (0 % en FRL1, 4 % en FRL2), tous les autres étant utilisés moins de 10 fois dans chaque groupe.

> « On compte 43 verbes différents en FRL1, contre 32 en FRL2, mais seulement 16 verbes sont communs aux deux corpus. »

La figure 1 illustre les classements des vidéos obtenus à l'issue de l'analyse en cluster, mis en évidence par les embranchements et les couleurs. Chaque branche représente une vidéo (voir la numérotation des vidéos dans le tableau 1). La longueur des branches renseigne sur la similarité entre les verbes employés : de courtes branches indiquent que les vidéos sont décrites par un nombre limité de mêmes verbes.

En FRL1, l'algorithme a classé les vidéos en deux groupes : les vidéos 1 à 8 (*+contiguïté*) ont été classées dans le même groupe en raison de l'emploi fréquent de verbes comme *toucher/taper*, qui expriment le contact, tandis que les vidéos 9 à 16 (*-contiguïté*) sont classées ensemble en raison de l'emploi de verbes comme *jeter/envoyer,* qui expriment le mouvement. On peut donc considérer que la contiguïté affecte les décisions lexicales des locuteurs en FRL1, qui vont alors choisir de mettre en avant à travers le verbe des aspects différents de l'événement : le contact ou le mouvement. Certains verbes spécifiques, comme *balayer sur* (vidéo 3), conduisent à l'apparition de sous-groupes. Les branches relativement longues des arbres de classification indiquent tout de même une grande diversité dans les verbes employés dans chaque groupe. Il apparaît néanmoins clairement que les situations où les objets sont envoyés dans les airs ne sont pas décrites de la même manière que les autres. Pour les locuteurs FRL1, *jeter sur* est employé quand il n'y a pas de contiguïté. Au contraire, les verbes comme *toucher/frapper* sont quasiment réservés au cas où le personnage tient toujours l'objet à la main. L'emploi des verbes est donc distribué selon un critère de contiguïté.

En FRL2, les analyses révèlent l'absence de distinction entre les items à partir des verbes employés, car les premiers groupements classent la vidéo 14 à part, tandis que le reste des vidéos appartient

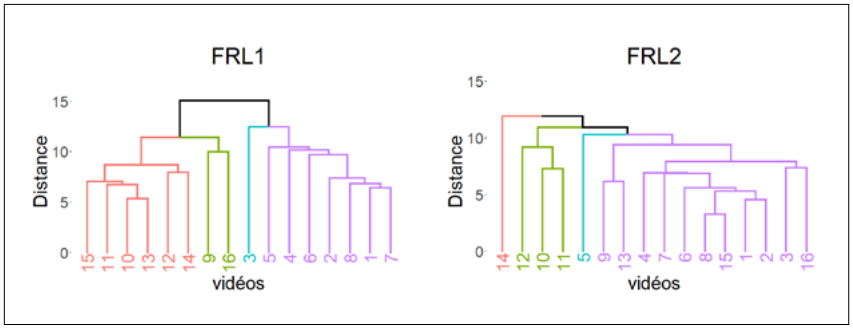


Figure 1
Arbres de classification des vidéos selon les verbes employés (FRL1 et FRL2)

au même groupe. Certains sous-groupes émergent, mais les différences entre chaque sous-groupe sont marginales. L'impossibilité de faire des classements selon la situation ne vient pas seulement de l'utilisation indifférenciée de *toucher* ou *jeter* pour les deux situations, mais surtout du fait que beaucoup de verbes n'apparaissant qu'une ou deux fois, ils ne peuvent pas servir à distinguer les vidéos.

Dans un premier temps, le niveau de maîtrise de l'apprenant et ses connaissances lexicales peuvent jouer un rôle dans l'absence de groupements comparables aux données en FRL1. Cependant, ces différences dans les classements peuvent également venir d'une conception différente de l'événement à décrire, selon leur L1.

Ainsi la différence de contiguïté ne semble pas ou peu impacter les apprenants dans leur choix de verbe en FRL2. En effet, ils emploient les verbes *toucher* et *frapper* quelle que soit la situation, même quand l'objet est envoyé en l'air, et même s'ils connaissent *jeter* et *lancer* (cf. 2.2). Rappelons le premier exemple donné :

> 1) « y a une [dame] qui est frappé l'autre par sa balle aussi » (FRL2, CLS14, n°9).

L'énoncé en (1) décrit une situation où le premier personnage ne tient plus l'objet à la main quand celui-ci touche l'autre personnage. Cependant en FRL1, comme nous l'avons mentionné, les verbes *toucher* ou *heurter* sont préférés pour une situation où le personnage tient toujours l'objet à la main, comme en (2) et (3) :

> 2) une avec un cerf volant. euh qui heurte l'autre avec son cerf volant (FRL1, CLF02, n°1)

> 3) celle qui ouvre le parapluie. touche celle. qui met son écharpe (FRL1, CLF08, n°5)

Néanmoins, il y a 14 exemples en FRL1 où ces verbes apparaissent dans la condition où l'objet est envoyé en l'air, comme dans les exemples (4) et (5). Mais ce qui est intéressant, ce sont les compléments admis par le verbe dans ce contexte précis. En effet, le verbe est utilisé en FRL1 avec un seul complément (*A heurte B)* sans jamais employer l'ajout « *avec O* ». La manière dont A est entrée en contact avec B n'est donc pas mentionnée, même si elle pourrait être inférée via l'évocation de l'objet en périphérie.

> 4) euh une femme heurte une autre femme. en enlevant sa chaussure (FRL1, CLF14, n°10)

> 5) euh une femme qui jouait avec une bouteille d'eau. heurte euh une femme devant elle (FRL1, CLF14, n°13)

Le fait de préciser l'action avec une indication de l'instrument (*frapper B avec O*) n'est jamais attesté en FRL1 pour les items 9 à 16 (- *contiguïté*). En FRL1, la précision de l'instrument dans la structure argumentale des verbes exprimant le contact comme *frapper* est spécifique aux items 1 à 8 *(+ contiguïté*), ce qui semble indiquer que cet ajout est fortement associé au fait de toujours tenir l'objet à la main. Le sémantisme verbal est donc lié à la structure argumentale du verbe. Ainsi dans notre corpus FRL1, les verbes utilisés avec deux compléments sont les verbes de déplacement provoqué (*A jette O sur B*) ou les verbes de contact (*A heurte B avec O*), mais uniquement lorsque le personnage A tient toujours l'objet à la main pour ces derniers (items 1 à 8, *+ contiguïté)*. A l'inverse, les apprenants FRL2 emploient non seulement plus de verbes comme *frapper* pour les situations des items 9 à 16 (- *contiguïté*), mais en outre, ils les utilisent souvent en précisant l'objet dans un complément prépositionnel, comme en (1).

Réemploi des stimuli vidéos en classe de langue

Sans données empiriques sur la description d'un type d'événement en L1, cela nécessite un travail de la part des enseignants pour déterminer l'emploi de certains verbes. Or, c'est un prérequis pour concevoir des activités didactiques susceptibles de favoriser l'appropriation de ces distinctions sémantiques subtiles. La psycholinguistique permet de mettre en évidence des distinctions sémantiques à partir d'analyses de l'usage. L'utilisation de méthodes expérimentales pour constituer un corpus est pertinente, car elle permet de révéler la diversité et la stabilité des descriptions pour une même situation, en demandant de produire plusieurs énoncés pour des situations relativement similaires. Cela permet notamment de s'assurer que c'est bien le

« [S]i les différences entre les productions en L1 et en L2 viennent aussi d'une conceptualisation différente de l'événement, une activité mobilisant les compétences méta-linguistiques pourrait permettre de faire prendre conscience de cette différence de perspective, en attirant explicitement l'attention sur les caractéristiques de la situation. »

critère de contiguïté, et non pas la caractéristique propre à une seule des vidéos présentées, qui conduit à l'utilisation de verbes différents (*frapper* ou *jeter sur*).

Ensuite, l'intérêt de constituer des corpus en L1 et en L2 à partir des mêmes vidéos est double : cela permet d'observer les usages en L1, qui peuvent servir d'exemples et de matériel pour l'apprentissage, tout en évaluant les productions des apprenants, afin de savoir si ces usages sont connus. La méthodologie expérimentale permet d'étudier les critères précis responsables de la variation intra et inter groupe. Nous avons vu qu'en FRL1, le choix du verbe est influencé par une certaine conceptualisation de l'événement, qui varie ici selon la contiguïté. En FRL2, où les réalisations verbales sont largement contraintes par les compétences lexicales des apprenants (niveau en FRL2, types d'inputs reçus, etc.), on peut remarquer la présence de nombreux verbes qui, à première vue, ne donnent pas d'indications sur une conceptualisation partagée des événements par tous les apprenants FRL2. L'analyse des données en arabe syrien (la L1 des apprenants) est en cours et renseignera sur les catégorisations des événements éventuellement mises en avant dans les réalisations linguistiques. Les analyses en clusters permettent de situer les potentielles difficultés des apprenants, et mettent en évidence le critère de distinction qu'il convient d'enseigner entre *jeter sur* et *frapper* en français pour une scène de collision.

Nos analyses quantitatives illustrent des classements inconscients entre les différents verbes qui existent chez les locuteurs du FRL1. Cela corrobore les propositions de Garcia-Debanc & Aurnague (2020) concernant l'utilisation de tâches de classification pour favoriser l'apprentissage du vocabulaire. Ces propositions sont d'autant plus pertinentes que si les différences entre les productions en L1 et en L2 viennent aussi d'une conceptualisation différente de l'événement, une activité mobilisant les compétences méta-linguistiques pourrait permettre de faire prendre conscience de cette différence de perspective, en attirant explicitement l'attention sur les caractéristiques de la situation. Par ailleurs, les analyses des énoncés confirment la nécessité de prendre en compte les liens entre le sens des verbes et les compléments qui les suivent (objet direct, groupe prépositionnel).

Enfin, les vidéos sont à la fois des prompts pour la génération des corpus[3] et partie intégrante de ceux-ci en tant qu'illustration des situations de collision et de toutes les caractéristiques qui peuvent avoir un impact sur les choix sémantiques de la description. De la même manière, il serait intéressant de s'appuyer sur les vidéos illustrant les emplois en FRL1 lors d'un travail d'enseignement du lexique à partir d'un corpus., Si nos vidéos n'ont pas été créées avec une visée didactique, elles peuvent néanmoins servir de base pour la création de matériel. David et al. (2022) soulignent que les images sont un support particulièrement intéressant en classe de FLE, car elles contextualisent le sens en permettant de dépasser l'aspect opaque des mots et des expressions.

Ainsi il est possible de mettre en place une démarche inductive, à la manière de Garcia-Debanc & Aurnague (2020) durant laquelle les apprenants s'appuient sur un corpus en FRL1 et les vidéos qui ont permis leur production pour accéder au(x) sens des verbes employés et leurs règles d'utilisation. Une telle pédagogie correspond à la tendance actuelle dans l'enseignement du vocabulaire, qui vise à mettre l'apprenant au cœur des apprentissages, en encourageant l'observation et la formulation d'hypothèses.

3 Ces données font partie d'un sous-corpus du projet CLASS.

Références

Bellingham, E., Evers, S., Kawachi, K., Mitchell, A., Park, S., Stepanova, A., & Bohnemeyer, . (2020). Exploring the Representation of Causality Across Languages : Integrating Production, Comprehension and Conceptualization Perspectives. In Bar-Asher Siegal, E., Boneh, N. (eds), *Perspectives on Causation: Selected Papers from the Jerusalem 2017 Workshop* (pp. 75-119). Springer International Publishing.

David, C., Gala, N., Leconte, A., & Roubaud, M.-N. (2022). Contextualiser pour faciliter l'accès au sens : Focus sur l'enseignement du lexique en FLM, FLS et FLE. *TIPA. Travaux interdisciplinaires sur la parole et le langage*, *38*, Article 38.

Garcia-Debanc, C., & Aurnague, M. (2020). Quelle programmation des activités d'étude de la langue sur le lexique en fin d'école primaire pour susciter le réemploi en production écrite? *Repères. Recherches en didactique du français langue maternelle*, *61*, 17-33.

Grossmann, F. (2011). Didactique du lexique : État des lieux et nouvelles orientations. *Pratiques. Linguistique, littérature, didactique*, *149-150*, Article 149-150. https://doi.org/10.4000/pratiques.1732

Haiman, J. (1983). Iconic and economic motivation. *Language*, 59(4), 781–819.

Leeman, D. (2000) Le vertige de l'infini ou la difficulté de didactiser le lexique, Le français aujourd'hui, 131, p. 42-52.

Majid, A., Boster, J. S., & Bowerman, M. (2008). The cross-linguistic categorization of everyday events : A study of cutting and breaking. *Cognition*, *109*(2), 235-250.

Sardier, A., & Roubaud, M.-N. (2020). Construire la compétence lexicale : Quelles avancées vers le réemploi aujourd'hui ? *Repères. Recherches en didactique du français langue maternelle*, *61*. https://doi.org/10.4000/reperes.2537

Talmy, L. (2000). *Toward a cognitive semantics* (Vol. 2). MIT press.

Tremblay, O., & Ronveaux, C. (2018). Aimer les mots, discipliner le lexique. *La Lettre de l'AIRDF*, *64*(1), 15-19. https://doi.org/10.3406/airdf.2018.2246

Tema

UN CORPUS DE PRODUCTIONS ÉCRITES EN FRANÇAIS LANGUE ÉTRANGÈRE – MATÉRIAU POUR MIEUX COMPRENDRE LES CHOIX LEXICAUX D'APPRENANTS MULTILINGUES

The aim of this paper is to present a corpus containing 105 Swedish learners' written production in French and to discuss some possible didactic applications of such a corpus. Learners from four different grades (age 11-15) were invited to retell a short picture story in writing. One of the main characteristics of the corpus is that it is multilingual in nature, with many instances of influences from the learners' previously acquired languages. It is argued that this type of corpus can be useful in class, allowing for potential learning situations, linguistic development and metalinguistic reflection.

Christina Lindqvist
| Université de Göteborg et Østfold university college

Christina Lindqvist est professeure de français à l'Université de Göteborg, Suède, et Østfold university college, Norvège. Ses recherches se situent dans les domaines de l'acquisition d'une troisième langue et de l'acquisition du vocabulaire chez les apprenants suédois du français. Elle enseigne la didactique des langues étrangères, la linguistique, la grammaire et le vocabulaire. Elle dirige des thèses de doctorat en didactique/apprentissage des langues étrangères.

Introduction

L'objectif de cet article est de présenter des recherches récentes menées sur un corpus de productions écrites d'apprenants suédois du français, en vue de mettre en évidence quelques applications didactiques possibles à partir de ce genre de corpus. Le corpus, qui a été constitué à des fins de recherche en premier lieu, et qui sera décrit plus amplement dans la section suivante, contient des narrations écrites à partir d'une série d'images racontant une petite histoire de deux enfants qui partent en pique-nique (voir l'annexe). Les apprenants sont des élèves du secondaire I (11 à 15 ans) en Suède. Nous allons montrer que ce type de corpus permet, entre autres, d'analyser les compétences linguistiques des apprenants à différents niveaux d'apprentissage et de mieux comprendre le processus d'apprentissage en se focalisant en premier lieu sur les aspects lexicaux et le multilinguisme des apprenants. Les résultats des recherches précédentes sur l'impact des langues déjà acquises sur le français écrit nous permettront notamment de mettre en avant l'utilité d'un corpus d'apprenants multilingues dans le cadre de l'enseignement des langues.

Présentation du corpus

Le corpus consiste en 105 productions écrites d'élèves suédois, âgés de 11/12 à 14/15 ans, qui ont été récoltées dans quatre classes, de la 6e à la 9e année du système scolaire suédois. Il convient de noter qu'il s'agit de quatre classes différentes, ce qui veut dire que le corpus est de nature transversale et non longitudinale. Les élèves ont été invités à écrire un texte à partir d'une série d'images intitulée « The Dog Story », qui est l'histoire de deux enfants qui partent en pique-nique, mais lorsqu'ils arrivent dans la forêt, il s'avère que leur chien s'était caché dans leur panier et qu'il a mangé la nourriture qui s'y trouvait. Les élèves disposaient de 20 minutes pour cette tâche. Ils ne disposaient que d'un stylo et d'une feuille de papier, et n'étaient pas

Classe	Nombre d'élèves	Âge	Semestres d'études de français	Mots produits au total	Nombre min et max de mots	Moyenne de mots par texte	Langue maternelle	Langues étudiées ou parlées à la maison
6	17	11/12	1	503	7–74	26	Suédois	Anglais (17) Finnois (1) Norvégien (1) Polonais (1)
7	26	12/13	3	1835	18–164	66	Suédois	Anglais (26) Danois (1)
8	35	13/14	5	3516	34–197	100	Suédois	Anglais (35) Chinois (2) Espagnol (16) Allemand (1)
9	27	14/15	7	3066	62–263	114	Suédois	Anglais (27) Allemand (2) Espagnol (12) Chinois (4)

Tableau 1
Le corpus

autorisés à utiliser des dictionnaires, ni à collaborer. Conformément à la méthode pédagogique utilisée à l'école, ils devaient essayer de se concentrer sur la transmission du message et de communiquer de manière aussi détaillée que possible ce qu'ils voyaient sur les images. L'exemple 1 présente un texte écrit par un élève en 9e. On voit que le texte est rédigé en français, mais contient des éléments d'autres langues comme « mountain » ou « años ».

Ex. 1. Texte d'un élève en 9e

C'est un garçon et une fill et un chien et une mére dans la picture . Les enfants faisons sandwiches. Le chien manges les sandwiches. Puis les enfants allons en un picnic avec zero sandwiches. Les enfants allon en un en un mountain avec deux animeaux grande. Apres les enfants mangais les sandwiches il est zero sandwiches et un chien le chien mangais tout les sandwiches et les enfants mangais zero sandwiches.

PS. En le mountain il est trois trés grand trees et le chien etais blanc et marron. Et les enfants etais sanq y sept años.

Le tableau 1 rend compte des caractéristiques du corpus. En ce qui concerne les connaissances d'autres langues des apprenants, ils avaient tous étudié l'anglais depuis l'âge de 7, 8 ou 9 ans, soit entre trois et huit ans lors de la collecte des données. Un trait caractéristique important des apprenants de ce corpus est donc que l'anglais représente leur première et principale langue étrangère. De plus, d'autres langues étrangères sont aussi présentes, car en 8e, il est possible d'opter pour une langue étrangère supplémentaire comme l'allemand ou l'espagnol. Certains élèves parlent aussi d'autres langues à la maison (p.ex. le norvégien ou le danois). Voilà pourquoi le nombre de langues supplémentaires augmente à partir de la 8e année, contribuant au caractère multilingue du corpus.

Comme il ressort du tableau 1, il s'agit d'un corpus relativement restreint en termes de nombre total de mots et de participants. Néanmoins, il sert d'exemple d'un type de corpus d'apprenants qui pourrait facilement être conçu par le chercheur/enseignant à des fins différentes. En effet, la collecte des données écrites présente plusieurs avantages, dont le temps nécessaire à la collecte et au traitement des données, moindre que pour des données orales (cf. Tracy-Ventura & Paquot, 2020).

Résultats de recherche

Le corpus que nous venons de présenter permet d'analyser de nombreux aspects différents de l'interlangue des apprenants. Comme il ressort du tableau 1, une des caractéristiques des élèves de ce corpus est leur multilinguisme. En effet, alors qu'ils ont le suédois comme langue maternelle, et qu'ils ont tous commencé l'apprentissage de l'anglais bien avant celui du français, ils ont en outre des connaissances d'autres langues, telles

Ce type de corpus permet, entre autres, d'analyser les compétences linguistiques des apprenants à différents niveaux d'apprentissage et de mieux comprendre le processus d'apprentissage en focalisant en premier lieu sur l'aspect lexical et le multilinguisme des apprenants

que l'allemand, l'espagnol ou le chinois Ainsi, les études menées sur ce corpus se sont surtout concentrées sur la manière dont les langues déjà acquises influencent l'écriture et ceci en examinant les choix lexicaux[1]. L'analyse des choix lexicaux est efficace, car elle permet d'examiner plusieurs aspects de la compétence lexicale des apprenants en même temps (Jarvis & Pavlenko, 2008). Effectivement, le type de tâche exige que différents mots soient utilisés pour que le message soit transmis, ce qui implique qu'il est nécessaire d'avoir un vocabulaire assez étendu. Cependant, certains mots centraux du récit vont forcément manquer aux apprenants. Alors, que faire devant des problèmes de lexique ? Les résultats des études antérieures montrent que plusieurs langues sont utilisées lors du processus d'écriture pour combler les lacunes lexicales. Ainsi, dans Lindqvist (2015), qui a uniquement inclus les apprenants ayant des connaissances en anglais à part le suédois, il s'est avéré que l'anglais était la source dominante du *transfert lexical*, c'est-à-dire les emplois lexicaux manifestant des influences d'autres langues que le français, et ceci dans toutes les classes. En effet, 70 % du total des occurrences de transfert provenaient de l'anglais. De plus, une différence qualitative a été notée dans le sens où les cas transférés du suédois étaient surtout des changements de langue non adaptés au français, p.ex. *berg* (fr. *montagne*, ang. *mountain*), *korg* (fr. *panier*, ang. *basket*), ou *borta* (fr. *disparu*, ang. *gone*). En revanche, ceux transférés de l'anglais étaient souvent des adaptations comme *basquette* (crée à partir du mot anglais *basket*, mot-cible français : *panier*) ou *travaille* (ang. *travel*, mot-cible français *vont*). Dans Lindqvist (2019) également, incluant cette fois tous les apprenants du corpus, les transferts lexicaux provenaient des langues étrangères dans une plus large mesure que de la langue maternelle. Ainsi, alors que l'anglais dominait en tant que source d'influence, l'espagnol était aussi utilisé par un certain nombre d'apprenants.

Dans une étude visant à examiner d'autres aspects lexicaux à partir du même corpus, Lindqvist (2021) s'est focalisée sur certains items spécifiques dans la série d'images et sur la manière dont ils ont été produits par les apprenants dans leurs narrations écrites. Cette approche méthodologique a permis une analyse de la *profondeur* du vocabulaire, c'est-à-dire de l'aspect qualitatif des connaissances lexicales (à la différence de l'*étendue* du vocabulaire, qui concerne l'aspect quantitatif). Ainsi, l'analyse a porté sur l'orthographe, la relation forme-sens et les parties des mots, trois des aspects dans le modèle bien connu de Nation (« How is the word written and spelled ? », « What word form can be used to express this meaning ? », « What word parts are needed to express the meaning ? », Nation, 2020, p. 16). Les items inclus dans l'analyse étaient *mère, fille* et *garçon* (pour examiner l'orthographe), *chien* et *panier* (forme-sens), et *manger* et *dire* (parties des mots, soit morphèmes grammaticaux). Pour l'orthographe, les résultats ont mis en évidence un pattern assez variable en termes de difficulté et de degré d'exactitude. On note, par exemple, en principe les mêmes variantes d'orthographe de *mère* (*méré*, *mere*, *mére*, *maire*, *merè, mèrè*) dans toutes les classes, mais le degré d'exactitude est de plus en plus élevé à partir de la 7e. Pour *fille*, en revanche, ce mot est correctement écrit sans exception en 6e, alors que deux autres variantes apparaissent dans les autres classes (*fill*, *filles*). Concernant *garçon*, cette variante est accompagnée de *garcon* dans toutes les classes, tandis qu'il y a deux variantes supplémentaires en 6e : *garquon*, *garzon*. Puis, pour l'analyse de la relation forme-sens, il est clair que le mot *panier* pose problèmes aux apprenants, résultant en l'emploi du suédois et de l'anglais (cf. plus haut), surtout différentes variantes du mot anglais *basket* : *basquette*, *basquet*, *bascet*, *baskuette*, *basquete*, *baskét*, *baskèt*. Pour *manger*, le suédois (*äta*) et l'espagnol (*comer*) sont utilisés en 6e et en 7e, alors que les autres apprenants utilisent le mot cible sans exception. Enfin, pour les morphèmes grammaticaux, il s'est avéré que *dire* est très peu utilisé, mais

1 Les aspects grammaticaux ont également été analysés dans certaines publications, cf. Lindqvist (2015 ; 2019).

que les formes sont toujours correctes. *Manger*, de l'autre côté, est employé dans toutes les classes avec une augmentation de flexions différentes et une maitrise croissante au fur et à mesure des années.

Pour finir ce tour d'horizon des recherches menées sur le corpus en question, mentionnons aussi l'étude de Falk & Lindqvist (2022). Dans cette étude, les chercheures examinent les attitudes d'enseignants vis-à-vis du multilinguisme en salle de classe des langues étrangères en Suède. Des extraits du corpus sont utilisés dans le but d'inviter les enseignants à réfléchir et à exprimer leurs attitudes sur la présence des langues déjà acquises dans les textes. Il s'est avéré que les attitudes divergent en fonction des langues : l'usage de l'anglais est généralement vu comme une stratégie efficace, contribuant à la compréhension du lecteur potentiel, tandis que l'emploi du suédois est considéré comme moins réussi.

Applications didactiques

Jusqu'ici, le corpus a surtout été exploré dans une perspective lexicale et translinguistique. En général, les résultats indiquent clairement que les langues déjà acquises sont utilisées dans une large mesure pour combler les lacunes lexicales. Un des résultats les plus intéressants d'un point de vue didactique est probablement que les langues étrangères sont souvent activées et employées de différentes manières lors de la rédaction d'un texte. Ainsi, à côté du suédois – la langue maternelle – les apprenants recourent à l'anglais ainsi qu'à l'espagnol à un haut degré. Notons que ces deux langues sont relativement proches du français, ce qui pourrait contribuer au fait qu'elles sont utilisées. En fait, les résultats d'une enquête utilisée dans l'étude de Lindqvist (2015) font ressortir que les apprenants perçoivent plus de similarités entre l'anglais et le français qu'entre le suédois et le français. Nous avons aussi pu constater que les enseignants suédois considèrent que l'emploi de l'anglais constitue une stratégie efficace de production écrite, alors qu'ils ne sont pas aussi positifs vis-à-vis du suédois. Ces résultats rappellent ceux obtenus dans le domaine des recherches sur l'acquisition d'une troisième langue, où maintes études ont montré que toutes les langues acquises peuvent impacter l'apprentissage d'une langue étrangère supplémentaire. De surcroit, les chercheurs dans le domaine insistent souvent sur l'importance de reconnaitre ce phénomène en salle de classe. De manière générale, une application didactique serait alors d'encourager l'emploi des langues déjà acquises. Les possibilités sont multiples, mais nous allons limiter notre discussion à l'utilité d'un corpus d'apprenants écrit en classe multilingue. Prenons à cette fin un exemple d'un texte écrit par un élève en 8e.

Ex. 2. Texte d'un élève en 8e

Une fille et un garçon adore sa mère. Ils sont manger avec sa mère. La filles c'est le soer de le garçon. Ils sont un chien qui s'apelle Pedde. Pedde aime manger beaucop avec le garçon e la fille. Le garçon s'apelle Markus et la fille s'appelle Sofie. Sofie et Markus sur le **park** . Le **gras** c'est très verte et la souleille c'est très jaune. Ils sont une picnic . Dans le **basket** avec la **sallade** le chien Pedde être car Pede adore manger. Pede a mangé c'est tout dans le **basket**. Mais son bouteille au lau c'est n'ai pas manger. Sofie et Markus ne manger pas. Mais Sofie et markus sont son chien Pede. Sofie a cheveux blond ave une **dresse** blanche. Markus a cheveux noir ave une pulle blanche. Ils sont cinq ans et ils sont très gentille. Ils adore son chien et sa mère. Ils aimes le jour car c'est le très chaud et le soileille brille.

Un des résultats les plus intéressants d'un point de vue didactique est probablement que les langues étrangères sont souvent activées et employées de différentes manières lors de la rédaction du texte.

Dans l'exemple 2, nous avons mis en caractères gras les mots manifestant des influences d'autres langues : *park* (ang./sué. *park*), *gras* (ang. *grass*, sué. *gräs*), *basket* (ang. *basket*), *sallade* (sué. *sallad*) et *dresse* (ang. *dress*). Ces exemples pourraient constituer la base d'une discussion en classe autour des différences et similarités entre différentes langues,

Ces exemples pourraient constituer la base d'une discussion en classe autour des différences et similarités entre différentes langues, par exemple au niveau de l'orthographe (parc/park, sallad/salade) et de la relation forme-sens (dress/robe, basket/panier).

par exemple au niveau de l'orthographe (*parc/park, sallad/salade*) et de la relation forme-sens (*dress/robe, basket/panier*). Ce genre de discussion pourrait contribuer à une situation d'apprentissage, non seulement pour ce qui est du développement linguistique mais aussi de la compétence métalinguistique et multilingue des apprenants. Ce corpus d'apprenants pourrait aussi être utilisé dans le cadre de l'écriture. En classe, l'enseignant pourrait choisir d'éviter de parler des influences d'autres langues comme une erreur et les soulever comme des exemples d'un trait caractéristique de la compétence linguistique de l'apprenant multilingue. En effet, le texte pourrait être traité en classe comme une première ébauche où l'apprenant a fait des efforts pour communiquer l'histoire en puisant dans son répertoire langagier contenant plusieurs langues, plutôt que d'abandonner devant des lacunes lexicales. Les élèves pourraient discuter des instances actuelles dans le texte et proposer des révisions, que ce soit au niveau du vocabulaire ou d'autres aspects de la langue. En résumé, adopter une approche multilingue en salle de classe implique plusieurs avantages et l'emploi d'un corpus écrit d'apprenants multilingues peut s'avérer une ressource utile à cette fin.

Conclusion

Le but de cet article était de présenter un corpus d'apprenants contenant des productions écrites et les applications didactiques possibles de ce type de corpus. A partir des recherches précédentes sur le corpus, nous avons mis en exergue le caractère multilingue du corpus, permettant diverses applications didactiques en salle de classe. Nous avons notamment présenté les opportunités d'apprentissage et de réflexion métalinguistique à partir de textes contenant des influences de langues déjà acquises, trait inhérent au développement linguistique d'apprenants multilingues.

Bibliographie

Falk, Y. & Lindqvist, C. (2022). Teachers' attitudes towards multilingualism in the foreign language classroom: The case of French and German in the Swedish context. In: A. Krulatz, G. Neokleous & A. Dahl (Dirs), *Theoretical and applied perspectives on teaching foreign languages in multilingual settings: pedagogical implications*. Bristol: Multilingual Matters, pp. 154-169.

Jarvis, S. & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. New York: Routledge.

Lindqvist, C. (2015). Do learners transfer from the language they perceive as most closely related to the L3? The role of psychotypology for lexical and grammatical cross-linguistic influence in French L3. In: G. De Angelis, U. Jessner & M. Kresic (Dirs), *Multilingualism: Crosslinguistic influence in language learning*. London: Bloomsbury Publishing, pp. 231-251.

Lindqvist, C. (2019). Didactic challenges in the multilingual classroom. The case of French as a foreign language. In: M. J. Gutierrez-Mangado, M. Martínez Adrián, F. Gallardo Del Puerto (Dirs), *Cross-Linguistic Influence: From Empirical Evidence to Classroom Practice*. Springer International Publishing Switzerland, pp. 87-99.

Lindqvist, C. (2021). Vocabulary knowledge in L3 French: A study of Swedish learners' vocabulary depth. *Languages* 6, no. 1: 26. https://doi.org/10.3390/languages6010026

Nation, P. (2020). The different aspects of vocabulary knowledge. In: S. Webb (Dir), *The Routledge handbook of vocabulary studies*. New York: Routledge, pp. 15-29.

Tracy-Ventura, N. & Paquot, M. (2020). Second language acquisition and corpora: An overview. In: N. Tracy-Ventura & M. Paquot (Dirs), *The Routledge handbook of second language acquisition and corpora*. New York: Routledge, pp. 1-8.

Annexe

The Dog Story

Tema

TEACHING AND LEARNING FOREIGN LANGUAGES: INSIGHTS FROM CLASSROOM CORPUS RESEARCH

Un corpus de classe L2 est un ensemble d'enregistrements audio ou vidéo de leçons authentiques, qui ont été transcrits et préparés pour l'analyse. Un tel corpus donne un aperçu direct de la manière dont le programme L2 est vécu en classe en temps réel.

Dans cette contribution, nous présentons d'abord quelques exemples de recherches antérieures utilisant des corpus de classe, en soulignant la variété des perspectives qui peuvent les motiver. Nous décrivons ensuite le corpus «Apprendre le français» de leçons enseignées à de jeunes débutants anglophones, et les aperçus sur l'input, l'engagement et l'apprentissage de la L2 qui découlent de différentes analyses de ce corpus. En conclusion, nous discutons des contributions potentielles de la recherche sur les corpus de classe à notre compréhension de la pédagogie et à l'amélioration de la pratique.

Rosamond Mitchell
| University of Southampton
Florence Myles
| University of Essex

Introduction

Learner corpora are the main focus of this special issue. However in this contribution we turn our attention to a rather different though related type of corpus: the L2 classroom corpus.

A classroom corpus is a set of audio- or videorecordings of authentic L2 lessons, which have been transcribed and prepared for analysis. Such a corpus provides direct insight into how the L2 curriculum is delivered and experienced in the classroom in real time.

In this contribution we first present some examples of past research using classroom corpora, emphasising the variety of perspectives adopted. We then describe the 'Learning French' corpus of lessons taught to young Anglophone beginners, and the insights concerning input, engagement and L2 learning deriving from different analyses of this corpus. In conclusion we discuss the potential contributions of classroom corpus research to our understanding of pedagogy and the improvement of practice.

Approaches to classroom corpus research

An important strand of classroom corpus research uses conversation analysis (CA) to interpret interactions among classroom participants in a bottom-up fashion (e.g. Kunitz et al, 2021). CA research on classroom discourse has, for example, examined the L2 socialisation of young children (Cekaite, 2022), and the developing interactional competence of older classroom learners (Pekarek Doehler & Fasel Lauzon, 2015). A significant corpus in this tradition is the Multimedia Adult English Learner Corpus (MAELC), a collection of adult English as a Second Language (ESL) lessons videorecorded in Portland, USA (Reder et al., 2003). Researchers using this corpus have examined the development of students' linguistic and interactional competence. Through longitudinal case studies of

individual students, they have shown how this development is embodied in the routines of classroom life. For example, Hellermann (2008) described how students learn to initiate, manage and conclude interactions with peers during classroom small group activities. Eskildsen (2012) traced the development of L2 English negation by two students. He argued that the individuals' rather different learning pathways are shaped by their interactional experiences, which at times may consolidate non-standard forms (e.g. *you no* + Verb), and at other times disrupt them. Eskildsen and Wagner (2015) analysed students' gestures as well as speech to document their growing understanding and use of English prepositions with complex meanings (*under, across*).

Our second corpus example was created by Collins et al (2009). This corpus comprises audio- and videorecordings of Grade 6 English lessons for Francophone children (aged 11-12) in Quebec. For these younger learners, the classroom was effectively the only source of English input. The researchers recorded and transcribed ca. 40 hours of instructional input (the students' own speech was not the focus of their research). Their research took a usage-based perspective on L2 acquisition. According to usage-based theory, item frequency, lexical properties and perceptual saliency in L2 input are important factors influencing learner uptake. Collins et al examined the frequency and saliency in teacher talk of three English features: the simple past (regular *-ed*), the possessive determiners *his/her*, and progressive *-ing*. Previous studies have shown that *-ed* and *his/her* are difficult for French-speaking learners to acquire, while progressive *-ing* is more easily learned; could this apparent difference in learnability be explained by differences in their input profiles? Regarding raw frequency, it turned out that there was little difference between the three features; all were relatively infrequent in ongoing classroom input. However, progressive *-ing* occurred with a good variety of common lexical verbs and was always fully articulated. It was therefore judged to be salient both in terms of lexis and phonology. In contrast, regular simple *-ed* past (e.g. *asked*) was much rarer than irregular past forms (e.g. *made, did*), and occurred frequently with only four different verbs. When it did occur, the *-ed* ending was normally unstressed and frequently elided. The possessive determiners *his/her* occurred very rarely with semantic transparency (e.g. *his wife, her father*); they were rarely stressed, and almost always unaspirated (*he rode 'is bike*). Overall, these researchers claimed that the learnability differences between these elements could most likely be explained by their lexical properties and other aspects of saliency in L2 classroom input; their pedagogical recommendations focus on how best to promote saliency.

A classroom corpus is a set of audio- or videorecordings of authentic L2 lessons, which have been transcribed and prepared for analysis. Such a corpus provides direct insight into how the L2 curriculum is delivered and experienced in the classroom in real time.

Taken together, these two examples of classroom corpora illustrate the light shed on classroom input and interaction, and their potential for informing pedagogical strategies. It should be noted however that neither of these studies collected complementary data from the classroom learners in the form of standardised tests, so the only available evidence on L2 development was the classroom contributions of the learners themselves. The studies by Hellermann, Eskildsen and Wagner use qualitative analysis of student contributions over time to illustrate aspects of L2 development, while Collins et al. do not investigate students' actual learning of the items which were investigated. Clearly where classroom corpora can be complemented with test evidence, more powerful conclusions could be drawn about the relationship between classroom experience and L2 development.

The 'Learning French' (*LF*) corpus

The classroom corpus presented here comprises 33 L2 French lessons taught to a class of 26 children aged 7-8 in an English primary school. The data was collected as part of the longitudinal research project 'Learning French from ages

Rosamond Mitchell is Professor Emeritus of applied linguistics at the University of Southampton. She has research interests in classroom language learning, and in study abroad and its impact. Together with Florence Myles, she has for many years promoted corpus-based approaches to study L2 learning.

Florence Myles is Professor Emerita of Second Language Acquisition at the University of Essex. Her recent research has compared how children of different ages learn foreign languages in the classroom, in terms of grammar, vocabulary, attitudes and motivation.

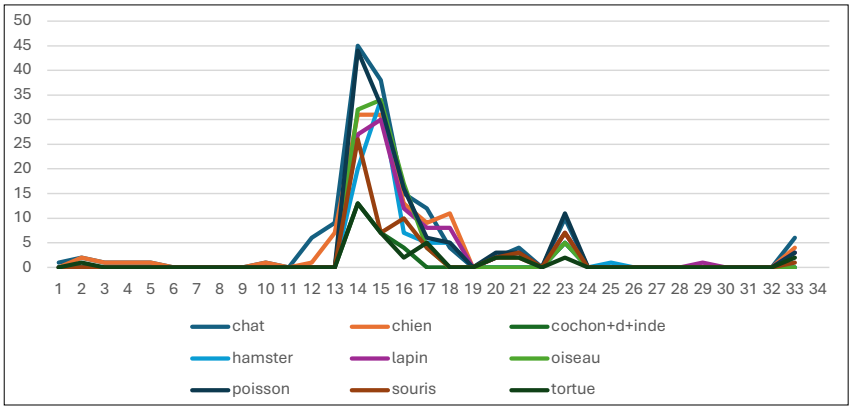


Figure 1

Distribution of animal vocabulary in *LF* corpus

This investigation of the vocabulary of classroom input raises key pedagogic issues. Firstly, how far the curriculum should be tailored specifically to the current interests of young learners, how far it should look ahead to future needs (e.g. as reflected in a reference corpus such as LLB); and secondly, how 'rich' and personalised classroom input should be, once curriculum topics have been selected.

5, 7 and 11: An investigation into starting ages, rates and routes of learning' (Myles, 2017). The project also repeatedly collected data on children's L2 development through an Elicited Imitation task, oral interviews and a Receptive Vocabulary Test (RVT). Data was also gathered on learners' L1 literacy and their working memory capacity. The learners were monolingual L1 English speakers, with no extramural contact with French; a pretest had established that the children knew no French at all before teaching began. The specialist teacher used an oracy-led approach, and spoke both English and French during instruction. The 33 lessons were audio- and videorecorded, and transcribed using CHAT/CLAN conventions (MacWhinney, 2000).

Exploitation of the corpus

Here we bring together a number of studies conducted using this corpus. The first study to be presented investigated lexical characteristics of the teacher's L2 input (for a fuller account of this study, see Mitchell & Myles, 2023). The second study investigated children's learning of French vocabulary, and its relationship with aspects of teacher input (Mitchell & Rule, 2022). The third study took a case study approach to explore children's classroom engagement, and its relationship with L2 development (Mitchell & Myles, 2019).

Teacher speech as L2 input

The teacher's speech was the prime source of L2 input for learners in this setting. To analyse her spoken French, we initially used the CLAN computer programs. This provided counts of word types (n = 653) and tokens (n = 44,316), which were also tagged for parts of speech. Token frequency was highest for function words including definite and indefinite articles, the first-person singular pronoun *je* [I] , the verbs *avoir* [to have] and *être* [to be], and the discourse marker *bien* [good, fine]. Many content words related to curriculum topics such as food and drink (60 word types), the body (27), animals (30), colours (15) etc. Within these thematic groups, word frequency was quite variable; typically a smaller cluster was the focus of intentional instruction and those words occurred with high frequency, while others might occur once or twice only. For example, among animal vocabulary, the items *chat, chien, lapin, poisson* [cat,

dog, rabbit, fish] occurred over 100 times each, while *canard, dinosaure* [duck, dinosaur] occurred only once. While function words were distributed relatively evenly through the entire corpus, the occurrence of content words was connected to lesson topics. For example, Figure 1 shows the distribution of selected animal names; these occurred with high frequency in Lessons 14-17, when the curriculum topic was 'pets', animal stories were read and animal songs were sung, but were rarely retrieved and re-practised later.

Next, the teacher's French vocabulary was compared with the vocabulary of two French reference corpora, created to support L2 acquisition. These corpora were the *Frequency Dictionary of French* by Lonsdale and Le Bras (*LLB*: 2009), and *FLELex*, from the University of Louvain (François et al., 2014; Pintard & François, 2020). The *LLB* dictionary presents the 5,000 commonest words in a large corpus of (adult) contemporary French, while *FLELex* provides wordlists related to the six proficiency levels of the Common European Framework of Reference for Languages (*CEFR*), derived from a corpus of French teaching materials. Our analysis explored the relationship between the teacher's vocabulary in the *LF* corpus, the first 2,000 words from *LLB*, and the wordlists for *CEFR* levels A1 and A2 from *FLELex* (which together total 1,925 words).

This analysis showed an overlap of just under 60 % of word types between the *LF* wordlist and the *LLB* 1-2k lists, and an overlap of just under 70 % with *FLELex* AI-A2. (Calculated on the basis of word tokens, the overlaps were much higher, over 80 % in both cases.) The differences derived partly from the curriculum, which gave much richer treatment to themes such as animal names (30 types in *LF*, 11 in *FLELex* A1-A2, and just two in *LLB* 1-2k), foods, etc. Words relating to the classroom environment and classroom management were also richer in *LF* (90+ words in LF, 60+ in each reference corpus); an important group of words relating to emotions were uncommon in *LLB* 1-2k, though more likely to be found in *FLELex* A1-A2 (see Table 1 for details).

This investigation of the vocabulary of classroom input raises key pedagogic issues. Firstly, how far the curriculum should be tailored specifically to the current interests of young learners, how far it should look ahead to future needs (e.g. as reflected in a reference corpus such as *LLB*); and secondly, how 'rich' and personalised classroom input should be once curriculum topics have been selected. As illustrated in Table 1, a good proportion of the word types in teacher speech occur with very low frequencies (fewer than 10 times, and thus less likely to lead to incidental learning: Peters, 2020); what balance could/should be struck between such incidental enrichment, on the one hand, compared to the need for recycling and distributed practice of 'target' vocabulary?

Item	**Word class**	**Frequency in *LF* input**	***LLB* band**	***FLELex* CEFR band**
aimer	v	345	1k	A1
ami	n	220	1k	A1
mal	adj/n	101	1k	A1
adorer	v	59	3k	A1
calin	n	57	Ø*	C2
coeur	n	53	1k	A1
bisou	n	43	Ø	A2
désirer	v	35	2k	A1
détester	v	30	3k	A1
beau	adj	26	1k	A1
parfait	adj	16	2k	A1
génial	adj	12	4k	A1
joli	adj	12	3k	A1
joyeux	adj	10	4k	A1
méchant	adj	9	4k	B1
anniversaire	n	9	3k	A1
barbant	adj	8	Ø	Ø
laid	adj	7	Ø	B1
amour	n	7	1k	A1
délicieux	adj	4	Ø	A1
mignon	adj	4	Ø	A2
fatigué	adj	3	4k	A1
triste	adj	3	2k	A1
embrasser	v	3	4k	A1
content	adj	2	2k	A1
gentil	adj	2	3k	A1
amuser	v	2	3k	A1
fatigué	adj	1	3k	A1
dégoutant	adj	1	Ø	C2
moche	adj	1	Ø	C1

*The symbol Ø denotes non-occurrence

Table 1
Emotion words in *LF* teacher talk compared with banding in reference corpora.

Item	Facility at Post-Test	Facility at Delayed Post-Test	LF input frequency	Number of lessons	Pedagogic activities	Multimodal support
poisson 'fish'	96.15	76.92	124	9	Focused oral practice Meta-comment Incidental use (song, film, story, game) Drawing and labelling	Iconic gesture (swimming) Image (flashcards, story) Text (image labels) Text (story)
glace 'ice cream'	92.31	69.23	35	5	Focused oral practice Incidental use (games, role play) Drawing and labelling	Image (flashcards, whiteboard images) Imitation foods Text (image labels)
frapper 'to clap'	69.23	73.08	37	10	Incidental use (song, game)	Action (handclapping)
parler 'to speak'	34.62	34.62	187	18	Incidental use (song, classroom management)	None
main 'hand'	30.77	19.23	160	20	Focused oral practice Incidental use (song, classroom management, game)	Actions (handclapping, hand raising) Pointing/touching own body

Table 2
Attributes of well-learned and poorly-learned words (after Mitchell & Rule, 2022)

Drivers for children's vocabulary learning

The children's L2 vocabulary development was assessed in the wider 'Learning French' project by the Receptive Vocabulary Test (RVT). This 50-item test included nouns and verbs only, was specially constructed to reflect the teacher's own vocabulary use, and was administered 3 times (Mid-programme, as a Post-test, and as a Delayed Post-test. Forty-four RVT items were drawn from the teacher's lexical inventory; the remaining 6 items were not heard in class. Ten items were judged to be cognates (e.g. *bébé* [baby]). The computer-based test had a multiple-choice format; children saw a selection of 4 images, heard the target word, and selected the matching image. At Post-test and Delayed Post-test, the group mean score was just over 50 % each time. However, of key interest to the study was the variation in learning of individual items. On the basis of their facility values, i.e. proportion correctly known, all words in the test were allocated to one of 3 categories: 'well-learned', 'moderately-learned', 'poorly-learned'.

To explore the drivers promoting learning of vocabulary, a quantitative approach was adopted first. This investigated the relationship between a word's facility value in RVT, showing how well it had been learned, its frequency in classroom L2 input, and its status as cognate or non-cognate. We know that input frequency is positively related to vocabulary learning as far as L2 reading is concerned, though there is less evidence for aural input (Peters, 2020). Cognate words are also generally easier to learn than non-cognates (e.g. Cobb, 2000). A statistical test was carried out on the Post-Test results which explored how far success in learning individual words could be attributed to frequency in teacher input and to cognate status. The results were significant and showed that these two variables in combination explained around one-third of the variation in learning success for the tested words[1].

To exploit the classroom corpus more fully, a qualitative approach was also adopted, to more closely examine a subset of five words which had proved easier/more difficult to learn, according to their facility values at Post-test and Delayed Post-test. All lessons in which the selected words occurred were identified, and all occasions of use were scrutinized on video to identify the activity types in which the words appeared, as well as any supporting gestures, images, objects, or printed text. Table 2 provides an overview of these five items, their facility values at Post-test and Delayed Post-test,

1 The test was a standard multiple regression with teacher input frequency and cognate status as predictor variables, and facility values on the Post-test as criterion variable. Results were significant and showed that the two variables in combination explained 35.3 % of the variance in test scores (R^2 = 0.353, adjusted R^2 = 0.323, p < 0.001).

	Bruno* (m)	Roseline (f)	Capucine (f)	Maxence (m)	Faustine (f)	Xavier (m)
Birth date	Sept 2001	Feb 2002	Feb 2002	Jan 2002	Aug 2002	May 2002
WM score (max 28)	24	20	18	15	9	6
L1 literacy score (max 9)	8	9	8	6	2	3
RVT Post-test/ Delayed Post-test score (max 50)	40/33	34/32	33/36	18/20	25/27	14/13
EI test score (max. 465)	391	349	289	262	282	244

*Names are pseudonyms

Table 3
Attainment of 6 case study children
(after Mitchell & Myles, 2019)

input frequency, distribution through the lesson sequence, related activities and multimodal support.

Table 2 shows that the best-learned words, *poisson* and *glace*, were the subject of focused oral practice. In addition, they were used incidentally and were supported by written text (including children's own label-writing) as well as by images and gestures. The less-well learned verbs *frapper* and *parler* were never the subject of focused practice, nor were they supported with writing or images. The worst-learned word, *main*, was, however, the subject of focused oral practice and associated with gestures (notably, hand-raising in response to the classroom command *levez la main*! [hands up!]). While each individual word may present a slightly different learning challenge, it seems that combinations of focused practice, multimodal support, and the chance to link oral and written forms can promote the learnability of L2 words, but do not guarantee it.

Learner engagement and L2 development

As mentioned earlier, the wider 'Learning French' project collected data on children's working memory, using a Non-Word Repetition test (Gathercole & Baddeley, 1996); the school provided assessments of their L1 literacy levels on a standard scale from 1-9. These factors were found to be significant predictors of children's L2 development as measured by the RVT, accounting for almost half the variance in learners' test scores (Mitchell & Rule, 2022). However, the classroom corpus allowed us to explore another potentially very important factor which could influence learner development, i.e. their ongoing engagement in classroom activities.

Learner engagement is central to classroom learning more generally; it is a complex construct, with behavioural, emotional, and cognitive dimensions (Fredricks, Blumenfeld & Paris, 2004). To study learner engagement, six case study children were selected, with high, mid, and low scores for L1 literacy and working memory. Table 3 provides details together with the scores achieved by these children on the RVT and the Elicited Imitation test.

Table 3 shows that the three children with the highest WM and L1 literacy scores also have the highest French test scores. However, the youngest child, Faustine, with very low WM and L1 literacy scores, considerably out-performs expectations on the French tests. So did classroom engagement have a role to play in moderating the influence of WM and L1 literacy?

While each individual word may present a slightly different learning challenge, it seems that combinations of focused practice, multimodal support, and the chance to link oral and written forms can promote the learnability of L2 words, but do not guarantee it.

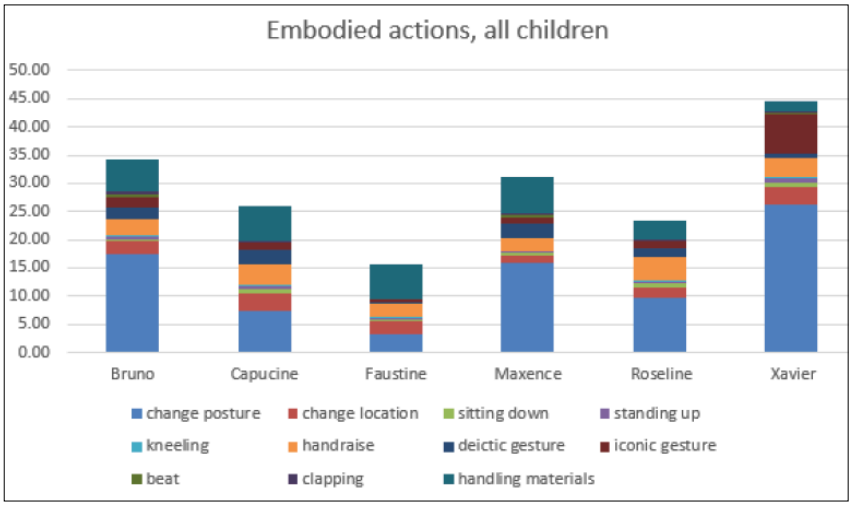


Figure 2
Bodily movement as behavioural (dis)engagement (after Mitchell & Myles, 2019)

To analyse classroom engagement, we tracked the children individually through a video sample of 100 minutes each, drawn from 5 spaced lessons. Using the computer program ELAN (Wittenburg et al., 2006) we developed a coding scheme to track behavioural engagement, including children's gaze and bodily movements. Analysis of children's gaze showed that the four highest achievers were attending either to the teacher or to the whiteboard or screen at least 70 % of the time. The two lowest achievers (Maxence, Xavier) were attending to these French input sources just over 60 % of the time. The children's bodily movements are summarized in Figure 2 (scale is % of observed time). Some movements such as hand-raising indicate engagement, but 'change posture' covers behaviours such as self-touching, shifting in seat, which likely indicate mild distraction. Figure 2 shows a noticeable gender difference, with boys generally more restless than girls; it seems that bodily restlessness up to quite a high level is not necessarily a barrier to learning (Bruno), but in the case of Xavier, may have passed a threshold beyond which learning is depressed.

Emotional and cognitive engagement were studied through qualitative analysis of critical incidents. All six children were well engaged emotionally, seeking teacher approval and taking part in games, competitions and other 'fun' activities. For all four higher achievers, incidents involving planning and reflection were observed and interpreted as indicators of cognitive engagement. For example, Bruno was observed privately rehearsing new vocabulary and associated gesture; Capucine volunteered a planned monologue of several utterances in French, following a holiday break; Faustine was seen organising materials (unasked) in response to teacher detailed instructions for a new activity. For the two lowest achievers, no such incidents were observed.

Discussion

This paper has illustrated some possibilities of research using classroom corpora. Corpus creation must follow relevant ethical guidance, which may be particularly strict regarding young learners (see e.g. BAAL, 2021; BERA, 2018). Thus for example, in the case of the *LF* corpus, only transcripts but not videos may be generally shared. However once created, and subjected to different types of analysis, many pedagogical implications can be drawn, and some examples are given below.

Firstly, the studies discussed here show the need for instructors to reflect on the most appropriate balance between explicit instruction and incidental exposure, for both grammar and vocabulary. Saliency is always crucial, and pedagogic strategies which improve the saliency of new material, including multimodal strategies, need to be developed and evaluated.

Secondly, instructors need to reflect on the choice of curriculum topics and implications for vocabulary instruction. To what extent should topics reflect the current interests of students or rather prepare them for future L2 communicative needs? What should the balance be between prescribed wordlists derived from reference corpora, and 'personalised' vocabulary? Given what is known about the importance of both frequency in input, and dispersed practice, for vocabulary acquisition, how 'rich' should topic-specific vocabulary be? These are complex questions and corpus findings point us to some extent in different directions. For example, Collins et al. suggest that a richer selection of (regular) verbs in teacher input would be helpful in promoting learning of English past tense morphology. On the other hand, some of the rich topic-related vocabulary

in the 'Learning French' corpus (e.g. for animals, and foods) was used only incidentally and never practised; it seems likely that working with more limited lists could promote learning more systematically. More classroom research is clearly needed here.

Finally, the direct study of learner engagement made possible in a video corpus reminds us firstly of the challenges of managing young learners' behavioural engagement. We also see the importance of emotional engagement through positive teacher encouragement and 'fun' activities. And lastly, but of greatest importance, we see the need to work actively with children to develop the qualities of cognitive engagement, such as the ability to think ahead, manage resources and plan their own learning, to learn with and from peers, or to envision longer term learning goals and develop the perseverance needed to attain them.

References

British Association for Applied Linguistics (BAAL) (2021). *Recommendations on good practice in applied linguistics*. 4th Edition. Accessible at https://www.baal.org.uk/who-we-are/resources/

British Educational Research Association (BERA) (2018). *Ethical guidelines for educational research*. 4th Edition. Accessible at https://www.bera.ac.uk/publication/ethical-guidelines-for-educational-research-2018

Cekaite, A. (2022). Early language education and language socialization. In M. Schwartz (Ed.), *Handbook of Early Language Education* (pp. 143-165). Cham: Springer International Publishing.

Cobb, T. (2000). One size fits all? Francophone learners and English vocabulary. *The Canadian Modern Language Review*, 57(2), 295-324.

Collins, L., Trofimovich, P., White, J., Cardoso, W., & Horst, M. (2009). Some input on the easy/difficult grammar question: An empirical study. *The Modern Language Journal, 93*(3), 336-353.

Eskildsen, S. W. (2012). L2 negation constructions at work. *Language Learning, 62*(2), 335-372.

Eskildsen, S. W., & Wagner, J. (2015). Embodied L2 construction learning. *Language Learning, 65*(2), 268-297.

François, T., Gala, N., Watrin, P. & Fairon, C. FLELex: A graded lexical resource for French foreign learners. In *the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavik, Iceland, 26-31 May.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59-109.

Hellermann, J. (2008). *Social practices for language learning*. Clevedon, UK: Multilingual Matters.

Kunitz, S., Markee, N., & Sert, O. (2021). *Classroom-based conversation analytic research: theoretical and applied perspectives on pedagogy*. Cham, Switzerland: Springer Nature.

Lonsdale, D., & Le Bras, Y. (2009). *A frequency dictionary of French: Core vocabulary for learners*. Abingdon: Routledge.

MacWhinney, B. (2000). *The CHILDES project: Tools for analysing talk. 3rd Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mitchell, R., & Myles, F. (2019). Learning French in the UK setting: Classroom engagement and attainable learning outcomes. *Apples: Applied Language Studies, 13*(1), 69-93.

Mitchell, R. & Myles, F. (2023). Lexical input in the primary languages classroom. Paper presented at Annual Meeting of the British Association for Applied Linguistics, York, August 2023.

Mitchell, R., & Rule, S. J. (2022). Learning vocabulary in the primary languages classroom: What corpus analysis can tell us. In K. McManus & M. Schmid (Eds.), *How special are early birds? Foreign language teaching and learning (EuroSLA Studies 6)* (pp. 37-61). Berlin: Language Science Press.

Myles, F. (2017). Learning foreign languages in primary schools: is younger better? *Languages, Society and Policy, 1*(1), 1-8.

Pekarek Doehler, S., & Fasel Lauzon, V. (2015). The development of L2 interactional competence: Evidence from turn-taking organization, sequence organization, repair organization, and preference organization. In N. Markee (Ed.), *The handbook of classroom discourse and interaction* (pp. 409-424). Malden, MA: Wiley Blackwell.

Peters, E. (2020). Factors affecting the learning of single-word items. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 125-142). Abingdon/ New York: Routledge.

Pintard, A. and François, T. (2020). Combining expert knowledge with frequency information to infer CEFR levels for words. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)* (pp. 85-92).

Reder, S., Harris, K., & Setzler, K. (2003). The multimedia adult ESL learner corpus. *TESOL Quarterly, 37*(3), 546-557.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. (2006). *ELAN: a Professional Framework for Multimodality Research*. In: Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.

Primula

GRADED READERS IN THE EFL CLASSROOM – THE EXAMPLE OF MARY SHELLEY'S *FRANKENSTEIN*

Graded Readers bieten Schülerinnen und Schülern im EFL-Unterricht die Möglichkeit, sich mit literarischen Texten auseinanderzusetzen, die im Original für sie zu anspruchsvoll wären. Doch zu welchem Preis wird diese Lernmöglichkeit erkauft? Die linguistische Analyse, welche im Rahmen der Masterarbeit durchgeführt wurde, vergleicht verschiedene Graded Readers von Mary Shelleys *Frankenstein* mit dem Original und zeigt erhebliche Vereinfachungen auf lexikalischer und syntaktischer Ebene auf. Dies hat Auswirkungen auf die literarische Aussagekraft dieses faszinierenden Prosatextes. Unter anderem werden die Emotionen der Figuren in den Graded Readern weniger deutlich herausgearbeitet. Um diesen Verlust an literarischem Material auszugleichen, wurde ein Unterrichtsprojekt entwickelt, welches den Prinzipien des TBLT folgt.

Janina Liechti
University of Zurich

Janina Liechti earned her MA in Secondary Education from the Zurich University of Teacher Education (PHZH) and is currently enrolled in Educational Sciences at the University of Zurich. She is working as a tutor at the institute of Educational Sciences and is interested in (digital) media education.

Although the usage of literature in the EFL classroom has long been advocated for because of its potential to foster both linguistic and transversal competences (cf. Claridge, 2012; Hill, 1997; Grimm et al., 2015; Richards & Smiths 2002), the in-classroom application of literature has yet to find broader appeal as it still faces many challenges. Two of these challenges concern the thematical appeal of stories and the linguistic difficulty of texts (Grimm et al., 2015). In order to overcome the latter, teachers often look to graded readers. The master's thesis presented here aims to examine two research questions regarding graded readers in the EFL classroom. Firstly, it examines in what ways graded reader versions differ from the original text in terms of their linguistic and literary material. Secondly, on the basis of these findings, it explores how the literary material which is lost during the process of adaptation can be compensated for through adequate tasks. The literary text used for this master's thesis is Mary Shelley's *Frankenstein* because it offers compelling storylines and characters. In addition, it deals with relatable topics such as relationships and hardships. Therefore, the pupils are likely to be interested in the story as it is thematically appealing to them (cf. Grimm et al., 2015). Secondly, Shelley's *Frankenstein* was selected because it has been turned into a variety of graded readers, which permits teachers to easily adapt the linguistic difficulty of the text to the pupils' language proficiencies. In order to examine how these versions differ from one another, a linguistic and literary analysis have been conducted, which then served as the basis for the development of classroom material.

Three key passages of Shelley's *Frankenstein* were compared to the respective passages in different graded readers. The scenes include the birth of Victor's creature, the creature's request for a companion, and the death of Elizabeth. Each of those scenes also received its respective worksheet in the classroom project. All findings of this thesis will be illustrated through the example of the

first passage. In this scene, Victor brings the creature to life for the first time. His initial excitement over witnessing the fruits of his labour is quickly replaced by horror before he flees and leaves the creature alone.

Linguistic analysis

Based on publishers' principles of grading texts as well as on claims by Dirks (2004) and Hill (2008), three hypotheses were formulated. The first one focuses on the lexical level and claims that the difficulty of words decreases along with the language level of the respective graded reader (cf. Hill, 2008). More difficult versions should therefore use more difficult words on average. As seen in Table 1, this hypothesis could only be partially verified. Although the original clearly uses the most difficult lexis whilst the easiest version provided by Starry Forest books uses the least difficult lexis (as can be seen in the percentage of A1 to C2 words used in the respective texts), Richmond Readers' and Macmillan's versions barely differ. However, across all graded reader versions different kinds of lexical adaptations could be observed. In many cases, difficult words have been either replaced or omitted (with the latter occurring more frequently). For example, in the original, Victor's initial view of the creature is described as follows: "How can I describe my emotions at this catastrophe, or how delineate the wretch whom with such infinite pains and care I had endeavoured to form?" (Shelley, 1993). The Richmond Reader version, on the other hand, describes the scene as follows: "I cannot describe my feelings at that moment" (Shelley, 2012). Although the gist stays the same, Victor's efforts and a first description of his creature have been omitted in the simplified version.

Regarding parts of speech, the second hypothesis claims that the relative number of adverbs and adjectives decreases along with a decrease in language level, as publishers tend to omit anything they consider to be less important to the story during the process of simplification (Lucas, 1991). For the same reason, it is expected that the relative number of verbs increases in the simplified versions. As graded readers reduce descriptions in favour of driving the plot forward, the number of verbs relative to the number

	Original reader Part I, Chapter IV, p. 38 - 39	Richmond Readers (B1) Chapter 5, p. 19 - 21	Macmillan Readers (A2) Chapter 2, p. 14 - 15	Starry Forest Books (A1) p. 3 - 6
Number of words	311 (100%)	153 (100%)	221 (100%)	17 (100%)
Difficulty of words	**A1** 193 (62%) **A2** 36 (11,6%) **B1** 20 (6,4%) **B2** 13 (4,1%) **C1** 7 (2,3%) **C2** 7 (2,3%) **Unlisted** 35 (11,3%)	**A1** 113 (73,9%) **A2** 21 (13,9%) **B1** 15 (9,9%) **B2** 2 (1,3%) **Unlisted** 2 (1,3%)	**A1** 156 (70,6%) **A2** 35 (15,8%) **B1** 23 (10,4%) **B2** 4 (1,8%) **C2** 3 (1,4%)	**A1** 7 (41,2%) **A2** 2 (11,7%) **B1** 2 (11,7%) **B2** 1 (5,9%) **Unlisted** 5 (29,4%)
Number of sentences	13	15	34	5
Average number of words per sentence	23,9	10,2	6,5	3,4
Main clauses	18	18	41	4
Subordinate clauses	11	3	1	0
Noun	68 (21,1%)	35 (25,9%)	47 (21,3%)	6 (35,3%)
Article / determiner	50 (15,6%)	20 (13%)	40 (18,0%)	2 (11,8%)
Adjective	31 (9,6%)	14 (9,2%)	22 (10%)	0 (0%)
Verb	52 (16,1%)	30 (19,6%)	49 (22,1%)	4 (23,5%)
Adverb	19 (5,9%)	10 (6,5%)	13 (5,8%)	1 (5,9%)
Preposition	37 (11,5%)	14 (9,2%)	10 (4,5%)	0
Pronoun	27 (8,4%)	17 (11,1%)	19 (8,5%)	1 (5,9%)
Conjunction	22 (6,8%)	9 (5,8%)	16 (7,2%)	0
Other (Interjections, numerals, infinitive markers)	5 (1.6%)	4 (2,6%)	5 (2,3%)	4 (23,5%)

Table 1

Linguistic analysis. Adapted from "Graded readers in the EFL classroom. A literary and linguistic analysis of Mary Shelley's Frankenstein, or The Modern Prometheus *and its graded reader versions and a practical implication into the EFL classroom."* (Justus, 2023)

of words should automatically increase. The second hypothesis could not be verified. The total relative number of adverbs and adjectives does not vary significantly between the different versions. If we examine the numbers, adverbs account for 5,9% of all words in the original, for 6,5% in the Richmond Readers edition, and for 5,8% in the Macmillan version. The relative number for adverbs is therefore similar throughout all versions.

The third hypothesis, which claims that the relative number of verbs in each passage increases with a decrease in CEFR language level, could also only be partly verified. While it is the case in Table 1 with verbs accounting for 16,1% of all words in the original, 19,6% in Richmond Readers, 22,1% in Macmillan, and 23,5% in Starry Forest books, another passage has shown that the original may even use the most verbs relative to the total number of words.

Lastly, with a focus on sentence structure, three claims made by Hill (2008) and Dirks (2004) were examined, namely that 1) the relative number of main and subordinate clauses increases to the relative number of sentences with increasing language level, 2) that the sentence length decreases with language level, and 3) that easier versions use more dialogue to tell the plot. All three claims have been proven to be true. The original scores highest in relative number of main and subordinate clauses to relative number of sentences and highest in sentence length. Both factors also increase or decrease along with the CEFR language level. Furthermore, the easier versions also use more dialogue to tell the plot. However, the difference between the original, Richmond Readers, and Macmillan are not substantial. Starry Forest Books' version is the only one that mainly uses dialogue to tell the plot.

Literary analysis

The lexical and syntactical differences also influence the representation of the literary material. Overall, great changes could be discovered on the literary level. While there are substantial overlaps with regard to direct characterisation, changes can be observed on the level of indirect characterisation, as most characters talk, behave, and look differently than in the original. For example, in the first passage, Victor is mainly characterised indirectly through his speech throughout all versions. However, whereas his speech pattern and vocabulary indicate that he is eloquent, observant, smart, lonely, as well as fanatic, anxious, disgusted and disappointed in the original, he is reduced to being smart, observant, anxious and disappointed in Richmond Readers. In the Macmillan version, his personality undergoes further changes. While he remains smart and observant, his emotions switch from anxious, disgusted and disappointed to excited and shocked. In Starry Forest Books he even loses all linguistic characterisation. Moreover, it is important to note that emotions are far less prevalent in the graded reader versions than in the original. This is due to the linguistic adaptation. As the texts are significantly abbreviated, fewer words are available to explore many of Victor's difficult emotions. Also, by using simpler synonyms, nuances of Victor's emotions are lost. Given that Shelley's original is strongly driven by the characters' emotions, the reduction and alteration of emotions is a huge literary loss.

While there are substantial overlaps with regard to direct characterisation, changes can be observed on the level of indirect characterisation, as most characters talk, behave, and look differently than in the original.

Classroom Project Outline

Based on the previous analyses, a classroom project was designed. Firstly, it wants to recover the literary meaning which has been lost during the linguistic simplification by implementing tasks

which heavily focus on the characters' emotions. Secondly, it aims to foster both subject-specific and transversal competences of curriculum 21, namely *Writing, Speaking, Reading,* and *Empathy*.

The classroom project has been outlined for secondary school students in the 13 – 15 age bracket with varying English language skills. The heterogeneous composition of the class enables showcasing the potential of graded readers in heterogeneous settings. To foster both the subject-specific and transversal skills, the entire classroom project follows the guidelines of Task-based language teaching (TBLT). The specific activities vary depending on learning goal and targeted competency. The strong focus of TBLT on meaning- and comprehension-focused in- and output served as the main reason for choosing this method (Grimm et al., 2015). It strongly correlates with findings that attribute a great learning potential to graded readers when they are used in connection with meaning-based approaches (Nation & Deweerdt, 2001). With regard to fostering pupils' ability to empathise with literary characters, TBLT is also a suitable choice, as this teaching approach fosters holistic learning (Vogt, 2017).

The classroom project itself was designed around Macmillan's and Richmond Readers' graded reader versions of Shelley's *Frankenstein*. In order to develop the pupils' language skills, a combination of pre-, while- and post-activities encourage the learners to emotionally and cognitively engage with the material and the characters. With regard to empathy learning, Jamieson (2015) points out that pupils need to be presented with learning activities that focus on the characters' thoughts and emotions for the learners to develop empathy skills. This is achieved through a variety of tasks that ask the pupils different questions about the characters (see Figure 1, for example).

In order to ensure pupils' emotional and cognitive engagement with the characters, their literal understanding of the text needs to be secured first.

In order to ensure pupils' emotional and cognitive engagement with the characters, their literal understanding of the text needs to be secured first. Therefore, all empathy-focused tasks are preceded by a variety of other tasks that focus on the text as a whole. There, pupils work on their vocabulary, make assumptions, check the validity of different statements, summarise the passages, among other

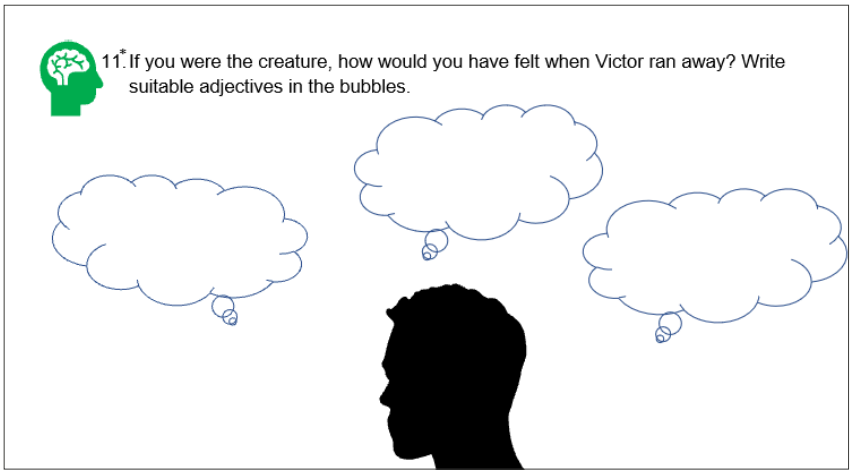


Figure 1

Example task. From *"Graded readers in the EFL classroom. A literary and linguistic analysis of Mary Shelley's* Frankenstein, or The Modern Prometheus *and its graded reader versions and a practical implication into the EFL classroom."* (Justus, 2023)

things. Only after the pupils have developed an understanding of the passage are they encouraged to talk, think, and write about the emotions and thoughts of the characters. Most tasks were also designed to be completed or discussed cooperatively in order to allow pupils to further work on their own empathy skills. By discussing their interpretations with others, they need to verbalise their thoughts whilst also listening and acknowledging other people's interpretations of the same passage.

Discussion

As seen in the results above, the linguistic simplifications have consequences on both the linguistic and literary material. Although some scholars argue against the use of graded readers in the EFL classroom for these reasons (cf. Yano et al, 1994, Nation & Ming-Tzu, 1999), this master's thesis heavily encourages the use of graded readers in meaning- and comprehension-focused settings. Moreover, graded readers are an optimal learning resource for those approaches. The linguistic changes enable pupils to engage with texts and themes they would not encounter otherwise, as the linguistic requirements are too demanding. Additionally, even though the literary material deviates from the original, teachers can recover the most important themes and meanings through a variety of tasks. If complemented with tasks such as the ones outlined above, graded readers can be a compelling and effective learning resource to develop both language and transversal competences.

Bibliography

Claridge, Gillian. 2012. "Graded Readers: How the Publishers Make the Grade." *Reading in a Foreign Language* 24 (1): 106 – 119.

Grimm, Nancy, Michael Meyer and Laurenz Volkmann. 2015. *Teaching English*. Tübingen: Narr Francke Attempto.

Hill, David R. 2008. "Survey Review: Graded Readers in English." *ELT Journal* 62 (2): 184 – 204.

Dirks, Ines. 2004. *Graded Readers: Text Structure, Use and Perception*. Lizentiatsarbeit. Zürich: Universität Zürich.

Lucas, Michael A. 1991. "Systematic Grammatical simplification." *International Review of Applied Linguistics in Language Teaching (IRAL)* 29 (3): 241-248.

Jamieson, Alanna. 2015. "Empathy in the English Classroom: Broadening Perspectives Through Literature." *LEARNing Landscapes* 8 (2): 229 – 244.

Justus, Janina. 2023. *Graded Readers in the EFL Classroom*. Zürich: Pädagogische Hochschule Zürich.

Nation, Paul and Jean Paul Deweerdt. 2001. "A Defence on Simplification." *Prospect* 16 (3): 55 – 67.

Nation, Paul and Karen Wang Ming-Tzu. 1999. "Graded readers and Vocabulary." *Reading in a Foreign Language* 12 (2): 355 – 379.

Richards, Jack. C. and Richard Schmidt. 2002. *Longman Dictionary of Language Teaching and Applied Linguistics*. 3rd ed. London: Longman, Pearson Education.

Shelley, Mary. 1993. *Frankenstein, or The Modern Prometheus. The 1818 Text.* Ed. Marilyn Butler. Oxford: Oxford University Press.

Shelley, Mary. 2005. *Frankenstein*. Adapted by Margaret Tarner. London: Macmillan Readers.

Shelley, Mary. 2012. *Frankenstein*. Adapted by Pam Davies. Oxford: Richmond Readers.

Shelley, Mary. 2021. *Frankenstein*. Adapted by A.H. Hill. New York: Starry Forest Books.

Vogt, Karin. 2017. "Inklusion und Heterogenität im Englischunterricht der Sekundarstufe." In *Inklusion, Diversität und das Lehren und Lernen fremder Sprachen*, published by Eva Burwitz-Melzer, Frank G. Königs, Claudia Riemer, Lars Schmelter, 326 – 336. Tübingen: Narr Francke Attempto.

Yano, Asukata, Michael H. Long and Steven Ross. 1994. "The Effects of Simplified and Elaborated Texts on Foreign Language Reading Comprehension." *Language Learning* 44 (2): 189 – 219.

The linguistic changes enable pupils to engage with texts and themes they would not encounter otherwise, as the linguistic requirements are too demanding.

IMPRESSUM

EDITORE

Association Babylonia Suisse

WWW.BABYLONIA.ONLINE

GRAFICA

Acadabra communication visuelle
www.acadabra.ch

CONCETTO GRAFICO

Filippo Gander | distillerie grafiche
www.distillerigrafiche.ch
filippo.gander@gmail.com

AUTORI E AUTRICI DI QUESTO NUMERO

Gaëtanelle Guilquin | UCLouvain
gaetanelle.gilquin@uclouvain.be
Nina Hicks | Université de Fribourg
nina.hicks@unifr.ch
Karine Lichtenauer | Université de Genève
karine.lichtenauer@unige.ch
Christina Lindqvist | Göteborgs universitet
christina.lindqvist@sprak.gu.se
Gwendoline Lovey | PH FHNW
gwendoline.lovey@fhnw.ch
Rosamond Mitchell | University of Southampton
r.f.mitchell@soton.ac.uk
Florence Myles | University of Essex
fmyles@essex.ac.uk
Mireille Copin | Université de Toulouse Jean Jaurès
mireille.copin@univ-tlse2.fr
Matthias Schwendemann | Universität Leipzig
matthias.schwendemann@uni-leipzig.de
Isabelle Racine | Université de Genève
isabelle.racine@unige.ch
France Rousset | Université de Fribourg
france.rousset@unifr.ch
Inès Saddour | Université de Toulouse Jean Jaurès
ines.saddour@univ-tlse2.fr
Thomas Studer | Université de Fribourg
thomas.studer@unifr.ch
Anita Thomas | Université de Fribourg
anita.thomas@unifr.ch
Franziska Wallner | Universität Leipzig
f.wallner@rz.uni-leipzig.de
Katrin Wisniewski | Universität Leipzig
katrin.wisniewski@uni-leipzig.de