CSP **Center scientific da cumpetenza per la plurilinguitad** Cogniziun Società Formation Bildung Migration Furmaziun Gesellschaft
CSP **Centro scientifico di competenza per il plurilinguismo** Scuola Arbeit Politique Communitad School Travail Ecole Community
CSP **Centre scientifique de compétence sur le plurilinguisme** Migrazione Furmaziun Societad Cognition Society Scola Migration
KFM **Wissenschaftliches Kompetenzzentrum für Mehrsprachigkeit** Societé Cognizione Migraziun Schule Communauté Kognition
RCM **Research Centre on Multilingualism** Formazione Lavoro Politics Comunità Work Politik Lavur Politica Formation Gemeinschaft

Das Schweizer Lernerkorpus SWIKO

Forschungsbericht

Corpus suisse des apprenant·e·s SWIKO

Rapport de recherche

The SWIKO Swiss learner corpus

Research report

Nina Selina Hicks, Thomas Studer

2026

Bericht des Wissenschaftlichen Kompetenzzentrums für Mehrsprachigkeit
Rapport du Centre scientifique de compétence sur le plurilinguisme
Rapporto del Centro scientifico di competenza per il plurilinguismo
Report of the Research Centre on Multilingualism

INSTITUT FÜR MEHRSPRACHIGKEIT
INSTITUT DE PLURILINGUISME
ISTITUTO DI PLURILINGUISMO
INSTITUT DA PLURILINGUITAD
INSTITUTE OF MULTILINGUALISM

UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Herausgeber | Publié par
Institut für Mehrsprachigkeit
www.institut-mehrsprachigkeit.ch
—
Institut de plurilinguisme
www.institut-plurilinguisme.ch

Autor*innen | Auteur·e·s
Nina Selina Hicks, Thomas Studer

Wissenschaftliche Mitarbeit | Collaboration scientifique
Katharina Karges

Das vorliegende Projekt wurde im Rahmen des Forschungsprogramms 2021–2024 des Wissenschaftlichen Kompetenzzentrums für Mehrsprachigkeit durchgeführt. Für den Inhalt dieser Veröffentlichung sind die Autor*innen verantwortlich.

Le projet dont il est question a été réalisé dans le cadre du programme de recherche 2021-2024 du Centre scientifique de compétence sur le plurilinguisme. La responsabilité du contenu de la présente publication incombe à ses auteur·e·s.

Übersetzungen | Traductions
Anaïk Horii - Traduction et révision, tran-scribe (Mary Carozza)

Fribourg | Freiburg, 2026

Layout
Billy Ben, Graphic Design Studio

Unterstützt von | avec le soutien de

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

2026.02 Wetland

Das Schweizer Lernerkorpus SWIKO

Forschungsbericht

Corpus suisse des apprenant·e·s SWIKO

Rapport de recherche

The SWIKO Swiss learner corpus

Research report

Nina Selina Hicks, Thomas Studer

2026

Bericht des Wissenschaftlichen Kompetenzzentrums für Mehrsprachigkeit
Rapport du Centre scientifique de compétence sur le plurilinguisme
Rapporto del Centro scientifico di competenza per il plurilinguismo
Report of the Research Centre on Multilingualism

Inhalt

Das Schweizer Lernerkorpus SWIKO

Forschungsbericht

Nina Selina Hicks, Thomas Studer

1 Einleitung

Das Schweizer Lernerkorpus SWIKO wurde im Rahmen des gleichnamigen Forschungsprojektes (2016-2019) sowie des Folgeprojektes „WETLAND – Weiterentwicklung und Anwendungen“ (2021-2024) am Wissenschaftlichen Kompetenzzentrum für Mehrsprachigkeit erarbeitet. Das übergeordnete Ziel bestand darin, Lernersprache am Ende der obligatorischen Schulzeit mithilfe von Konzepten und Methoden der Korpuslinguistik zu dokumentieren, recherchierbar aufzubereiten und exemplarisch vertieft zu analysieren. Im Fokus standen dabei die Landessprachen Deutsch und Französisch und Englisch als Fremdsprachen. Ein besonderes Augenmerk galt dem Zusammenspiel von Aufgaben, sprachlichen Merkmalen der schriftlichen und mündlichen Produktionen sowie Bewertungen der Produktionen.

Ausgangspunkt für das Forschungsprojekt war der Wandel in der Fremdsprachendidaktik um die Jahrhundertwende hin zum kommunikativen Ansatz, der sich heute, massgeblich beeinflusst vom Gemeinsamen Europäischen Referenzrahmen für Sprachen[1] (Europarat, 2001), in den sprachregionalen Lehrplänen (CIIP, 2023; D-EDK, 2016; Passepartout, 2015) und Unterrichtsmaterialien (z. B. *New World,* Arnet-Clark et al., 2013) widerspiegelt. Der Sprachunterricht zielt seither darauf ab, „Lernende zum Handeln in lebensweltlichen Situationen zu befähigen und sich in ihnen [ihren Sprachen] auszudrücken sowie Aufgaben unterschiedlicher Art erfolgreich auszuführen“ (Europarat, 2020, S. 33; vgl. 2001, Kap. 7). In der Folge hat sich der Stellenwert von Wortschatz und Grammatik verändert: Linguistische Kompetenzen stehen nicht mehr im Mittelpunkt des Unterrichts, sondern sind Mittel zum Zweck des Erreichens kommunikativer Ziele (Ende et al., 2013).

Gleichzeitig wurde die Organisation des Fremdsprachenunterrichts an öffentlichen Schulen der Schweiz mit dem HarmoS-Projekt vereinheitlicht (Konsortium HarmoS Fremdsprachen, 2009). So lernen die Schüler/innen der meisten Kantone ab der 5. und 7. Klasse[2] zwei Fremdsprachen, nämlich eine Landessprache plus Englisch. Am Ende der obligatorischen Schulzeit – in der 11. Klasse im Alter von ca. 15 Jahren – wird von den Schüler/innen erwartet, dass sie in beiden Fremdsprachen insgesamt das GER Niveau A2.2 und im Schreiben das GER Niveau A2.1 erreichen. Während ein nationales Monitoring (Konsortium ÜGK, 2019) sowie verwandte Projekte (u.a. Peyer et al., 2016) das Erreichen dieser kommunikativen Standards untersuchten, ist über spezifische *sprachliche* Kompetenzen wenig bekannt.

Vor diesem Hintergrund lautete die Grundfrage von SWIKO, wie sich die sprachlichen Kompetenzen – insbesondere Wortschatz und Grammatik – im Zuge des kommunikativen Ansatzes am Ende der obligatorischen Schulzeit präsentieren. Mit dieser Fragestellung

1 Im Folgenden kurz GER.

2 Seit HarmoS umfasst die obligatorische Schulzeit insgesamt 11 Jahre: Ab dem 4. Altersjahr zuerst 2 Jahre Kindergarten (ISCED 0), 6 Jahre Primarschule (ISCED 1) und 3 Jahre Sekundarschule (ISCED 2). Folglich sind Teilnehmende der 10. Klasse ungefähr 14-15 Jahre alt und besuchen das zweite Jahr der Sekundarstufe I.

möchte SWIKO einen empirischen Beitrag zum besseren Verständnis des Erwerbs sprachlicher Strukturen im „neuen“ Fremdsprachenunterricht leisten und zu realistischen Erwartungen an die Leistungen der Schüler/innen im sprachformalen Bereich beitragen.

In der ersten Forschungsperiode (2016-2020) wurde eine Datenbank aufgebaut, die aus aufgabenbasierten mündlichen und schriftlichen Lernertexten besteht. Für die Datenerhebung bearbeiteten Lernende der Sekundarstufe I insgesamt acht systematisch variierte Aufgaben unter verschiedenen Produktionsbedingungen (vgl. Kap. 2.3). Anschliessend wurden die Daten zuerst verschriftlicht und dann mit korpuslinguistischen Methoden aufbereitet (vgl. Kap. 2.4).

In der zweiten Forschungsperiode (2021-2024) wurde insbesondere das deutsche Teilkorpus weiter ausgebaut und durch eine halbautomatische Fehlerannotation ergänzt (vgl. Kap. 2.4). Zudem wurden sämtliche fremdsprachlichen Produktionen anhand der Niveaus des GER (Europarat 2001, 2020) eingestuft (vgl. Kap. 2.4). Auf diesen Grundlagen wurden einerseits das deutsche Teilkorpus vertieft analysiert (vgl. Kap. 3.1 und 3.2) und andererseits, darauf basierend, korpusbezogene Lehr-Lern-Aktivitäten für die Sekundarstufe I entwickelt (vgl. Kap. 3.3).

2 Korpus

Als Korpus gilt eine strukturierte, zweckorientierte Sammlung digitalisierter schriftlicher oder mündlicher Texte. Neben den Texten (Primärdaten) selbst bestehen Korpora aus linguistischen Beschreibungen (Annotationen) der Texte und Metadaten zur Charakterisierung der Aufgaben und der Lernenden (Lemnitzer & Zinsmeister, 2015).

Im SWIKO-Korpus stehen die Lernerproduktionen in drei Sprachen (Deutsch, Französisch, Englisch) im Zentrum. Diese basieren auf acht systematisch variierten Aufgaben (*tasks*), welche unter verschiedenen Produktionsbedingungen bearbeitet wurden (vgl. Kap. 2.3). Diese Erhebungsparameter, ebenso wie Merkmale der Lernenden (u. a. Alter, Geschlecht, Sprachkenntnisse), werden in den Metadaten festgehalten.

Abb. 1 veranschaulicht die Hauptkomponenten des SWIKO-Korpus und den Arbeitsprozess im Projekt. Die auf Basis der acht Aufgaben entstandenen Produktionen wurden nach korpuslinguistischen Methoden aufbereitet und analysiert. Nach der manuellen Transkription wurden sprachliche Informationen automatisch annotiert, darunter das Lemma[3] (z. B. gehören die Token *geh, gehst* und *ging* alle zum Lemma *gehen*) und die Wortart (Verb, Nomen usw.). Mit solchen Informationen lässt sich u. a. analysieren, wie lang, vielfältig oder dicht die Produktion ist. Zudem wurden in allen schriftlichen deutschen Produktionen orthographische und grammatische Fehler annotiert, sodass Fehlertypen erfasst und Strukturen identifiziert werden können, die für Lernende besonders herausfordernd sind.

Da die Produktionen von angehenden Fremdsprachenlehrpersonen nach den Niveaus des GER (Europarat, 2001, 2020) bewertet wurden (Ratings), lässt sich schliesslich auch untersuchen, welche Aufgabenmerkmale und welche sprachlichen Merkmale der Lernertexte mit den Ratings korrelieren.

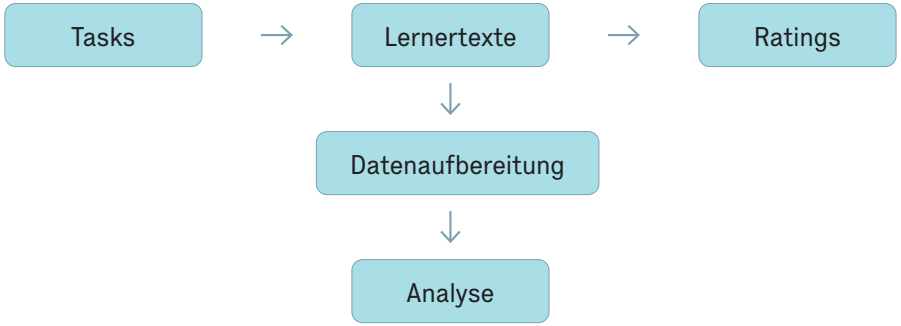


Abbildung 1. *Komponenten und Workflow des SWIKO/WETLAND Projekts*

3 Ein Token (oder auch: laufende Wortform) bezeichnet dabei eine lexikalische Einheit oder einzelne Wortform, die im Text vorkommt. Meistens lassen sich diese auf ein Lemma, d. h. eine kanonische Form oder Grundform zurückführen, wie sie z. B. in einem Wörterbuch zu finden wäre (Hass-Zumkehr, 2002).

2.1 Umfang

Ende 2024 umfasste das Korpus Daten aus drei verschiedenen Lehrplan-Regionen, welche gemäss der Erhebungschronologie in vier Teilkorpora erfasst sind (Tabelle 1):

- SWIK017 (Romandie, Deutsch als erste und Englisch als zweite Fremdsprache)
- SWIK018 (Deutschschweiz, Französisch als erste und Englisch als zweite Fremdsprache)
- SWIK019 (Deutschschweiz, Deutsch und Englisch als Schulsprachen)
- SWIK022 (Romandie, Deutsch als erste und Englisch als zweite Fremdsprache).

Textsprache[4]	Deutsch		Französisch		Englisch		TOTAL
	FL	LoS	FL	LoS	FL	LoS	
Klasse	10 & 11	11 & 12	11 & 12	10 & 11	10-12	10	
Mündlich							
Originaltexte n[5]	49	72	57	64	140	28	410
Transkripte n	42	72	7	0	108	8	237
Token n[6]	2’174	13’530	177	0	15’118	3’028	34’027
Schriftlich							
Originaltexte n	566	355	396	426	770	103	2’616
Transkripte n	543	347	322	384	684	102	2’382
Token n	23’737	23’667	17’567	27’584	45’173	8’556	146’284

Tabelle 1. *Umfang SWIKO Korpus (Stand Dez. 2024).*

4 FL steht für *foreign language*. LoS steht für *language of schooling* oder Unterrichtssprache, was oftmals der oder einer der Erstsprache(n) (L1) der Lernenden entspricht (andere Sprachenkombinationen wurden in den Metadaten festgehalten).

5 Bei den Originaltexten werden auch leere oder unlesbare Dokumente sowie Texte ohne Bezug zur Aufgabe mitgezählt. Nicht alle erhobenen Texte wurden transkribiert (z. B. mündliche französische Schulsprache).

6 Die angegebene Anzahl Tokens bezieht sich auf bereits transkribierte Texte.

2.2 Zugang und Nutzung

Auf das Korpus kann über die SWIKOweb Plattform (https://ifm-swiko.unifr.ch) zugegriffen werden. Die Webseite bietet einerseits Informationen zum Projekt, einschliesslich detaillierter Transkriptions- und Annotationsrichtlinien. Andererseits ist das Lernerkorpus mit Metadaten und massgeschneiderten Visualisierungstools zugänglich. Die Daten können nach einer Vielzahl von Kriterien gefiltert werden, von den Sprachen oder dem Geschlecht des Autors/der Autorin über verschiedene Aufgabenmerkmale und Produktionsbedingungen der Texte bis hin zum GER-Rating. Dank der Mehrebenen-Annotation können die Daten auf verschiedenen Ebenen durchsucht und angezeigt werden, und diese Suchanfragen können mit spezifischen Abfragen auf anderen Ebenen kombiniert werden. Beispielsweise könnte man nach allen Belegen des Lemmas *kein* suchen, die einen Flexionsfehler enthalten und von einem Nomen gefolgt werden. Standardmässig werden die Suchergebnisse tabellarisch inklusive Konkordanzen angezeigt; alternativ können auch Häufigkeitsverteilungen generiert werden.

2.3 Datenerhebung

Um der Vielfalt der Kommunikation im Klassenzimmer Rechnung zu tragen, wurden acht systematisch variierte Aufgaben für die Datenerhebung entwickelt (orientiert am task-Konzept des Task-Based Language Teaching [TBLT]-Ansatzes, vgl. Ellis et al., 2020). So mussten sich die Lernenden zu alltäglichen und schulischen (unten auch „akademischen“) Themen auf Basis von offenen und restriktiven Aufgaben sowohl beschreibend als auch argumentierend äussern. Tabelle 2 gibt einen Überblick über die Aufgaben und zeigt die Variation nach Texttyp (deskriptiv vs. argumentativ), Thema (akademisch vs. persönlich) und Struktur (restriktiv vs. offen).

Kürzel	Beschreibung der Aufgabe	Texttyp		Thema		Struktur	
		des	*arg*	*akad*	*pers*	*res*	*off*
SWI01	Kurze persönliche Fragen beantworten	×			×	×	
SWI02	Grafik zu Schweizer Haustieren beschreiben	×		×		×	
SWI03	Liste versch. Ferienmöglichkeiten diskutieren		×		×	×	
SWI04	Liste der wichtigsten Erfindungen diskutieren		×	×		×	
SWI05	Freies Selbstportrait für Klassenaustausch	×			×		×
SWI06	Ein Thema kurz präsentieren (8 zur Auswahl)	×		×			×
SWI07	Späteren Schulbeginn/ -schluss diskutieren		×		×		×
SWI08	Sprachaustausch statt Fremdsprachenunterricht diskutieren		×	×	×		×

Tabelle 2. *Aufgabenvariation für die SWIKO Datenerhebung.*

Abbildungen 2a und 2b (unten) illustrieren die Variation von Aufgaben und Produktionsbedingungen anhand von zwei komplementären Beispielen. SWI04 (2a) ist als argumentative, akademische und restriktivere Aufgabe kategorisiert. Im Beispiel mussten Lernende mit Deutsch als Schulsprache einen Text in Französisch als Fremdsprache schreiben, und zwar auf Papier. SWI05 (2b) hingegen ist eine deskriptive, persönliche und offene Aufgabe, die von Lernenden mit Französisch als Schulsprache in Deutsch als Fremdsprache bearbeitet werden musste, und zwar mündlich und am Computer.

Abbildung 2a. *Aufgabe SWI04 (argumentativ, akademisch, restriktiv).*

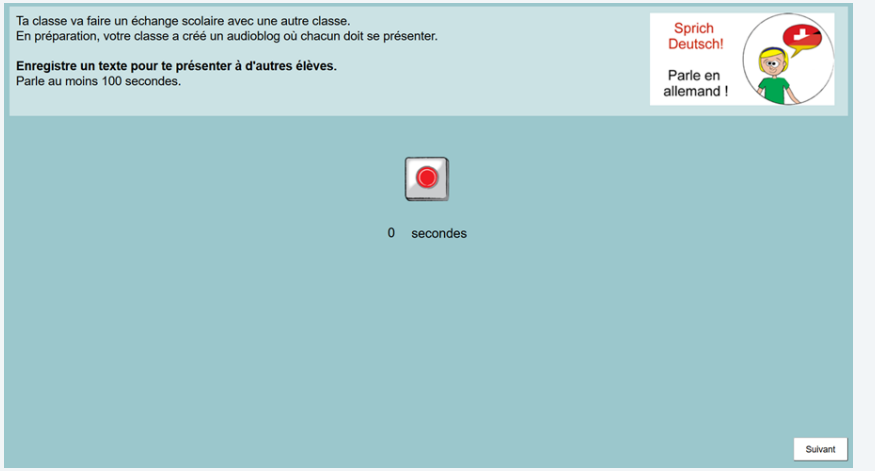


Abbildung 2b. *Aufgabe SWI05 (deskriptiv, persönlich, offen) - Screenshot.*

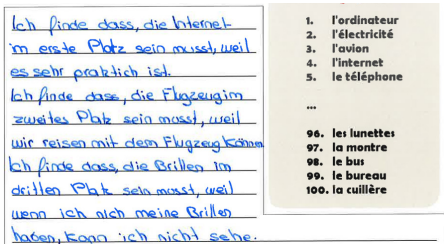
Zwischen 2017 und 2022 wurden Daten bei 340 Schüler/innen aus 17 Klassen der Sekundarstufe I (10.-12. Schuljahr nach der HarmoS-Einteilung) in der deutsch- und französischsprachigen Schweiz erhoben. Die Datenerhebung fand innerhalb von zwei Unterrichtsstunden (2×45min) während der regulären Schulzeit statt. Alle Lernenden bearbeiteten selbständig je vier Aufgaben in zwei Sprachen, z. B. je 4 Aufgaben in der ersten und in der zweiten Fremdsprache. Die meisten Lernenden bearbeiteten die Aufgaben schriftlich auf Papier oder am Computer. Pro Klasse lösten ausserdem vier zufällig ausgewählte Lernende dieselben Aufgaben in einem separaten Raum mündlich. Zwei von ihnen arbeiteten einzeln am Computer, während die anderen beiden an einer Interviewsituation mit einem Mitglied des Forschungsteams teilnahmen.

2.4 Datenaufbereitung

Die korpuslinguistische Datenaufbereitung bestand grob aus drei Schritten (in Abbildung 3 exemplarisch dargestellt für deutsche, schriftliche, papier-basierte Produktionen): (1) Manuelle Transkription des Originaltextes (0), (2) automatische Annotation sprachlicher Merkmale und (3) halbautomatische Fehlerannotation. Zudem wurden die Texte auf Basis des GER eingestuft (s. u.).

Im Einzelnen wurden die Originaltexte in einem ersten Schritt in zwei Versionen manuell transkribiert, wobei eine dem Original möglichst nahekommt (*Original Text*) und eine den Wortlaut in einer orthographisch korrekten Version wiedergibt (*Tagged Text*). Für mündliche Produktionen wurde der EXMARaLDA Partitur-Editor (Schmidt & Wörner, 2022) genutzt, für schriftliche Texte XMLmind (Shafie, 2021).

Danach wurden die Transkripte mit eigens entwickelten R-Skripten (R Core Team, 2022) in csv-Dateien umgewandelt. Mit dem Tool TreeTagger (Schmid, 2013) und dem koRpus-Paket (Michalke, 2019) wurden anschliessend automatisch Lemma und Wortart annotiert (POS-Annotation).

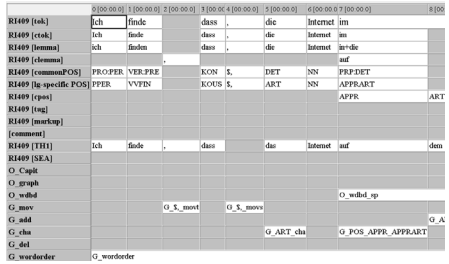


0) Originaltext

1) Transkript

	A	B	C	D	E	F
1	original	doc_id	token	common.F	de.POS.ta	lemma
2	Ich	SWI04_fD_	Ich	PRO:PER	PPER	ich
3	finde	SWI04_fD_	finde	VER:PRE	VVFIN	finden
4	dass	SWI04_fD_	dass	KON	KOUS	dass
5	,	SWI04_fD_	,	$,	$,	,
6	die	SWI04_fD_	die	DET	ART	die
7	Internet	SWI04_fD_	Internet	NN	NN	Internet
8	im	SWI04_fD_	im	PRP:DET	APPRART	in+die
9	erste	SWI04_fD_	erste	ADJ	ADJA	erst
10	Platz	SWI04_fD_	Platz	NN	NN	Platz
11	sein	SWI04_fD_	sein	VER:INF	VAINF	sein

2) POS-Annotation



3) Fehlerannotation

Abbildung 3. *Datenaufbereitung im SWIKO Projekt.*

Schliesslich wurden die Dateien in EXMARaLDA konvertiert und sämtliche schriftlichen deutschen Produktionen durch eine halbautomatische Fehlerannotation angereichert. Dazu wurde zuerst eine minimale Zielhypothese (TH1) formuliert, d. h. eine möglichst originalgetreue, aber orthographisch und grammatisch korrekte Version des Lernertextes (Lüdeling & Hirschmann, 2015). Im nächsten Schritt wurde die Differenz zwischen Originaltext und Zielhypothese zur automatischen Fehlerannotation genutzt. Annotiert wurden auf Basis von Tagsets orthographische (u. a. Gross-/Kleinschreibung), syntaktische (Satzbau) und grammatische (u. a. Flexion) Fehlschreibungen. Zuletzt wurde die automatische Annotation manuell überprüft und, wo nötig, korrigiert.

In einem letzten Schritt wurden zwischen 2020-2022 alle 1550 schriftlichen fremdsprachlichen Texte (DaF, FLE, und EFL) von 47 geschulten Rater/innen anhand der Niveaus des GER bewertet. Für das analytische Rating wurde ein validiertes Raster, basierend auf Deskriptoren aus *lingualevel* (Lenz & Studer, 2008) und dem Begleitband zum GER (Europarat, 2020), genutzt. Vier sprachliche Kriterien wurden berücksichtigt: *Wortschatz* legt den Fokus auf die Breite und Tiefe der verwendeten Wörter, *Grammatik* auf Merkmale wie Konjugation, *Rechtschreibung* auf orthographische Genauigkeit und *Text* auf Kohäsion. Die Bewertungen wurden anschliessend mit Mehrfacetten-Rasch-Analysen in Facets (Eckes, 2015; Linacre, 2022) ausgewertet, um für jeden Text einen *fair score* zu berechnen. *Fair scores* wurden sowohl für jedes Bewertungskriterium einzeln als auch über alle vier Kriterien hinweg berechnet. Damit steht für jeden Lernertext ein sprachliches GER-Profil zur Verfügung. Einzelheiten zu den Ratings können über https://ifm-swiko.unifr.ch unter Datenverarbeitung / Rating eingesehen werden.

3 Ausgewählte Resultate

3.1 Effekte von Produktionsbedingungen

Nach den ersten Datenerhebungen wurde in 1452 schriftlichen Texten der Einfluss des Mediums auf die Textlänge und Vielfalt des Wortschatzes untersucht (Karges et al., 2020). Sowohl bei Deutsch- als auch Französischsprachigen waren in der Schulsprache mündliche und schriftliche Texte gleich lang. Jedoch schrieben französischsprachige Lernende in Deutsch als Fremdsprache kürzere Texte am Computer als auf Papier, während deutschsprachige Lernende in Englisch als Fremdsprache längere Texte am Computer als auf Papier verfassten. Texte in der Schulsprache waren lexikalisch vielfältiger als in der Fremdsprache, während sich zwischen den zwei Fremdsprachen diesbezüglich kaum Unterschiede zeigten.

Verglichen wurden auch sprachliche Merkmale zwischen 110 mündlichen und 505 schriftlichen Produktionen in Deutsch als Schul- und Fremdsprache (Karges et al., 2022). Wie zu erwarten waren Produktionen in der Schulsprache im Schnitt länger und lexikalisch vielfältiger als fremdsprachliche Texte. Jedoch sagten die schulsprachlichen Teilnehmenden erheblich mehr als ihre Mitschüler/innen schrieben, während diese Unterschiede bei den DaF-Lernenden marginal waren. In mündlichen Produktionen griffen fremdsprachliche Lernende, mutmasslich bei Wortfindungsschwierigkeiten, (noch) häufiger auf andere Sprachen zurück als in schriftlichen Produktionen, wobei der französische Anteil in DaF-Texten in beiden Modalitäten doppelt so hoch war wie der englische.

3.2 Zusammenspiel von Aufgabe, sprachlichen Merkmalen der Produktionen und Bewertungen am Beispiel schriftlicher DaF-Texte

Auf der Basis von 544 schriftlichen Texten in Deutsch als Fremdsprache wurden Zusammenhänge zwischen den Aufgaben, den sprachlichen Merkmalen der Produktionen und den Bewertungen untersucht (Hicks, 2023; Hicks & Studer, 2024; Studer & Hicks, 2022). Bei den Aufgaben wurden die drei systematisch variierten Aufgabenmerkmale Texttyp (deskriptiv vs. argumentativ), Thema (persönlich vs. akademisch) und Struktur (offen vs. restriktiv) berücksichtigt. Als sprachliche Indikatoren wurden auf Basis des CAF-Frameworks (Housen et al., 2012; Michel, 2017) lexikalische und syntaktische Komplexität (*complexity*), Korrektheit (*accuracy*) und Textlänge (*fluency*) berechnet.

Alle drei Aufgabeneigenschaften beeinflussten sowohl die Länge als auch die lexikalische Dichte (*density*) und Elaboriertheit (*sophistication*) der Lernertexte, nicht aber die lexikalische Vielfalt (*diversity)*. Zudem prägte der Texttyp die syntaktische Komplexität,

während sich die Vertrautheit mit dem Thema auf die Korrektheit auswirkte. Zwei anonymisierte Beispiele derselben Person vermögen diese Zusammenhänge zu illustrieren (Tabelle 3): Im ersten stellte sich die Person vor (SWI05, links), im zweiten äusserte sie ihre Meinung zu einer Liste der wichtigsten Erfindungen (SWI04, rechts). Das Selbstportrait besteht aus einer Reihe einfacher, kurzer Hauptsätze, die viele Substantive (verhältnismässig seltenere Wörter) umfassen. Wenn es hingegen um akademische Argumentationen geht, wird die Sprache syntaktisch komplexer: die Autorin verwendet häufiger Nebensätze und schreibt längere Sätze, die mehr Adjektive und Funktionswörter enthalten. Diese Herausforderung führte aber auch zu mehr Fehlern.

SWI05: Selbstportrait (anonymisiert) deskriptiv, persönlich, offen	**SWI04: Erfindungen** argumentativ, akademisch, restriktiv
Hallo! Ich heisse Sandra und ich bin 15 Jahre alt. Ich liebe die Natur, Sport und ich lese gern aber ich spiele nicht gern Fussball und Bascket. "Shadow hunters" ist mein Lieblingserie und "La passe Miraire" ist mein Lieblingserie von Bücher. Ich habe zwei katzen, Sie heissen Simba und Luna. Mein Schwester heisst Laura und sie ist 18 aber ich habe keine Brüder. Ich liebe Ski fahren und Rad fahren aber mein Lieblingssport ist Klettern. Ich denke dass, ich Freundlisch, Neugierig, Schüchtern und Hilfsbereit bin. Bis bald!	*Für mich die Electricite ist im ersten platz weil ohne electricite es keine Lampe, keine Computer, keine Smartphone mehr gibt. Ich denke brille ist wichtiger als die sechs-und-neunzehnsten platz weil, für personnen wie kann nicht gut sehen ist ein sehr wichtiger punkt. Der Flieger ist für mich in die richtige platz und der bus auch aber der Stieft muss nicht in die liste bin weil es ist nicht ein sehr grossen invention.*

Tabelle 3: *Zwei Produktionen derselben Person Ri513 (links Selbstportrait, rechts Erfindungen).*

Die Art der Aufgabe hatte auch einen Einfluss auf die Bewertung: Texte, die auf persönlichen und offenen Aufgaben basierten, wurden höher eingestuft als Texte, die auf akademischen und restriktiven Aufgaben basierten. Während beispielsweise 75 % der Selbstportraits die Bewertung von A2.1 (der gemäss HarmoS zu erreichende Standard) oder höher erhielten, wurde nur gut ein Drittel der Texte zur Erfindungsliste diesen Stufen zugeordnet.

Schliesslich wurde auch der Zusammenhang zwischen der Bewertung und den sprachlichen Merkmalen untersucht. Längere Texte mit vielfältigem Wortschatz wurden dabei deutlich höher eingestuft; Texte mit vielen orthographischen und grammatischen Fehlern sowie nicht-zielsprachlichen Wörtern dagegen deutlich schlechter. Syntaktische Aspekte waren weniger relevant.

3.3 Didaktische Nutzung

Erkenntnisse wie in 3.2 berichtet können eine anschauliche Datenbasis für die Lehrerausbildung bieten. Die nah-authentischen Textbeispiele ermöglichen ein realistisches und differenziertes Verständnis der Leistungen der Schüler/innen. Zugleich schärfen sie das Bewusstsein für den grossen Einfluss der Aufgabe. Wie Tabelle 3 zeigt, werden verschiedene Aufgaben dem Leistungsvermögen der Lernenden unterschiedlich gerecht. Während Aufgaben wie das Selbstportrait die aktuellen Sprachkompetenzen der Lernenden gut abrufen, können Problemstellungen wie die Diskussion von Erfindungen ein Fenster zur nächsten Entwicklungszone öffnen.

SWIKO kann aber auch direkt im Klassenzimmer der Sekundarstufe I genutzt werden – zum Beispiel für das Erstellen von didaktischem Material zur Negation für den DaF-Unterricht (vgl. Abb. 4; Einzelheiten in Hicks & Studer, 2024). Mit SWIKOweb (vgl. Kap. 2.2) generierte Konkordanzen können in einer Einführungsphase genutzt werden, um Regeln zum Gebrauch von „*nicht*“ und „*kein*“ von den Lernenden ableiten zu lassen. In einem Mindmap können anschliessend typische Kombinationen gesammelt werden, sei es individuell, in Gruppen oder im Klassenverband. Auch Lückenübungen bieten sich an.

Abbildung 4. *Arbeitsblätter zur Negation für den DaF-Unterricht, basierend auf Konkordanzen aus dem SWIKO-Korpus.*

4 Fazit

Das Schweizer Lernerkorpus SWIKO ist eine umfangreiche Datenbank mit mündlichen und schriftlichen Produktionen von Lernenden in den schulischen Fremd- und den Schulsprachen auf der Sekundarstufe I. Die aufgabenbasierten Produktionen wurden mit korpuslinguistischen Methoden aufbereitet und, bezogen auf den GER, verlässlich eingestuft. Ausführliche Informationen zum entstandenen Korpus und zur Vorgehensweise im Projekt sind in einem gesonderten Korpusbericht (Hicks, Studer & Karges, i.V.) verfügbar.

In der Landschaft der Lernerkorpora zeichnet sich das mehrsprachige SWIKO Korpus besonders dadurch aus, dass es den öffentlichen Schulkontext und Lernende mit niedrigen Sprachniveaus adressiert, Aufgaben und Produktionsbedingungen systematisch variiert und sowohl die schulischen Fremdsprachen als auch die Schulsprache in den Blick nimmt.

Das Portal SWIKOweb (https://ifm-swiko.unifr.ch) ermöglicht Zugang zur Datenbank und ausserdem zu Werkzeugen, mit denen sich die über 2600 annotierten Lernertexte gruppieren, analysieren und visualisieren lassen. So kann das Korpus für wissenschaftliche und für didaktische Zwecke genutzt werden. Weiter erforscht werden kann etwa, welche Kombinationen von sprachlichen Merkmalen der Zielsprache auf welchen GER-Niveaus auftauchen oder bereits beherrscht werden. Didaktisch können u. a. weitere datenbasierte Übungen in Analogie zur Negation (Abb. 4) konstruiert werden, die das Sprachenlernen in herausfordernden Bereichen unterstützen.

Corpus suisse des apprenant·e·s SWIKO

Rapport de recherche

Nina Selina Hicks, Thomas Studer

1 Introduction

Le corpus suisse des apprenant·e·s SWIKO a été élaboré dans le cadre du projet de recherche SWIKO (2016-2019) et du projet subséquent WETLAND – Développement et applications (2021-2024) au Centre scientifique de compétence sur le plurilinguisme. L'objectif principal était de documenter le langage des apprenant·e·s à la fin de la scolarité à l'aide de concepts et de méthodes issus de la linguistique de corpus, de le rendre consultable et de l'analyser de manière approfondie sur la base d'exemples. L'accent a été mis sur les langues nationales allemand et français ainsi que l'anglais comme langues étrangères. Une attention particulière a été accordée à l'articulation entre les tâches, les caractéristiques linguistiques des productions écrites et orales et leur évaluation.

L'idée du projet de recherche est née du changement intervenu dans l'enseignement des langues étrangères qui, au tournant du siècle, adopte une approche communicative. Ceci se reflète aujourd'hui dans les plans d'études régionaux (CIIP, 2023 ; D-EDK, 2016 ; Passepartout, 2015) et le matériel pédagogique (p. ex. New World, Arnet-Clark et al., 2013) sous l'influence déterminante du Cadre européen commun de référence pour les langues (CECR: Conseil de l'Europe, 2001)[1]. Depuis lors, l'enseignement des langues vise à « permettre aux apprenant·e·s d'agir dans des situations de la vie quotidienne, de s'exprimer dans leurs langues et de réaliser avec succès des tâches de nature diverse » (Conseil de l'Europe, 2020, p. 33 ; voir aussi 2001, chap. 7). Par conséquent, l'importance accordée au vocabulaire et à la grammaire a évolué : les compétences linguistiques ne sont plus au centre de l'enseignement, mais constituent des ressources au service de la réalisation d'objectifs communicatifs (Ende et al., 2013).

Parallèlement, l'organisation de l'enseignement des langues étrangères dans les écoles publiques suisses a été harmonisée dans le cadre du projet HarmoS (concordat HarmoS Langues étrangères, 2009). Ainsi, dans la plupart des cantons, les élèves apprennent deux langues étrangères à partir de la 5e et de la 7e année[2], à savoir une langue nationale ainsi que l'anglais. À la fin de la scolarité obligatoire – en 11e année, vers l'âge de 15 ans – les élèves sont censés atteindre le niveau A2.2 du CECR globalement dans les deux langues étrangères et le niveau A2.1 à l'écrit. Alors qu'un suivi national (consortium HarmoS, 2019) et des projets connexes (entre autres Peyer et al., 2016) ont étudié dans quelle mesure ces standards étaient atteints, on en sait peu sur les compétences *linguistiques* spécifiques.

Dans ce contexte, la question fondamentale de SWIKO porte sur les compétences linguistiques des apprenant·e·s - en particulier en matière de vocabulaire et de grammaire – atteint à la fin de la scolarité obligatoire dans le cadre de l'approche actuelle de

1 Ci-après dénommé « CECR ».

2 Depuis HarmoS, la scolarité obligatoire dure au total 11 ans : dès l'âge de 4 ans révolus, les enfants suivent deux ans d'école enfantine (CITE 0), six ans d'école primaire (CITE 1) et trois ans d'école secondaire (CITE 2). Par conséquent, les participant·e·s en 10e année ont environ 14 - 15 ans et sont en deuxième année du secondaire I.

l'enseignement des langues. SWIKO souhaite ainsi, par son approche empirique, participer à améliorer notre compréhension de la manière dont les structures linguistiques sont acquises au travers de l'approche communicative et à formuler des attentes réalistes quant aux performances que les élèves peuvent réaliser s'agissant des aspects formels de la langue.

Au cours de la première période de recherche (2016-2020), une banque de données a été constituée à partir de productions d'apprenant·e·s, orales et écrites, recueillies sur la base de tâches. Dans le cadre de la récolte des données, les apprenant·e·s du secondaire I ont traité au total huit tâches variées de manière systématique dans différentes conditions de production (cf. chap. 2.3). Les données ont ensuite été transcrites, puis traitées à l'aide de méthodes appartenant à la linguistique de corpus (cf. chap. 2.4).

Au cours de la deuxième période de recherche (2021-2024), le sous-corpus allemand a notamment été élargi et complété par une annotation semi-automatique des erreurs (cf. chap. 2.4). De plus, toutes les productions en langue étrangère ont été classées selon les niveaux du CECR (Conseil de l'Europe, 2001, 2020) (cf. chap. 2.4). A partir de cela, le sous-corpus allemand a été soumis à une analyse approfondie (cf. chap. 3.1 et 3.2) et des tâches basées corpus ont été développées pour l'enseignement et l'apprentissage au niveau secondaire I (voir chap. 3.3).

2 Corpus

Un corpus est une collection structurée et ciblée de textes numériques, écrits ou oraux. Outre les textes eux-mêmes (données primaires), les corpus comprennent des descriptions linguistiques (annotations) des textes ainsi que des métadonnées permettant de caractériser les tâches et les apprenant·e·s (Lemnitzer & Zinsmeister, 2015).

Le corpus SWIKO se concentre sur les productions d'apprenant·e·s dans trois langues (allemand, français, anglais). Ces productions sont basées sur huit tâches (*tasks*) variées de manière systématique, réalisées dans différentes conditions de production (cf. chap. 2.3). Les paramètres de la récolte des données ainsi que les caractéristiques des apprenant·e·s (notamment l'âge, le sexe, les compétences linguistiques) sont consignés dans les métadonnées.

La figure 1 illustre les principales composantes du corpus SWIKO et le processus de travail dans le projet. Les productions basées sur les huit tâches ont été traitées et analysées selon des méthodes de la linguistique de corpus. Après la transcription manuelle, des informations linguistiques ont été automatiquement annotées, notamment le lemme[3] (p. ex., les tokens *mange*, *manges* et *mangeait* appartiennent tous au lemme *manger*) et la catégorie grammaticale (verbe, nom, etc.). De telles informations permettent notamment d'analyser la longueur, la diversité ou la densité de la production. De plus, l'ensemble des productions écrites en allemand a été annoté de manière à signaler les erreurs orthographiques et grammaticales afin de recenser les types d'erreurs et d'identifier les structures particulièrement exigeantes pour les apprenant·e·s.

Etant donné que les productions ont été évaluées par de futur·e·s enseignant·e·s de langues étrangères selon les niveaux définis dans le CECR (Conseil de l'Europe, 2001, 2020), il est finalement possible d'analyser quelles tâches et quelles caractéristiques linguistiques des productions des apprenant·e·s sont corrélées aux évaluations.

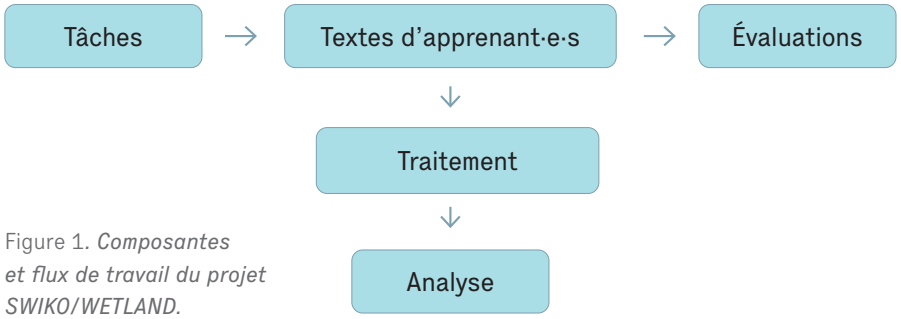


Figure 1. *Composantes et flux de travail du projet SWIKO/WETLAND.*

3 Un token (ou forme courante d'un mot) désigne une unité lexicale ou une forme individuelle de mot apparaissant dans le texte. La plupart du temps, les tokens peuvent être associés à un lemme, c'est-à-dire une forme canonique ou de base telle qu'on la trouverait, par exemple, dans un dictionnaire (Hass-Zumkehr, 2002).

2.1 Portée

Fin 2024, le corpus comprenait des données collectées dans trois régions aux plans d'études différents. Ces données ont été enregistrées dans quatre sous-corpus, suivant la chronologie des récoltes (tableau 1) :

- SWIKO17 (Suisse romande, allemand 1re langue étrangère, anglais 2e langue étrangère)
- SWIKO18 (Suisse alémanique, français 1re langue étrangère, anglais 2e langue étrangère)
- SWIKO19 (Suisse alémanique, allemand et anglais comme langue de scolarisation)
- SWIKO22 (Suisse romande, allemand 1re langue étrangère, anglais 2e langue étrangère)

Langue du texte[4]	Allemand		Français		Anglais		TOTAL
	FL	LoS	FL	LoS	FL	LoS	
Année	10 & 11	11 & 12	11 & 12	10 & 11	10-12	10	
Productions orales							
Textes originaux n[5]	49	72	57	64	140	28	410
Transcriptions n	42	72	7	0	108	8	237
Tokens n[6]	2'174	13'530	177	0	15'118	3 028	34'027
Productions écrites							
Textes originaux n	566	355	396	426	770	103	2'616
Transcriptions n	543	347	322	384	684	102	2'382
Tokens n	23'737	23'667	17'567	27'584	45'173	8'556	146'284

Tableau 1. *Portée du corpus SWIKO (état déc. 2024).*

4 FL signifie *foreign language* (langue étrangère). LoS signifie *language of schooling*, ou langue de scolarisation, et correspond souvent à la ou une des langues premières (L1) des apprenant·e·s (d'autres combinaisons linguistiques ont été consignées dans les métadonnées).

5 Les textes originaux incluent également les documents vides ou illisibles ainsi que les textes sans lien avec la tâche. Les textes recueillis n'ont pas tous été transcrits (par exemple, les productions orales en français langue de scolarisation).

6 Le nombre de tokens indiqué se réfère aux textes déjà transcrits.

2.2 Accès et utilisation

Le corpus est accessible via la plateforme SWIKOweb (https://ifm-swiko.unifr.ch). D'une part, le site web fournit des informations sur le projet, y compris des règles détaillées pour la transcription et l'annotation. D'autre part, il permet accès au corpus des apprenant·e·s, accompagné de métadonnées et d'outils de visualisation adaptés. Les données peuvent être filtrées selon une grande variété de critères, allant des langages ou du sexe de l'auteur·e à différentes caractéristiques des tâches et conditions de production des textes, jusqu'au niveau CECR attribué. Grâce à l'annotation multi-niveaux, les données peuvent être recherchées et affichées à différents niveaux, et ces requêtes peuvent être combinées à des recherches spécifiques à d'autres niveaux. Par exemple, on pourrait rechercher toutes les occurrences du lemme *kein* contenant une erreur de flexion et suivies d'un nom. Par défaut, les résultats de recherche sont affichés sous forme de tableau incluant les concordances ; il est également possible de générer la fréquence d'apparition.

2.3 Récolte des données

Afin de saisir la diversité des modes de communication en classe, huit tâches variées de manière systématique ont été élaborées en vue de récolter les données sur la base du concept de tâches tel que défini dans l'approche Task-Based Language Teaching (TBLT), (voir Ellis et al., 2020). Les apprenant·e·s devaient ainsi s'exprimer sur des thèmes quotidiens et scolaires, à partir de tâches ouvertes et restrictives, en mobilisant à la fois des modes d'expression descriptifs et argumentatifs. Le tableau 2 donne un aperçu des tâches et montre les variations selon le type de texte (descriptif ou argumentatif), le thème (académique ou personnel) et la structure (restrictive ou ouverte).

Sigle	Description de la tâche	Type de texte		Thème		Structure	
		des	*arg*	*acad*	*pers*	*res*	*ouv*
SWI01	Répondre à de courtes questions personnelles	×			×	×	
SWI02	Commenter des graphiques concernant les animaux domestiques en Suisse	×		×		×	
SWI03	Discuter d'une liste d'idées de vacances		×		×	×	
SWI04	Discuter d'une liste des inventions importantes		×	×		×	
SWI05	Créer son autoportrait pour une présentation en classe	×			×		×
SWI06	Présenter un sujet (parmi huit options proposées)	×		×			×
SWI07	Discuter de l'opportunité de décaler les horaires scolaires		×		×		×
SWI08	Discuter de l'opportunité de remplacer les cours de langues étrangères par un échange linguistique à l'étranger		×	×	×		×

Tableau 2. *Variation des tâches pour la collecte de données SWIKO.*

Les figures 2a et 2b (ci-dessous) illustrent la variation des tâches et les conditions de production à l'aide de deux exemples complémentaires. SWI04 (2a) est catégorisée comme une tâche argumentative, académique et plus restrictive. Dans cet exemple, les apprenant·e·s de langue de scolarisation allemande doivent rédiger, sur papier, un texte en français langue étrangère. SWI05 (2b), en revanche, est une tâche descriptive, personnelle et ouverte, que les apprenant·e·s de langue de scolarisation française doivent réaliser en allemand langue étrangère, à l'oral et sur ordinateur.

Figure 2a. *Tâche SWI04 (argumentative, académique, restrictive).*

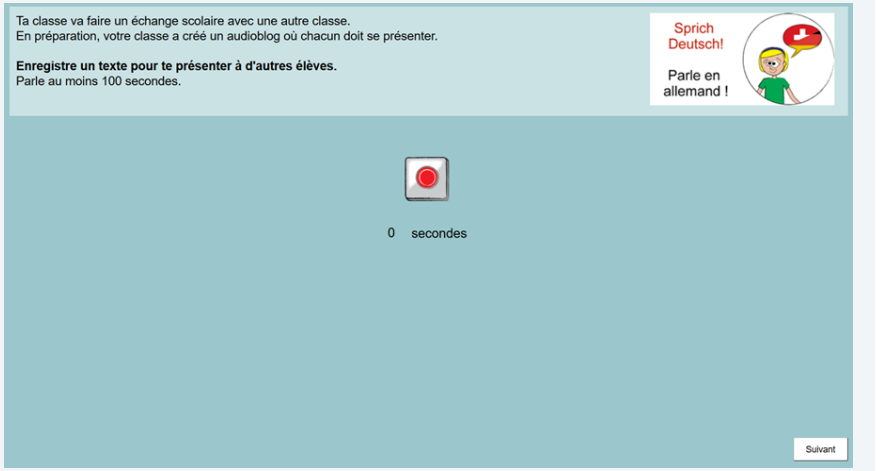


Figure 2b. *Tâche SWI05 (descriptive, personnelle, ouverte) - capture d'écran.*

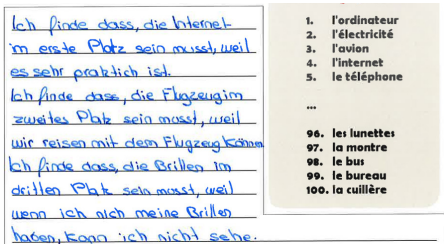
Entre 2017 et 2022, des données ont été recueillies auprès de 340 élèves issu·e·s de 17 classes du secondaire I (10e à 12e année selon la classification HarmoS) en Suisse alémanique et romande. La récolte des données a eu lieu pendant deux périodes de cours (2 × 45 min) pendant les heures de cours normales. Tous les apprenant·e·s ont traité de manière autonome quatre tâches dans deux des trois langues à leur disposition, par exemple quatre tâches dans la première et quatre tâches dans la deuxième langue étrangère. La plupart des apprenant·e·s ont réalisé les tâches par écrit, sur papier ou sur ordinateur. De plus, dans chaque classe, quatre apprenant·e·s choisi·e·s au hasard ont réalisé les mêmes tâches à l'oral dans une salle séparée. Deux d'entre eux·elles ont travaillé individuellement sur ordinateur, tandis que les deux autres ont participé à un entretien avec une personne de l'équipe de recherche.

2.4 Traitement des données

Le traitement des données linguistiques du corpus s'est déroulé en trois étapes (illustrées à titre d'exemple dans la figure 3 pour des productions écrites sur papier) en allemand : (1) transcription manuelle du texte original (0), (2) annotation automatique des aspects linguistiques et (3) annotation semi-automatique des erreurs. De plus, les textes ont été classés selon le CECR (voir ci-dessous).

Plus précisément, les textes originaux ont été transcrits manuellement en deux versions dans un premier temps : l'une proche de l'original (*Original Text*) et l'autre reproduisant le libellé avec une orthographe correcte (*Tagged Text*). L'on a utilisé EXMARaLDA Partitur Editor (Schmidt & Wörner, 2022) pour les productions orales et XMLmind (Shafie, 2021) pour les productions écrites.

Ensuite, les transcriptions ont été converties en fichiers CSV à l'aide de scripts R développés spécialement à cet effet (R Core Team, 2022). Les lemmes et les catégories grammaticales ont été annotés automatiquement (annotation morpho-syntaxique) avec TreeTagger (Schmid, 2013) et le package koRpus (Michalke, 2019),.

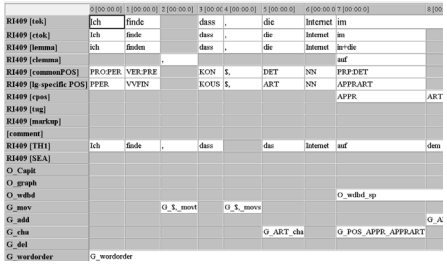


0) Texte original

1) Transcription

	A	B	C	D	E	F
1	original	doc_id	token	common.F	de.POS.ta	lemma
2	Ich	SWI04_fD_	Ich	PRO:PER	PPER	ich
3	finde	SWI04_fD_	finde	VER:PRE	VVFIN	finden
4	dass	SWI04_fD_	dass	KON	KOUS	dass
5	,	SWI04_fD_	,	$,	$,	,
6	die	SWI04_fD_	die	DET	ART	die
7	Internet	SWI04_fD_	Internet	NN	NN	Internet
8	im	SWI04_fD_	im	PRP:DET	APPRART	in+die
9	erste	SWI04_fD_	erste	ADJ	ADJA	erst
10	Platz	SWI04_fD_	Platz	NN	NN	Platz
11	sein	SWI04_fD_	sein	VER:INF	VAINF	sein

2) Annotation morpho-syntaxique



3) Annotation d'erreurs

Figure 3. *Traitement des données dans le cadre du projet SWIKO.*

Enfin, les fichiers ont été convertis dans EXMARaLDA et toutes les productions écrites en allemand ont été enrichies d'une annotation semi-automatique des erreurs. Pour ce faire, on a d'abord formulé une hypothèse cible minimale (TH1), c'est-à-dire une version du texte de l'apprenant·e aussi fidèle que possible à l'original, mais orthographiquement et grammaticalement correcte (Lüdeling & Hirschmann, 2015). À l'étape suivante, la différence entre le texte original, le libellé correctement orthographié et l'hypothèse cible a été utilisée pour l'annotation automatique des erreurs. Les fautes d'orthographe (notamment les majuscules/minuscules), de syntaxe (construction de phrases) et de grammaire (notamment la flexion) ont été annotées sur la base d'ensembles de balises. Enfin, l'annotation automatique a été vérifiée manuellement et corrigée lorsque cela était nécessaire.

Lors de la dernière étape, qui s'est déroulée entre 2020 et 2022, 47 évaluateurs et évaluatrices formé·e·s ont évalué les 1550 productions écrites en langue étrangère (DaF, FLE et EFL) sur la base des niveaux du CECR. L'évaluation analytique a été réalisée à l'aide d'une grille validée, basée sur les descripteurs *lingualevel* (Lenz & Studer, 2008) et le volume complémentaire du CECR (Conseil de l'Europe, 2020). Quatre critères linguistiques ont été pris en compte : le *vocabulaire* met l'accent sur l'étendue et la profondeur des mots utilisés, la *grammaire* sur des caractéristiques telles que la conjugaison, l'*orthographe* sur la précision orthographique et le *texte* sur la cohésion. Les évaluations ont ensuite été examinées au moyen d'analyses Rasch multi-facettes, effectuées avec le logiciel Facets (Eckes, 2015 ; Linacre, 2022), afin de calculer une note équitable (*faire score*) pour chaque texte. Les notes ont été calculées à la fois pour chaque critère d'évaluation pris individuellement et pour l'ensemble des quatre critères. Ainsi, un profil linguistique est disponible pour chaque texte d'apprenant·e. Les détails relatifs aux notations peuvent être consultés sur https://ifm-swiko.unifr.ch, sous la rubrique Traitement des données / Rating.

3 Résultats sélectionnés

3.1 Effets des conditions de production

Après les premières récoltes de données, l'équipe de recherche a étudié l'influence du support sur la longueur des textes et la diversité du vocabulaire au travers de 1452 textes écrits (Karges et al., 2020). Tant chez les germanophones que chez les francophones, les productions orales et écrites étaient de longueur égale dans la langue de scolarisation. Cependant, les apprenant·e·s francophones ont rédigé des textes plus courts à l'ordinateur que sur papier en allemand langue étrangère, tandis que les apprenant·e·s germanophones ont produit des textes plus longs à l'ordinateur que sur papier en anglais langue étrangère. Les textes rédigés dans la langue de scolarisation étaient lexicalement plus variés que dans la langue étrangère, alors que les différences à cet égard entre les deux langues étrangères étaient à peine perceptibles.

L'équipe a également comparé les caractéristiques linguistiques de 110 productions orales et de 505 productions écrites en allemand langue de scolarisation et allemand langue étrangère (Karges et al., 2022). Comme on pouvait s'y attendre, les productions dans la langue de scolarisation ont été en moyenne plus longues et lexicalement plus variées que les textes en langue étrangère. Dans leur langue de scolarisation cependant, les participant·e·s ont exprimé beaucoup plus de choses à l'oral que leurs camarades à l'écrit, tandis que ces différences étaient minimes chez les apprenant·e·s en allemand langue étrangère (DaF). Les apprenant·e·s en langue étrangère ont vraisemblablement eu (encore) plus souvent recours à d'autres langues dans les productions orales qu'à l'écrit lorsqu'ils et elles éprouvaient des difficultés à trouver leurs mots. Dans les deux modalités, la proportion de mots français dans les textes en allemand langue étrangère dans les textes en allemand langue étrangère était deux fois plus élevée que celle des mots anglais.

3.2 Liens entre tâche, caractéristiques linguistiques des productions et évaluations à partir de l'exemple de productions écrites en allemand langue étrangère

Sur la base de 544 productions écrites en allemand langue étrangère, les liens entre les tâches, les caractéristiques linguistiques des productions et les évaluations ont été étudiés (Hicks, 2023 ; Hicks & Studer, 2024 ; Studer & Hicks, 2022). Pour les tâches, les trois caractéristiques systématiquement variées ci-après ont été prises en compte : type de texte (descriptif vs argumentatif), thème (personnel vs académique) et structure (ouverte vs restrictive). Les indicateurs linguistiques, soit la complexité lexicale et syntaxique (*complexity*), l'exactitude (*accuracy*) et la longueur du texte (*fluency*), ont été calculés sur la base du cadre CAF (Housen et al., 2012 ; Michel, 2017).

Les trois caractéristiques des tâches ont influencé à la fois la longueur, la densité lexicale (*density*) et la sophistication (*sophistication*) des textes des apprenant·e·s, mais pas la diversité lexicale (*diversity*). De plus, le type de texte a influencé la complexité syntaxique, tandis que la familiarité avec le sujet a eu un impact sur l'exactitude. Deux exemples anonymisés de la même personne permettent d'illustrer ces liens (tableau 3) : dans le premier exemple, l'élève devait se présenter (SWI05, à gauche), dans le second exemple, elle devait donner son avis sur une liste des inventions les plus importantes (SWI04, à droite). L'autoportrait se compose d'une série de phrases principales simples et courtes, comprenant de nombreux noms (mots relativement plus rares). En revanche, lorsqu'il s'agit d'argumentations académiques, la langue se complexifie sur le plan syntaxique : l'auteure utilise plus fréquemment des subordonnées et rédige des phrases plus longues, contenant davantage d'adjectifs et de mots grammaticaux. Toutefois, ce défi a aussi conduit à un plus grand nombre d'erreurs.

SWI05 : autoportrait (anonymisé) descriptif, personnel, ouvert	SWI04 : inventions argumentatif, académique, restrictif
Hallo! Ich heisse Sandra und ich bin 15 Jahre alt. Ich liebe die Natur, Sport und ich lese gern aber ich spiele nicht gern Fussball und Bascket. «Shadow hunters» ist mein Lieblingserie und «La passe Miraire» ist mein Lieblingserie von Bücher. Ich habe zwei katzen, Sie heissen Simba und Luna. Mein Schwester heisst Laura und sie ist 18 aber ich habe keine Brüder. Ich liebe Ski fahren und Rad fahren aber mein Lieblingssport ist Klettern. Ich denke dass, ich Freundlisch, Neugierig, Schüchtern und Hilfsbereit bin. Bis bald!	*Für mich die Electricite ist im ersten platz weil ohne electricite es keine Lampe, keine Computer, keine Smartphone mehr gibt. Ich denke brille ist wichtiger als die sechs-und-neunzehnsten platz weil, für personnen wie kann nicht gut sehen ist ein sehr wichtiger punkt. Der Flieger ist für mich in die richtige platz und der bus auch aber der Stieft muss nicht in die liste bin weil es ist nicht ein sehr grossen invention.*

Tableau 3 : *deux productions réalisées par l'élève Ri513 (à gauche, autoportrait ; à droite, inventions).*

Le type de tâche a également influencé l'évaluation : les textes basés sur des tâches personnelles et ouvertes ont été mieux notés que ceux basés sur des tâches académiques et restrictives. Par exemple, tandis que 75 % des autoportraits ont été évalués au niveau A2.1 (le niveau standard à atteindre selon HarmoS) ou supérieur, un peu plus d'un tiers seulement des textes résultant de la tâche portant sur liste des inventions ont atteint ces niveaux.

Enfin, le lien entre l'évaluation et les caractéristiques linguistiques a également été examiné. Les textes plus longs incluant un vocabulaire varié ont été bien mieux notés tandis que les textes comportant de nombreuses fautes d'orthographe et de grammaire ainsi que des mots n'appartenant pas à la langue cible ont été nettement moins bien notés. Les aspects syntaxiques étaient moins pertinents.

3.3 Utilisation didactique

Les résultats présentés au point 3.2 peuvent constituer une base de données illustrative pour la formation des enseignant·e·s. Les exemples de textes presque authentiques permettent une compréhension plus réaliste et plus nuancée des performances des élèves. Parallèlement, ils sensibilisent au rôle déterminant de la tâche. Comme l'illustre le tableau 3, les différentes tâches font ressortir les capacités des apprenant·e·s de manière différente. Alors que des tâches telles que l'autoportrait permettent une bonne évaluation des compétences linguistiques existantes, des problématiques telles que la discussion sur les inventions peuvent faire apparaître de nouveaux domaines à développer.

SWIKO peut également être utilisé directement en classe au secondaire I, par exemple pour créer du matériel didactique traitant de la négation à utiliser dans le cadre de l'enseignement de l'allemand langue étrangère (cf. fig. 4 ; détails dans Hicks & Studer, 2024). Les concordances générées avec SWIKOweb (cf. chap. 2.2) peuvent être utilisées dans une phase d'introduction afin d'amener les apprenant·e·s à déduire eux·elles-mêmes les règles d'emploi de « *ne pas* » et « *aucun* ». Les combinaisons typiques peuvent ensuite être rassemblées dans une carte mentale, que ce soit individuellement, en groupes ou en classe. Les exercices à trous sont également intéressants.

Figure 4. *Fiches d'exercices sur la négation pour l'enseignement de l'allemand langue étrangère, basées sur des concordances issues du corpus SWIKO.*

4 Résumé

Le corpus suisse des apprenant·e·s SWIKO est une vaste banque de données contenant des productions orales et écrites dans les langues étrangères et les langues scolaires d'apprenant·e·s du niveau secondaire I. Les productions, basées sur des tâches, ont été traitées à l'aide de méthodes de la linguistique de corpus ainsi que classées de manière fiable selon le CECR pour les langues. Des informations détaillées sur le corpus créé et sur la méthodologie utilisée dans le cadre du projet sont disponibles dans un rapport distinct consacré au corpus (Hicks, Studer & Karges, à paraître).

Le corpus plurilingue SWIKO occupe une place unique dans le contexte plus large des corpus d'apprenant·e·s en ciblant le contexte scolaire public et des apprenant·e·s ayant un niveau de compétence linguistique limité, en variant systématiquement les tâches et les conditions de production, et en prenant en compte à la fois les langues étrangères apprises à l'école et la langue de scolarisation.

Le portail SWIKOweb (https://ifm-swiko.unifr.ch) offre un accès à la banque de données ainsi qu'à des outils permettant de regrouper, d'analyser et de visualiser plus de 2600 textes annotés annotés d'apprenant·e·s. Le corpus peut dès lors être exploité à des fins scientifiques et pédagogiques. Il serait également possible de poursuivre les recherches pour déterminer quelles combinaisons de caractéristiques linguistiques de la langue cible émergent ou sont maîtrisées selon les différents niveaux du CECR. Sur le plan didactique, d'autres exercices basés sur les données peuvent notamment être élaborés, à l'image des exercices sur la négation (fig. 4), afin de soutenir l'apprentissage des langues dans les domaines les plus exigeants.

The SWIKO Swiss learner corpus

Research report

Nina Selina Hicks, Thomas Studer

1 Introduction

The Swiss learner corpus SWIKO was developed at the Research Centre on Multilingualism in the context of the 2016–2019 SWIKO research project and as part of the follow-up project WETLAND – Weiterentwicklung und Anwendungen (Further development and applications, 2021–2024). The overarching aim was to document learner language at the end of compulsory schooling using concepts and methods from corpus linguistics, to make it available for further research, and to conduct detailed analyses of selected areas of linguistic competence. The focus was on national languages French and German as well as English as a foreign language. Particular attention was devoted to the interactions between tasks, the linguistic features of written and oral productions, and the ratings of these productions.

The starting point for the research project was informed by the shift towards a communicative approach in foreign language education at the turn of the century. Originating in the Common European Framework of Reference for Languages (CEFR: Council of Europe[1], 2001), the communicative approach is now reflected in both regional curricula (CIIP, 2023; D-EDK, 2016; Passepartout, 2015) and teaching materials (e.g. *New World,* Arnet-Clark et al., 2013). Language teaching has since been directed "towards enabling learners to act in real-life situations, expressing themselves and accomplishing tasks of different natures" (Council of Europe, 2020, p. 29; cf. 2001, ch. 7). This development also redefined the importance placed on vocabulary and grammar, which are no longer at the centre of language lessons; rather, they are viewed as a means to achieving communicative goals (Ende et al., 2013).

At the same time, the HarmoS project was introduced to standardise instructed foreign language education in Switzerland's public school system (Konsortium HarmoS Fremdsprachen, 2009). At present, students in most cantons learn two foreign languages (a national language and English) starting in Years 5 and 7.[2] At the end of compulsory schooling (Year 11, at the age of approximately 15), it is expected that students' proficiency in both foreign languages is a CEFR A2.2 overall, and CEFR A2.1 in writing. Although a national monitoring programme (Konsortium ÜGK, 2019) as well as related projects (e.g. Peyer et al., 2016) have investigated how well these communicative standards are achieved, little is known about specific *language* skills.

Against this backdrop, SWIKO investigates learners' linguistic competence – especially regarding vocabulary and grammar – at the end of mandatory schooling within the realms of a communicative approach to foreign language education. Through this line of inquiry, SWIKO aims to make an empirical contribution to better understanding the acquisition of

1 Hereinafter referred to as CEFR.

2 With the introduction of HarmoS, compulsory schooling is set to last a total of 11 years: starting at age 4, first 2 years of kindergarten (ISCED 0), 6 years of primary school (ISCED 1) and then 3 years of secondary school (ISCED 2). Consequently, students in Year 10 are generally 14 to 15 years old and in their second year of lower secondary school.

linguistic structures in the "new" foreign language lessons, and to support the formulation of realistic expectations regarding student achievement in formal aspects of language use. During the first research period (2016–2020), a database consisting of task-based oral and written learner texts was compiled. During the data collection phase, students in lower secondary schools completed a total of eight systematically varied tasks under different conditions (cf. Section 2.3). The data were then transcribed and subsequently prepared using corpus-linguistic methods (cf. Section 2.4).

During the second research period (2021–2024), the sub-corpus for German in particular was further enlarged and supplemented by semi-automatic error annotation (cf. Section 2.4). In addition, all foreign language productions were rated according to their CEFR level (Council of Europe 2001, 2020) (cf. Section 2.4). This work served as the basis for conducting detailed analyses of the German sub-corpus (cf. Sections 3.1 and 3.2), as well as for developing corpus-based teaching and learning activities for lower secondary schools (cf. Section 3.3).

2 Corpus

A corpus is a structured, purpose-oriented, electronic collection of written or oral texts. In addition to the texts themselves (primary data), corpora contain linguistic descriptions of the texts (annotations) and metadata to characterise tasks and learners (Lemnitzer & Zinsmeister, 2015).

The SWIKO corpus focuses on learner productions in three languages (German, French, English). These productions are based on eight systematically varied tasks, which were completed under different conditions (cf. Section 2.3). These survey parameters are recorded in the metadata, as are learner characteristics such as age, gender, language ability.

Figure 1 illustrates the main components of the SWIKO corpus and procedures used in the project. The learner productions resulting from the eight tasks were processed and analysed using corpus-linguistic methods. Following manual transcription, the linguistic information was automatically annotated; this information includes the lemma[3] (e.g. the tokens *geh, gehst* and *ging*, which belong to the lemma *gehen*) and the part of speech (e.g. verb, noun). This information can be used to analyse the length, diversity and density of a learner's production. In addition, orthographic and grammatical errors were annotated in all written productions in German to enable common error types to be documented, and structures that cause learners the most problems to be identified.

The productions were also rated by prospective foreign language teachers according to the levels defined in the CECR (Council of Europe, 2001, 2020). This makes it possible to analyse how tasks and linguistic features of the learner texts correlate with the ratings.

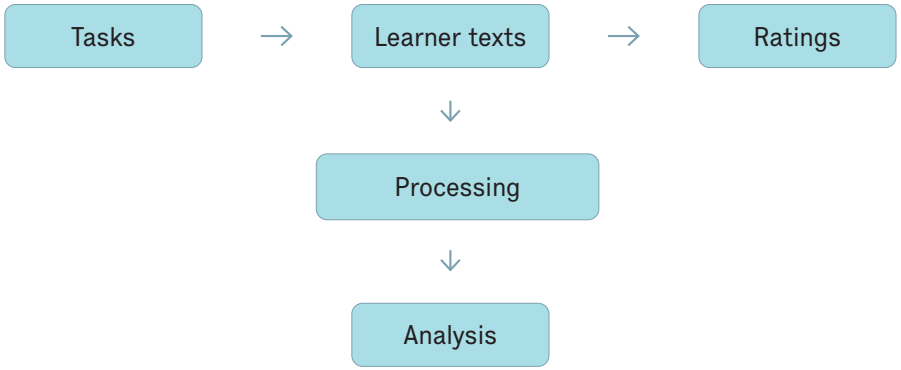


Figure 1: *Components and workflow of the SWIKO/WETLAND project.*

3 A token (also called a running word) refers to a single occurrence of a lexical unit or individual word form in a text. In most cases, tokens can be traced back to a lemma, i.e. a canonical or basic form of a word that would be found in a dictionary (Hass-Zumkehr, 2002).

2.1 Scope

At the end of 2024, the corpus comprised data from three different curricular regions, which are documented in four sub-corpora in the order the surveys were conducted (Table 1):

- SWIKO17 (French-speaking Switzerland, German as the first, English as the second foreign language)
- SWIKO18 (German-speaking Switzerland, French as the first, English as the second foreign language)
- SWIKO19 (German-speaking Switzerland, German and English as languages of schooling)
- SWIKO22 (French-speaking Switzerland, German as the first, English as the second foreign language)

Text language[4]	German		French		English		TOTAL
	FL	LoS	FL	LoS	FL	LoS	
Year	10 & 11	11 & 12	11 & 12	10 & 11	10-12	10	
Oral							
Original texts n[5]	49	72	57	64	140	28	410
Transcripts n	42	72	7	0	108	8	237
Tokens n[6]	2'174	13'530	177	0	15'118	3'028	34'027
Written							
Original texts n	566	355	396	426	770	103	2'616
Transcripts n	543	347	322	384	684	102	2'382
Tokens n	23'737	23'667	17'567	27'584	45'173	8'556	146'284

Table 1. *Scope of SWIKO corpus (as of Dec. 2024).*

4 FL stands for *foreign language*, and LoS for *language of schooling* or language of instruction, which often corresponds to the learners' first language(s) (L1) (other language combinations were recorded in the metadata).

5 Blank or illegible documents as well as texts unrelated to the task are also counted as original texts. Not all texts collected were transcribed (e.g. oral French school language).

6 The number of tokens given refers to already transcribed texts.

2.2 Access and use

The corpus can be accessed via the SWIKOweb platform (https://ifm-swiko.unifr.ch). On the one hand, the website provides information on the project, including detailed transcription and annotation guidelines. On the other hand, it allows access to the learner corpus including metadata and customised visualisation tools. The data can be filtered according to various criteria, including an author's languages or gender, task characteristics, production conditions of the texts, and CEFR rating. Thanks to multi-layer architecture, the data can be searched and displayed at different levels, and these search queries can be combined with specific queries at other levels. For example, a search can be initiated for all occurrences of the lemma *"kein"* that contain an inflection error and that are followed by a noun. By default, the search results are displayed in tabular form, including concordances; alternatively, frequency distributions can be generated.

2.3 Data collection

To reflect the scope of communication in the classroom, eight systematically varied tasks were developed for data collection on the basis of the task concept in task-based language teaching (TBLT-approach, cf. Ellis et al., 2020). In open and restrictive tasks, learners were asked to produce both descriptive and argumentative texts on everyday and school-related topics. Table 2 provides an overview of the tasks and depicts the variation according to text type (descriptive vs. argumentative), topic (academic vs. personal) and structure (restrictive vs. open).

Code	Task description	Text type		Topic		Structure	
		des	*arg*	*acad*	*pers*	*rest*	*open*
SWI01	Answer brief personal questions	×			×	×	
SWI02	Describe graphic about Swiss pets	×		×		×	
SWI03	Discuss list of holiday options		×		×	×	
SWI04	Discuss list of most important inventions		×	×		×	
SWI05	Free self-portrait for class exchange	×			×		×
SWI06	Present a topic (out of 8 options given)	×		×			×
SWI07	Discuss later start/end of school		×		×		×
SWI08	Discuss language exchange instead of foreign language lessons		×	×	×		×

Table 2. *Task variation for SWIKO data collection.*

Figures 2a and 2b (below) illustrate task variation and production conditions using examples of two contrasting text types. Task SWI04 (2a) is categorised as argumentative, academic and restrictive. In this example, learners with German as the language of schooling were asked to write a text in French as a foreign language, with the specific requirement that this be done on paper. Task SWI05 (2b), by contrast, is a descriptive, personal and open task that learners with French as the language of schooling were required to complete in German as a foreign language, this time with the specific requirement that it be done orally on a computer.

Figure 2a. *Task SWI04 (argumentative, academic, restrictive).*

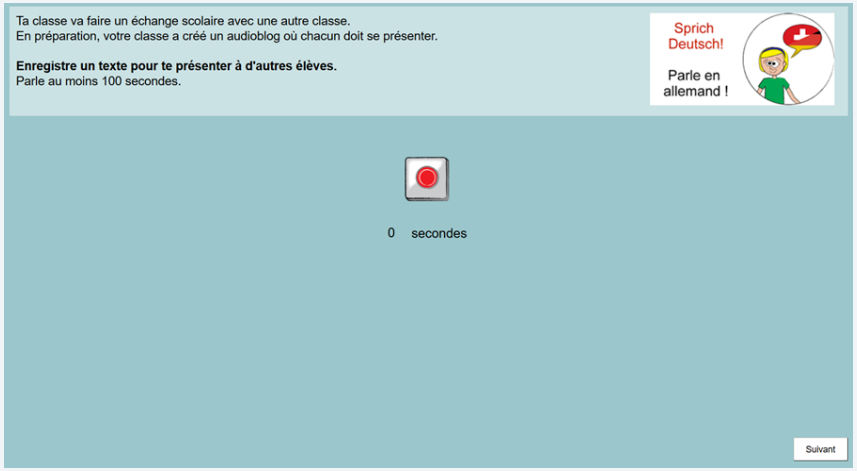


Figure 2b. *Task SWI05 (descriptive, personal, open) - Screenshot.*

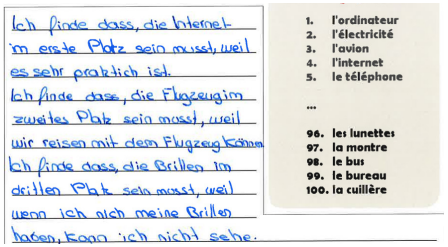
Between 2017 and 2022, data was collected among 340 students in 17 lower secondary school classes (HarmoS Years 10 to 12) in German- and French-speaking Switzerland. The data was collected over the course of two 45-minute lessons during regular school hours. All learners worked independently on four tasks each in two of the three languages available to them – for example, four tasks in the first, and four tasks in the second foreign language. Most learners completed the tasks in writing, either on paper or on a computer. In addition, four learners were randomly selected from each class to complete the same tasks orally in a separate room; two of these learners worked individually on a computer, while the other two participated in an interview situation with a researcher.

2.4 Data processing

Overall, the corpus linguistic data processing consisted of three steps (Figure 3 presents an example for written, paper-based productions in German): (1) manual transcription of the original text (0), (2) automatic linguistic annotation and (3) semi-automatic error annotation. The texts were also rated according to their CEFR levels (see below).

In a first step, each original text was transcribed manually in two versions, one that is as close as possible to the original (*Original Text*) and an orthographically correct reproduction (*Tagged Text*). The EXMARaLDA tool (Schmidt & Wörner, 2022) was used for oral productions, and XMLmind (Shafie, 2021) for written texts.

The transcripts were then converted into comma-separated value (csv) files using customised R scripts (R Core Team, 2022). Lemmas and part of speech were automatically annotated (POS annotation) using TreeTagger (Schmid, 2013) and the koRpus package (Michalke, 2019).

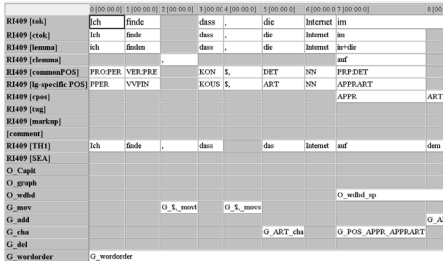


0) Original text

1) Transcript

	A	B	C	D	E	F
1	original	doc_id	token	common.F	de.POS.ta	lemma
2	Ich	SWI04_fD_	Ich	PRO:PER	PPER	ich
3	finde	SWI04_fD_	finde	VER:PRE	VVFIN	finden
4	dass	SWI04_fD_	dass	KON	KOUS	dass
5	,	SWI04_fD_	,	$,	$,	,
6	die	SWI04_fD_	die	DET	ART	die
7	Internet	SWI04_fD_	Internet	NN	NN	Internet
8	im	SWI04_fD_	im	PRP:DET	APPRART	in+die
9	erste	SWI04_fD_	erste	ADJ	ADJA	erst
10	Platz	SWI04_fD_	Platz	NN	NN	Platz
11	sein	SWI04_fD_	sein	VER:INF	VAINF	sein

2) POS annotation



3) Error annotation

Figure 3. *Data preparation in the SWIKO project.*

Afterwards, the files were converted in EXMARaLDA and all written German productions were augmented with a semi-automatic error annotation. To this end, a minimal target hypothesis (TH1) was first formulated, i.e., a version that is as close as possible to the original learner text, yet orthographically and grammatically correct (Lüdeling & Hirschmann, 2015). Then, the difference between the original text, the orthographically correct reproduction, and the target hypothesis was used to automatically annotate errors. Based on tagsets, orthographic (e.g. use of upper- and lowercase letters), syntactic (sentence construction) and grammatical (e.g. inflection) errors were tagged. The automated annotation was then manually reviewed and corrected where necessary.

In a final step, between 2020 and 2022, 47 trained raters assessed all 1550 written texts in German, French and English as a foreign language (DaF, FLE, and EFL), by assigning them a CEFR level. The analytical rating was conducted using a validated grid based on descriptors from *lingualevel* (Lenz & Studer, 2008) and the CEFR Companion Volume (Council of Europe, 2020). Four language criteria were assessed: *vocabulary*, with a focus on depth and breadth of words used; *grammar*, for features such as conjugation; *spelling*, for orthographic accuracy; and *text*, in terms of textual cohesion. To calculate a *fair score* for each text, the ratings were then analysed by means of a Many-facet Rasch measurement using Facets (Eckes, 2015; Linacre, 2022). These scores were calculated for each rating criterion individually as well as across all four criteria in the rubric. Thus, a CEFR language profile is available for each learner text in the corpus. Details of the ratings can be viewed at https://ifm-swiko.unifr.ch under Data processing / Rating.

3 Selected results

3.1 Impact of production conditions

Following the initial data collection, the influence of the medium on text length and lexical diversity was analysed in 1452 written texts (Karges et al., 2020). Texts produced by German- and French-speaking students in the language of schooling were equally long for both oral and written texts. However, French-speaking learners of German as a foreign language wrote shorter texts on the computer than on paper, whereas German-speaking learners of English as a foreign language produced longer texts on the computer than on paper. Texts composed in the school language were more lexically diverse than those written in the foreign language, while there were hardly any differences in this regard between texts written in the two foreign languages.

Linguistic features in 110 oral and 505 written texts in German as a language of schooling and as a foreign language were also compared (Karges et al., 2022). As expected, productions in the school language were on average longer and more lexically diverse than the foreign language texts. However, when using the language of schooling, participants spoke considerably more than their classmates wrote; by contrast, such differences were marginal among learners of German as a foreign language (DaF). In oral productions, foreign language learners resorted to other languages (even) more frequently than in written productions, presumably due to difficulties in finding the right words, with the proportion of French in German as a foreign language texts being twice as high as that of English in both modalities.

3.2 Interaction between task, linguistic features of productions, and ratings on the example of written DaF texts

The 544 written texts produced in German as a foreign language were analysed in order to shed light on the relationships between tasks, linguistic features of the productions, and ratings (Hicks, 2023; Hicks & Studer, 2024; Studer & Hicks, 2022). The tasks encompassed the three systematically varied task characteristics described above: text type (descriptive vs. argumentative), topic (academic vs. personal) and structure (restrictive vs. open). Linguistic indicators were calculated on the basis of the CAF framework: lexical and syntactic *complexity, accuracy,* and *fluency* (Housen et al., 2012; Michel, 2017).

All three task types were relevant factors in text length as well as in lexical *density* and *sophistication*, although no impact was seen with regard to lexical *diversity*. Furthermore, the text type influenced syntactic complexity, while a learner's familiarity with the subject matter also affected accuracy. Two anonymised examples written by the same learner illustrate these relationships (cf. Table 3 below): in the first text (SWI05, left), the learner was required to write a self-portrait, and in the second (SWI04, right) to express an opinion about the most important inventions. The self-portrait presented consists of a series of short simple sentences containing numerous nouns (relatively less frequent words). In the text that requires academic argumentation, the language becomes more syntactically complex: here, the learner wrote longer sentences, including subordinate or dependent clauses, that contain more adjectives and function words. However, the more demanding task also resulted in more errors.

SWI05: Self-portrait (anonymised) descriptive, personal, open	SWI04: Discussion of inventions argumentative, academic, restrictive
Hallo! Ich heisse Sandra und ich bin 15 Jahre alt. Ich liebe die Natur, Sport und ich lese gern aber ich spiele nicht gern Fussball und Bascket. «Shadow hunters» ist mein Lieblingserie und «La passe Miraire» ist mein Lieblingserie von Bücher. Ich habe zwei katzen, Sie heissen Simba und Luna. Mein Schwester heisst Laura und sie ist 18 aber ich habe keine Brüder. Ich liebe Ski fahren und Rad fahren aber mein Lieblingssport ist Klettern. Ich denke dass, ich Freundlisch, Neugierig, Schüchtern und Hilfsbereit bin. Bis bald!	*Für mich die Electricite ist im ersten platz weil ohne electricite es keine Lampe, keine Computer, keine Smartphone mehr gibt. Ich denke brille ist wichtiger als die sechs-und-neunzehnsten platz weil, für personnen wie kann nicht gut sehen ist ein sehr wichtiger punkt. Der Flieger ist für mich in die richtige platz und der bus auch aber der Stieft muss nicht in die liste bin weil es ist nicht ein sehr grossen invention.*

Table 3: *Two written productions by the same learner Ri513 (self-portrait on the left, discussion of inventions on the right).*

The task type also influenced the rating: personal and open texts received higher ratings than texts produced based on an academic and restrictive task. For example, 75 percent of the self-portraits were rated A2.1 (the HarmoS standard) or higher, while just over a third of the texts discussing inventions received a similar rating.

The relationship between rating and linguistic characteristics was also analysed. Longer and more lexically diverse texts were rated significantly higher, whereas texts with numerous spelling and grammatical errors as well as non-target words were rated significantly lower. Syntactic aspects were less relevant.

3.3 Pedagogical application

The findings reported in section 3.2 can serve as a useful data pool for teacher training programmes. The near-authentic texts enable a more realistic and nuanced understanding of students' performance. At the same time, they raise awareness for the significant impact of the type of task. As shown in Table 3, different tasks suit different levels of learner ability. While tasks such as self-portraits are good at assessing learners' current language skills, challenges such as discussing inventions can open a window to the next zone of development.

SWIKO can also be used directly in the classroom at lower secondary schools – for example to create teaching and learning materials for learning negation rules in DaF lessons (cf. Figure 4; details in Hicks & Studer, 2024). Concordances generated with the SWIKOweb platform (cf. section 2.2) can be used in an introductory phase to help learners deduce rules for the use of "*nicht*" and "*kein*" in implicit learning tasks. Learners can then collect typical combinations in a mind map, working either individually, in groups or as a class. Gap-fill exercises can also be created easily.

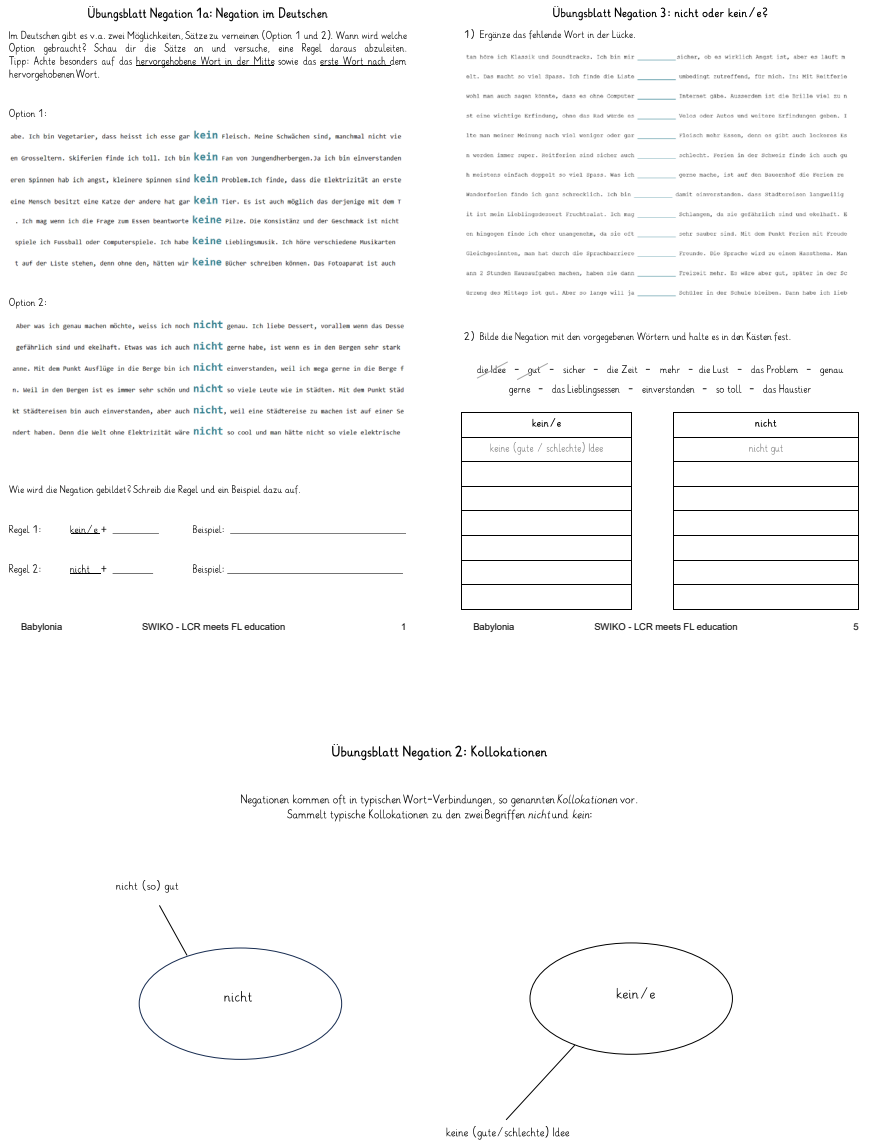


Figure 4*: Worksheets on negation rules in DaF, based on concordances from the SWIKO corpus.*

4 Final remarks

The Swiss learner corpus SWIKO is a comprehensive database containing oral and written texts produced by learners in foreign languages and school languages at lower secondary schools. The task-based productions were processed using corpus-linguistic methods, then reliably rated with reference to the CEFR. Detailed information on the resulting corpus and the project methodology is available in a separate corpus report (Hicks, Studer & Karges, forthcoming).

The multilingual SWIKO corpus occupies a unique position within the broader context of learner corpora, as it addresses the public-school setting and learners with limited language proficiency. In addition, the tasks and production conditions are systematically varied, and both the language of schooling as well as foreign languages taught at school are incorporated.

The SWIKOweb platform (https://ifm-swiko.unifr.ch) provides access to the database as well as additional tools that can be used to group, analyse and visualise the more than 2600 annotated learner texts contained in the corpus. Thus, the corpus is a valuable resource for both research and pedagogical purposes. Further research could investigate aspects such as which combinations of linguistic features in the target language emerge or are already mastered at which CEFR levels. From a pedagogical point of view, further data-based teaching material analogous to negation (Fig. 4) can be constructed to support language learning in particularly challenging areas.

5 Bibliografie
Bibliographie
Bibliography

5 Bibliografie Bibliographie Bibliography

Arnet-Clark, I., Frank Schmid, S., Ritter, G., & Rüdiger-Harper, J. (2013). *New World 1. English as a Second Foreign Language*. Klett & Balmer Verlag.

CIIP. (2023). *Plan d'études romand*. Portail CIIP. https://portail.ciip.ch

D-EDK. (2016). *Lehrplan 21 Fachbereich Sprache*. Deutschschweizer Erziehungsdirektoren-Konferenz. https://v-fe.lehrplan.ch/container/V_FE_DE_Fachbereich_SPR.pdf

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement*. Peter Lang. https://www.peterlang.com/document/1045610

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

Ende, K., Grotjahn, R., Kleppin, K., Mohr, I., & Ende, K. (with Goethe-Institut). (2013). *Curriculare Vorgaben und Unterrichtsplanung* (1. Auflage). Klett.

Europarat. (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen* (J. Quetz & G. Schneider, Übers.). Langenscheidt.

Europarat. (2020). *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lehren, lernen, beurteilen. Begleitband.* (J. Quetz & R. Camerer, Übers.). Klett.

Hicks, N. S. (2023). *Lexical features in adolescents' writing: Insights from the trilingual parallel corpus SWIKO*. Workshop on Profiling second language vocabulary and grammar, University of Gothenburg.

Hicks, N., & Studer, T. (2024). Learner corpora in foreign language education: Examples from the multilingual SWIKO corpus. *Babylonia Journal of Language Education*, *2*, 26–35. https://doi.org/10.55393/babylonia.v2i.388

Hicks, N. S., Studer, T. & Karges, K. (i.V.): The Swiss Leaner Corpus SWIKO. Young learner texts across languages, modes, and tasks.. *International Journal of Learner Corpus Research*.

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

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. https://doi.org/10.1515/zgl-2022-2050

Karges, K., Studer, T., & Wiedenkeller, E. (2020). Textmerkmale als Indikatoren von Schreibkompetenz. *Bulletin suisse de linguistique appliquée*, *No spécial Printemps 2020*, 117–140.

Konsortium HarmoS Fremdsprachen. (2009). *Fremdsprachen. Wissenschaftlicher Kurzbericht und Kompetenzmodell (provisorische Fassung)*. https://edudoc.ch/record/87025?ln=de

Konsortium ÜGK. (2019). *Überprüfung der Grundkompetenzen: Nationaler Bericht der ÜGK 2017: Sprachen 8. Schuljahr.* EDK & SRED. https://doi.org/10.18747/PHSG-coll3/id/385

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

Lenz, P., & Studer, T. (2008). *Lingualevel: Instrumente zur Evaluation von Fremdsprachenkompetenzen : 5.-9. Schuljahr*. Schulverlag.

Linacre, J. M. (2022). *Facets computer program for many-facet Rasch measurement* (Version 3.84.0) [Software]. Winsteps.com.

Lüdeling, A., & Hirschmann, H. (2015). Error annotation systems. In F. Meunier, G. Gilquin, & S. Granger (Hrsg.), *The Cambridge Handbook of Learner Corpus Research* (S. 135–158). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.007

Michalke, M. (2019, Mai 13). *Package «koRpus»*. [Software]. https://reaktanz.de/R/pckg/koRpus/koRpus.pdf

Michel, M. (2017). Complexity, Accuracy, and Fluency in L2 Production. In S. Loewen & M. Sato (Hrsg.), *The Routledge Handbook of Instructed Second Language Acquisition* (S. 50–68). Routledge.

Passepartout (Hrsg.). (2015). *Lehrplan Französisch und Englisch*. http://www.passepartout-sprachen.ch/services/downloads/download/533/get

Peyer, E., Andexlinger, M., Kofler, K., & Lenz, P. (2016). *Projekt Fremdsprachenevaluation BKZ: Schlussbericht zu den Sprachkompetenztests*. Institut für Mehrsprachigkeit.

R Core Team. (2022). *R: A Language and Environment for Statistical Computing* (Version 4.0.2) [Software]. R Foundation for Statistical Computing. http://www.R-project.org

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

Schmidt, T., & Wörner, K. (2022). EXMARaLDA – creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, 565–582. https://doi.org/10.1075/prag.19.4.06sch

Shafie, H. (2021). *XMLmind XML Editor* [Software]. XMLmind Software.

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.