QuAX-DaF

Quantitative Analyse von Texten
für den Deutsch-als-Fremdsprache-Unterricht

Daniel Jach

Letzte Änderung
16. März 2020

1 Einleitung

QuAX-DaF (**Qu**antitative **A**nalyse von Te**X**ten für **D**eutsch **a**ls **F**remdsprache) ist ein online Werkzeug für den DaF-Unterricht, erreichbar unter https://danieljach.shinyapps.io/quax-daf/. Das Programm analysiert deutschsprachige Texte mit Methoden der quantitativen Linguistik und erzeugt Material für Übungen zum Textverstehen und Wortschatzerwerb, angepasst an den Vokabelstand Ihrer Lernenden. QuAX ist für Lehrerinnen und Lehrer gemacht, möglichst praktikabel gestaltet und verlangt kein linguistisches oder statistisches Vorwissen. QuAX ist meines Wissens das erste DaF-Tool, das auf Erkenntnissen der gebrauchsbasierten Spracherwerbsforschung aufbaut und diese für den Unterricht nutzbar macht.

2 Gebrauchsbasierter Spracherwerb, Textverstehen und Wortschatz

Ergebnisse psycholinguistischer Forschung haben gezeigt, dass eine Fremdsprache – ähnlich wie eine Muttersprache – zu wesentlichen Teilen aus dem Gebrauch heraus gelernt wird (Ellis et al., 2015; Madlener, 2015). Solches erfahrungsbasiertes Lernen ist ein Nebenprodukt mentaler Aktivität und verläuft normalerweise implizit, das heißt ohne dass die Lernenden etwas davon mitbekommen. Gebrauchsbasierte Forschung geht dabei davon aus, dass Lernende einzelne Äußerungen, die ihnen im Sprachgebrauch begegnen, im Gedächtnis behalten und unbewusst hinsichtlich Form und Bedeutung miteinander vergleichen. Wiederkehrende Gemeinsamkeiten schreiben sich tiefer ins Gedächtnis ein, so dass sich sukzessive schematische Abbilder (Repräsentationen) geteilter grammatischer und anderer Eigenschaften herausbilden. Diese gebrauchsbasierten Abbilder unterscheiden sich von ihren Lehrbuch-Beschreibung wie Wildtiere von ihren im Zoo lebenden Artgenossen: Sie sind auf ihre Umgebung feinabgestimmt und in ein enges Netz ähnlicher Konstruktionen eingebunden (Diessel, 2019; Römer, 2004; Wagner, 2015). Die Häufigkeit, mit der eine bestimmte Konstruktion im Input erscheint, ist dabei ein entscheidender Faktor für ihren Erwerb und die Ausgestaltung ihrer Repräsentation (Diessel, 2016; H.-J. Schmid, 2018).

Der gebrauchsbasierte Ansatz ist nicht auf den Erwerb grammatischer Konstruktionen beschränkt, sondern bezieht auch den Erwerb von Wörtern ein. Korpuslinguistische Arbeiten haben etwa gezeigt, dass Wörter in Texten nicht wie einzelne Bausteine aneinander gereiht, sondern normalerweise in einem mehr oder weniger engen Verbund mit anderen Wörtern gebraucht werden (Hoey, 2004; Sinclair, 1991). Erwerb und Verarbeitung solcher Kollokationen hängen wiederum wesentlich von der Häufigkeit ab, mit der Lernende ihnen begegnen (Schmitt, 2012). Auch konnotative Anteile der Wortbedeutung werden überwiegend beiläufig aus dem Gebrauch heraus gelernt (Corrigan, 2004).

Auf diese Weise erwerben Lernende die Wörter und grammatischen Konstruktionen einer Sprache implizit aus ihren Gebrauchserfahrungen, ähnlich einem neuronalen Netzwerk, das wiederkehrende Muster in seinem Input erkennt, intern abspeichert und die entstehenden Abbilder kontinuierlich nachbessert, ohne dafür eines "Bewusstseins" zu bedürfen oder ein solches zu entwickeln (Elman, 1993). Damit werden die Existenz bewusster Lernvorgänge und die Wichtigkeit von Instruktionen keineswegs verneint, sondern ihre Rolle wird im Zusammenspiel mit impliziten Prozessen bestimmbar. Instruktionen haben in diesem Modell vor allem die Funktion, den Lernenden auf noch unbekannte Strukturen in seinem Input aufmerksam zu machen. Bewusste Lernvorgänge steuern ihre Verarbeitung so lange, bis sich ein internes Abbild herausbildet, das diese Aufgabe automatisch und unbewusst übernimmt (Ellis, 2015; Roehr-Brackin, 2018).

Für den Erwerb realistischer Konstruktionen und ihre idiomatische Feinabstimmung auf den Kon-

text ist also Erfahrung mit authentischem Sprachgebrauch von besonderer Bedeutung. Im gesteuerten Fremdsprachenerwerb im Herkunftsland ist authentischer Sprachgebrauch vor allem in Form von Texten (und aufgezeichneten Gesprächen) für Lernende erfahrbar. In bildungssprachlichen Domänen, zum Beispiel in der universitären Erwachsenenbildung, in der Vorbereitung auf ein Studium im Ausland oder im deutschsprachigen Fachunterricht an Auslandsschulen, dominieren Texte mit bildungs-, fach- und schriftsprachlichen Registern. Das Verstehen solcher Texte setzt aber normalerweise schon einen relativ großen Wortschatz voraus (Tschirner, 2019). Relevant für das Textverstehen sind zum einen besonders häufige Wörter, weil diese mehr als seltene Wörter zur Textdeckung beitragen (Nation, 2006). Korpuslinguistische Studien haben jedoch gezeigt, dass die oft für den Unterricht herangezogenen pragmatisch orientierten Wortlisten (zum Beispiel Langenscheidt, Goethe-Institut, Profile Deutsch) nur einen kleinen Teil solcher Wörter enthalten. Genauer gesagt: Statt häufiger Wörter, die wesentlich zum Textverstehen beitragen, enthalten diese Listen überwiegend relativ seltene Wörter, die meist für direkte Gespräche in bestimmten Alltagsdomänen gedacht sind (Tschirner, 2006). Auch ein häufigkeitsbasierter DaF-Vokabeltrainer existiert bislang nicht. Zum anderen spielen Wörter bildungs-, fach- und schriftsprachlicher Register eine entscheidende Rolle. Solche Wörter sind in den einschlägigen Wortlisten gewöhnlich nicht enthalten und kommen auch nur selten in alltagssprachlichen Texten vor. Sie sollten im Unterricht daher hervorgehoben und ihr Gebrauch gezielt vermittelt werden.

3 QuAX-DaF

3.1 Funktionsweise

QuAX nutzt Methoden und Einsichten der gebrauchsbasierten Spracherwerbsforschung, um deutschsprachige Texte statistisch zu analysieren und für den DaF-Unterricht auf unterschiedlichen Niveaustufen aufzubereiten. Eingegebene Texte werden zunächst lemmatisiert, das heißt flektierte Wortformen werden in ihre Grundform (Lemma) gebracht. Anschließend vergleicht QuAX die Häufigkeit der Wörter im Text mit ihrer Häufigkeit im Deutschen und bildet den Vergleich anschaulich ab. Das ermöglicht es den Lehrenden einzuschätzen, ob der eingegebene Text vor allem gebräuchliche oder ungewöhnlich viele seltene Wörter enthält und welche das sind.

Angepasst an den Vokabelstand der Lernenden erzeugt QuAX dann automatisch Übungsmaterial für den typischen unterrichtspraktischen Dreischritt: vor dem Lesen, beim Lesen und nach dem Lesen (Watkins, 2017, S. 20).

3.2 Textanalyse

QuAX eignet sich zur Analyse ganz unterschiedlicher bildungssprachlicher Texte, zum Beispiel sprach- und literaturwissenschaftlicher Fachtexte, journalistischer Nachrichten, Rezensionen und Reportagen, literarischer Werke oder Texte für den deutschsprachigen Fachunterricht. Geben Sie im Steuerungspaneel (siehe Abbildung 1) einen Text in das dafür vorgesehene Feld ein oder laden Sie einen Text im PDF-Format hoch. Sie können auch einen vorbereiteten Beispieltext aus der Dropdown-Liste darunter auswählen. Die Beispiele enthalten Texte, wie sie etwa in deutschsprachigem Geschichtsunterricht, Übungen zum Textverstehen oder sprach- und literaturwissenschaftlichen Seminaren im universitären DaF-Unterricht zum Einsatz kommen. Eine Analyse ist aus statistischen Gründen nur möglich, wenn Ihr Text eine gewisse Länge und Komplexität aufweist.

Wählen Sie für dieses Beispiel den Text *Kinderzeitmaschine: Über Ritter* aus. Der Text ist eine Zusammenstellung mehrerer Beiträge zum Thema *Ritter*, die auf der öffentlich geförderten Websei-

te *https://www.kinderzeitmaschine.de/* erschienen sind und sich an Kinder im Alter zwischen acht und dreizehn Jahren wenden. Der Text erscheint im Eingabefeld. Klicken Sie anschließend auf die Schaltfläche *Analyse beginnen* und zählen Sie leise bis zehn.

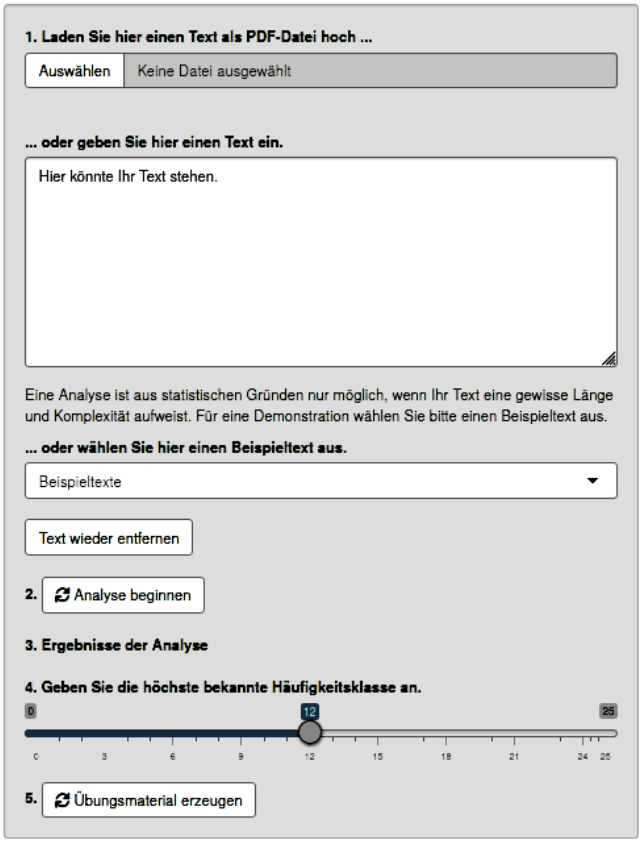

Abb. 1: Steuerungspaneel

QuAX lemmatisiert die Textwörter, ermittelt ihre Häufigkeit im Text und vergleicht sie mit ihrer Häufigkeit im Deutschen. Das Ergebnis der Analyse erscheint rechts neben dem Steuerungspaneel (siehe Abbildung 2).

Das Programm ermittelt die Anzahl der einzelnen Wörter in Ihrem Text (Tokens) und vergleicht sie mit der Anzahl verschiedener Wörter (Types). Das Type-Token-Verhältnis ist ein in der Korpuslinguistik häufig genutztes Maß, um die Komplexität eines Textes einzuschätzen. Der Wert steigt mit der Anzahl der Types und fällt mit der Anzahl der Tokens. Der analysierte Text erreicht einen Wert von 0,5. Das bedeutet, dass durchschnittlich jeder Type zweimal im Text als Token vorkommt. Zum Vergleich wird das Type-Token-Verhältnis in den „Sternchen"-Texten der bekannten PASCH-Initiative angezeigt[1], geordnet nach den Niveaustufen des Gemeinsamen Europäischen Referenzrahmens (GER). Der analysierte Text liegt zwischen den Niveaustufen B2/C1 und C1. Probieren Sie einige andere Beispieltexte aus. Der zweite Beispieltext, das Grimmsche Märchen *Rumpelstilzchen*, zum Beispiel liegt mit einem Type-Token-Verhältnis von 0,4 auf einem Niveau mit den PASCH-Texten der Niveaustufe B1. Die Reportage *Zu Hause im Baumhaus* übersteigt mit einem Wert von 0,57 den Wert der C1-PASCH-Texte. Relativ hohe Werte können auch von der Kürze der analysierten Texte bedingt sein. Ein längerer Auszug aus Siegfried Lenz' Roman *Deutschstunde* etwa, ein ohne Zweifel für fremdsprachige Leser schwieriger Text, erzielt mit 0,41 einen relativ niedrigen Wert, weil sich Types in längeren Texten häufiger wiederholen.

Anschließend werden Textwörter angezeigt, die QuAX unbekannt sind. Hier erscheinen häufig Eigennamen, technische Ausdrücke, Fremdwörter und seltene Nomen-Komposita. Im *Ritter*-Text sind QuAX zum Beispiel die Wörter *Lehnsherren, Schwertgürtel, Schwertleite, Kampfübungen, Nasalhelm* und *Kettenhaube* unbekannt. Außerdem werden im Gegenwartsdeutsch nicht mehr gebräuchliche Wörter wie *Turnei* und *Tjost* und äußerst seltene Wörter wie die Präteritum-Plural-Form *schützten* angezeigt. Daneben kommt es hin und wieder zu Missverständnissen. Zum Beispiel erkennt QuAX nicht, dass *Das* und *Die* Artikelwörter am Satzanfang bzw. am Beginn einer Überschrift sind. Bei der Analyse von Fachtexten werden hier auch Fachausdrücke aufgelistet. Aus der Definition von *Hermeneutischer Zirkel* etwa sind QuAX unter anderem die Wörter *Zirkelbewegung, Textverstehen, wirkungsgeschichtliches* und *Verstehenskonstituenten* nicht bekannt.

Wie sollten Sie mit diesen Wörtern im Unterricht umgehen? Das hängt auch von Ihren Lernzielen

[1] https://www.pasch-net.de/de/pas/cls/leh/unt/dst.html, heruntergeladen am 2. Januar 2020.

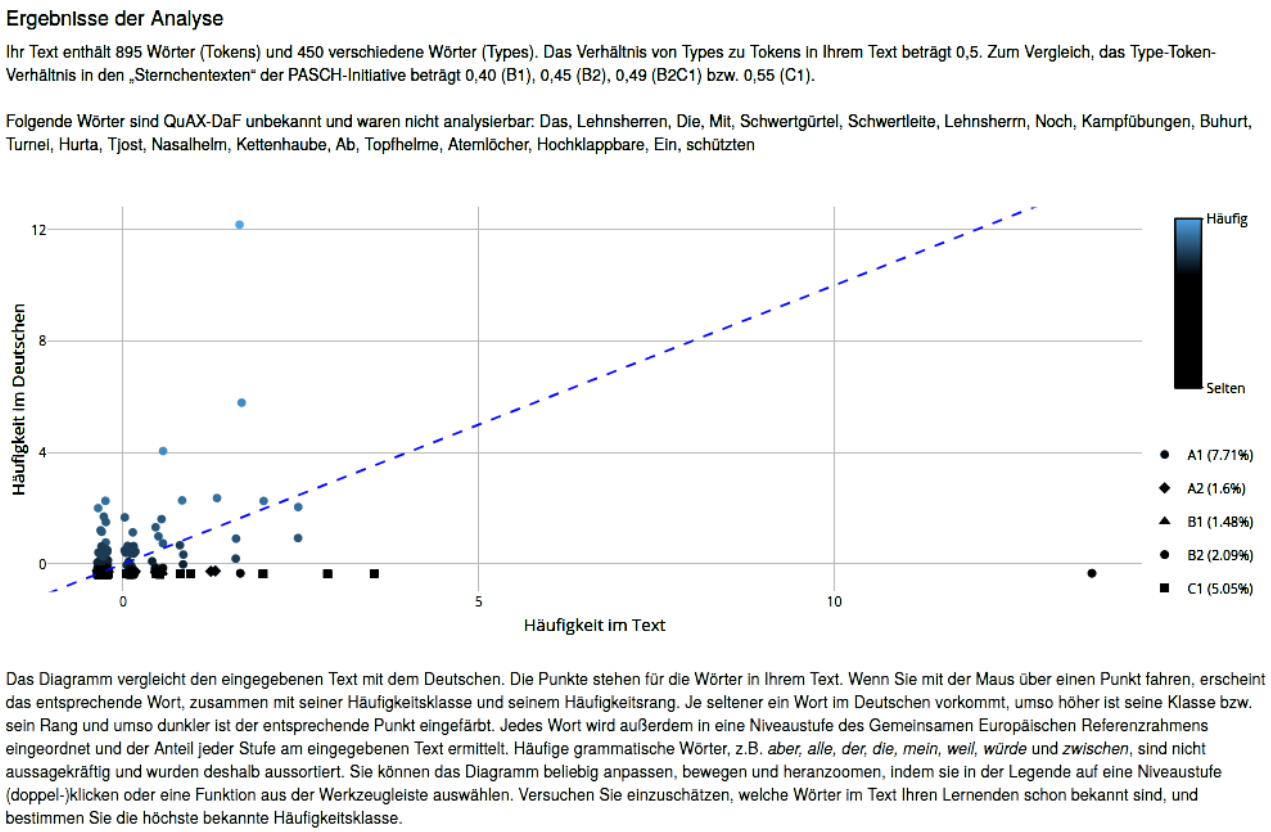

Abb. 2: Ergebnisse der Analyse

ab. Einige Wörter könnten im Text durch andere Ausdrücke ersetzt werden (z.B. *Kettenhaube* und *Nasalhelm* durch *Schutzhaube* und *Helm*), um die Lektüre zu vereinfachen. Andere Wörter beziehen sich auf sachlich relevante Konzepte, die vor der Lektüre eingeführt werden sollten (z.B. *Lehnsherr*). Wieder andere Wörter sachlich relevanter Konzepte sollten die Lernenden durch die Lektüre selbstständig erschließen (z.B. *Turnei* und *Tjost*, die im Text erklärt werden). Und Fachvokabular (z.B. *Textverstehen* und *Verstehenskonstituenten* aus dem *Hermeneutik*-Textbeispiel) könnten die Lernenden selbstständig während der Lektüre in Fachwörterbüchern nachschlagen.

Das anschließende Diagramm vergleicht den eingegebenen Text mit dem Deutschen. Die Punkte stehen für die Lemmas im Text. Die Lemmas erscheinen, wenn Sie mit der Maus über die Punkte fahren. Ihre Häufigkeit im Text ist auf der horizontalen Achse abgetragen, ihre Häufigkeit im Deutschen auf der vertikalen Achse. Das Wort *Ritter* zum Beispiel ist im analysierten Text häufig, im Deutschen dagegen eher selten, und erscheint dementsprechend in der rechten unteren Ecke des Diagramms. Angezeigt wird nicht die absolute Häufigkeit, sondern (grob gesagt) wie häufig oder selten ein Wort im Text bzw. im Deutschen verglichen mit einem durchschnittlich häufigen Wort vorkommt. Sogenannte Stoppwörter, die für den Textinhalt eine untergeordnete Rolle spielen, zum Beispiel *aber, am, denn, diesen, keine, manche, sonst, unser* und *zwischen*, wurden aussortiert, weil sie besonders häufig auftreten und das Diagramm darum stark verzerren würden. Sie können das Diagramm bewegen und heranzoomen (Werkzeugleiste).

Wörter, die im Deutschen häufig, aber im Text relativ selten vorkommen, erscheinen oberhalb der blauen Linie. Hierzu gehören im Beispieltext etwa *geben, sollen, kommen, neu, sagen, Jahr* und die Verschmelzungen von *in* und *zu* mit Artikelwörtern zu den Formen *im* und *zum*. Diese Wörter verursachen vermutlich keine besonderen Verstehensschwierigkeiten. Sofern sie Ihren Lernenden aber

noch nicht bekannt sind, sollten diese Wörter vorrangig gelernt werden. Sie sind zwar im untersuchten Text relativ selten, kommen aber in anderen Texten umso häufiger vor und tragen dort wesentlich zur Textdeckung bei.

Wörter, die im Deutschen selten, aber im Text relativ häufig vorkommen, erscheinen unterhalb der blauen Linie. Sie verdienen besondere Aufmerksamkeit. Sie tragen zur Textdeckung bei und spielen inhaltlich vermutlich eine wichtige Rolle in Ihrem Text. Dazu gehören im *Ritter*-Text die Wörter *Ritter, Schwert, Lanze, Knappe, Schild* und *Jahrhundert*. Eine Analyse von Grimms Märchen *Rumpelstilzchen* platziert an dieser Position bezeichnenderweise die Wörter *Männchen, Stroh, Gold* und *spinnen*. Die häufige direkte Rede im Märchentext zeigt sich in der überproportionalen Verwendung von *sprechen* (z.B. *Der König sprach zum Müller: ...*). Ein Auszug aus Anne Franks Tagebuch enthält besonders häufig die Wörter *Vater, Mutter* und den Namen von Annes Schwester, *Margot*, aber auch *Haus*, *kommen* und *gehen* kommen relativ häufig vor. Bei der Analyse von Fachtexten wie der Definition von *Hermeneutischer Zirkel* erscheint hier unter anderem fachspezifische Terminologie, die im normalen Sprachgebrauch nur selten verwendet wird, zum Beispiel *Text, Zirkel, Interpretation* und *Spirale*. Diese Wörter sind im Deutschen selten, erscheinen im untersuchten Text aber überproportional häufig und scheinen besonders prägend für den untersuchten Text zu sein.

Am Achsenschnittpunkt sammeln sich Wörter, die sowohl im Deutschen als auch im Text relativ selten sind, in einer dichten Wolke. Hierzu gehören im *Ritter*-Text zum Beispiel *kämpfen, lernen, Gesicht, allein, Holz, Gedicht, Krieg, dreieckig, zusammenbauen* und *ernennen*. Je größer und dichter diese Wolke ist, umso mehr verschiedene Wörter enthält der Text, einen umso größeren Wortschatz verlangt er von den Lernenden und umso schwieriger ist er dementsprechend zu verstehen.

Zusammen mit den Lemmas werden ihre Häufigkeitsklasse, ihr Häufigkeitsrang und eine Einordnung in die Niveaustufen des GER angezeigt. Je seltener ein Wort im Deutschen vorkommt, umso höher ist seine Häufigkeitsklasse und umso heller ist der entsprechende Punkt eingefärbt. Der Häufigkeitsrang bildet einen ähnlichen Zusammenhang kleinschrittiger ab, ist aber nicht so robust wie die Häufigkeitsklasse. Das Wort *Ritter* zum Beispiel erreicht die Häufigkeitsklasse 13 und den Häufigkeitsrang 3.248, beides relativ hohe Werte, die anzeigen, dass *Ritter* im Gebrauchsdeutsch der Gegenwart eher selten vorkommt.

Die GER-Niveaustufen werden mit verschiedenen Formen angezeigt, die in der Legende neben dem Diagramm beschrieben sind. Das Wort *Ritter* etwa wird der Niveaustufe C1 zugeordnet und erscheint im Diagramm daher als Quadrat. Wenn Sie in der Legende auf eine bestimmte Niveaustufe klicken, werden Wörter dieser Stufe aus dem Diagramm entfernt. Wenn Sie einen Doppelklick ausführen, werden nur Wörter dieser Stufe angezeigt. In der Legende wird auch der Anteil der Wörter verschiedener Niveaustufen im eingegeben Text angezeigt. Im *Ritter*-Text werden 15,59% der Textwörter auf dem Niveau C1 eingeordnet, 6,45% dem Niveau B2, 4,57% dem Niveau B1, usw. Diese Einordnung basiert *nicht* auf den pragmatisch orientierten Wortlisten des Goethe-Instituts, sondern auf dem Häufigkeitsrang der Wörter. Nach einer korpuslinguistisch gestützten Einschätzung von Tschirner (2019, S. 101) wächst der Wortschatz von Lernenden mit ihrem GER-Niveau ungefähr wie in Tabelle 1 abgebildet.

Tabelle 1: Zusammenhang von GER-Niveau und Wortschatz-Größe

GER-Niveau	A1	A2	B1	B2	C1
Wortschatz	800	1600	3200	4000	5500

QuAX ordnet dementsprechend die 800 häufigsten deutschen Wörter auf Niveau A1 ein, Wörter mit einem Häufigkeitsrang zwischen 801 und 1600 auf Niveau A2, usw. Ab einem Häufigkeitsrang von 4001 hat ein Wort das Niveau C1 erreicht.

Bestimmen Sie jetzt den Vokabelstand Ihrer Lernenden. Versuchen Sie einzuschätzen, welche Wörter im Text Ihren Lernenden schon bekannt und welche ihnen noch unbekannt sind, und ermitteln Sie so die höchste bekannte Häufigkeitsklasse. Wenn Ihren Lernenden zum Beispiel die Wörter *Ritter, Schild, Schwert, Gebrauch, Nase* und *Mut* schon bekannt sind, dann haben sie wahrscheinlich auch andere Wörter derselben Klasse (13), zum Beispiel *Wettkampf, Metall, Gesang, Ehre* und *erfreuen* sowie häufigere Wörter niedrigerer Klassen, etwa *Holz, Pferd, tödlich, Schlag* und *klein*, schon größtenteils erworben. Textwörter höherer Klassen (zum Beispiel *Langschwert, Lehnsherr, durchbohren, maßvoll, Lanze, adlig* und *Lösegeld*) sind ihnen dagegen möglicherweise nur vereinzelt oder nicht bekannt. Die höchste Ihren Lernenden noch bekannte Häufigkeitsklasse, zum Beispiel 13, nutzt QuAX als Annäherung an ihren Vokabelstand. Stellen Sie den Schieberegler im Steuerungspaneel entsprechend ein und klicken Sie auf die Schaltfläche *Übungsmaterial erzeugen.*

3.3 Übungsmaterial

Vor dem Lesen. QuAX erstellt eine Wortwolke aus den 25 häufigsten Wörtern Ihres Textes. Die Größe der Wörter bildet ihre Häufigkeit im Text ab: je häufiger, umso größer. Seltene Wörter, die Ihren Lernenden vermutlich unbekannt sind, und Stoppwörter erscheinen nicht in der Wortwolke. Wenn Sie auch unbekannte Wörter abbilden möchten, stellen Sie den Schieberegler im Steuerungspaneel auf die höchste Häufigkeitsklasse und klicken Sie noch einmal auf *Übungsmaterial erzeugen.* Sie können die Wolke herunterladen, indem Sie mit der rechten Maustaste auf sie klicken und „Bild speichern unter“ auswählen.

Vor dem Lesen können Sie die Wortwolke nutzen, um die Lernenden auf die Lektüre vorzubereiten. Auf der Grundlage der Wortwolke aktivieren die Lernenden nötiges Vokabular und bauen inhaltliche Erwartungen auf, die das anschließende Lesen unterstützen und ein globales Textverstehen stärken. Die folgende Übung ist, zum Teil wörtlich, Watkins (2017, S. 35) entnommen.

1. Teilen Sie den Lernenden mit, dass Sie ihnen für 15 Sekunden Wörter aus dem nächsten Lesetext zeigen werden. Sie sollen versuchen, sich in dieser Zeit so viel wie möglich zu merken.
2. Zeigen Sie die Wortwolke mit dem Beamer oder der digitalen Tafel für 15 Sekunden.
3. Lassen Sie die Lernenden anschließend in Gruppen diskutieren, was sie sich gemerkt haben.
4. Bitten Sie die Lernenden, vorherzusagen, wovon der Text handeln könnte.
5. Wenn nötig, zeigen Sie die Wortwolke noch einmal für 15 Sekunden.
6. Lassen Sie einige Gruppen ihre Vorhersagen im Plenum berichten.
7. Die erste Leseaufgabe sollte sein, zu überprüfen, ob die gemachten Vorhersagen richtig waren.

Wenn die Wortwolke unbekannte Wörter enthält, sollten Sie diese vor der Übung einführen. Die zeitliche Begrenzung erzeugt eine spielerische Spannung und erhöht den emotionalen Gehalt dieser Übung. Das bildet einen wünschenswerten Kontrast zur anschließenden Phase ruhiger Einzelarbeit beim Lesen.

Eine ähnliche Übung ohne Zeitlimit verlangt von den Lernenden, zur Vorbereitung der Lektüre in Gruppen möglichst viele Assoziationen zu brainstormen. Dazu gehören zum Beispiel Wörter ähn-

licher Bedeutung, emotionale Konnotationen, typische Kollokationen und Textsorten, in denen die abgebildeten Wörter gebräuchlich sind (Watkins, 2017, S. 32).

Beim Lesen. QuAX hebt Wörter, die Ihren Lernenden vermutlich unbekannt sind, im Text visuell hervor. So sind diese Wörter schnell auffindbar und ziehen außerdem die Aufmerksamkeit der Lesenden verstärkt auf sich. Abhängig von Ihren Lernzielen können Sie diese Wörter im Text durch gebräuchlichere Synonyme oder Paraphrasen ersetzen, um den Lernenden die Lektüre zu erleichtern; oder gezielt die Aufmerksamkeit der Lernenden auf diese Wörter lenken, um ihren idiomatischen Gebrauch im Kontext zu vermitteln oder ihre Bedeutung aus dem Kontext zu erschließen. Sie können den modifizierten Text mit gedrückter linker Maustaste markieren und dann wie gewohnt in ein Textbearbeitungsprogramm Ihrer Wahl kopieren.

Wenn Sie den Schieberegler so einstellen, dass auch schon bekannte inhaltliche Schlüsselwörter des Textes hervorgehoben sind, eignet sich der markierte Text auch für eine Übung, die globales Textverstehen und flüssiges Lesen stärkt (angelehnt an Watkins, 2017, S. 115).

1. Wenn nötig, führen Sie die Schlüsselwörter vor der Lektüre ein.
2. Bitten Sie dann die Lernenden, den Text zügig zu lesen und sich nicht auf jedes einzelne Wort zu konzentrieren, sondern ein globales Verständnis anzustreben. Geben Sie ein Zeitlimit vor, das eine zügige Lektüre verlangt. Wenn Sie den Text nicht bis zum Ende lesen, ist das auch in Ordnung.
3. Die Lernenden lesen den Text im vorgegebenen Zeitlimit.
4. Anschließend notieren sie die Ergebnisse ihrer Lektüre.
5. Die Lernenden lesen den Text erneut im vorgegebenen Zeitlimit.
6. Anschließend ergänzen sie ihre Notizen mit den hinzugewonnenen Erkenntnissen.
7. Wiederholen Sie diese Sequenz ein drittes Mal.
8. Teilen Sie die Lernenden in Gruppen ein, in denen sie ihre Notizen vergleichen und vervollständigen.

Die zeitliche Begrenzung bei der Lektüre führt im ersten Durchgang dazu, dass die Lernenden ihre Aufmerksamkeit nicht auf einzelne Wörter richten und linear von einem Wort auf das nächste verschieben, sondern weiter über den Text verteilen. Der Blick bleibt dabei an den visuell hervorgehobenen Schlüsselwörtern hängen. Der erste Versuch, auf dieser fragmentarischen Grundlage die Textbedeutung zu erschließen, stärkt deszendente („top-down“) Verstehensprozesse und damit ein globales Textverstehen. In den folgenden Durchgängen verschiebt sich der Aufmerksamkeitsfokus auf den unmarkierten Kontext, so dass die Lernenden ihr Verständnis der Schlüsselwörter nachbessern und weitere Bestandteile der Textbedeutung aszendent („bottom-up“) erschließen und ihr Modell der Textbedeutung ergänzen und gegebenenfalls korrigieren können. Das wiederholte Lesen unter Zeitdruck übt außerdem das flüssige, schnelle Lesen eines Textes ein.

Nach dem Lesen. Nach dem Lesen folgt häufig eine Aufgabe, die die Lernenden zur vertiefenden und vernetzenden Auseinandersetzung mit dem Textinhalt anleitet, zum Beispiel eine Diskussion oder ein Rollenspiel. QuAX erzeugt dagegen Material für eine Übung, die Erwerb und Verarbeitung seltener Wörter aus dem Text unterstützt. Die Übung ist Watkins (2017, S. 105) entliehen. Den englischen Namen der Übung, *Rogue words*, habe ich mit *Falsche Fuffziger* übersetzt. Falsche Fuffziger sind eigentlich gefälschte Geldstücke (Fünfzig-Pfennig-Stücke) einer nicht mehr gebräuchlichen deutschen Währung, bezeichnen aber umgangssprachlich im übertragenen Sinne unaufrichtige, verlogene Menschen.

QuAX erstellt eine Tabelle aus zehn bekannten Textwörtern, zehn unbekannten Textwörtern (also Wörtern, die Sie bei der Analyse als vermutlich unbekannt eingeschätzt haben) und zwanzig Pseudowörtern. Ein Pseudowort ist eine Abfolge von Buchstaben, die den phonotaktischen Regeln einer Sprache entspricht, also ein Wort dieser Sprache sein könnte, aber dennoch nicht in dieser Sprache existiert, für das Deutsche zum Beispiel *zünnen, hogt, veuten, verspreisen, Tusten, Zaul* und *Neklepen*. In psycholinguistischen Untersuchungen werden Pseudowörter verwendet, um ihre Verarbeitung mit der von existierenden Wörtern zu vergleichen.

1. Erklären Sie den Lernenden, dass die Tabelle bekannte und unbekannte Wörter enthält. Die Aufgabe der Lernenden lautet, die unbekannten Wörter in der Tabelle so schnell wie möglich zu identifizieren. Sie sollten die als falschen Fuffziger erkannten Einträge einkreisen.
2. Betonen Sie, dass die Lernenden schnell arbeiten sollten, indem sie ein eng begrenztes Zeitlimit vorgeben oder die Aufgabe im Stil eines Rennens gestalten, das derjenige gewinnt, der als erster alle falschen Fuffziger erkannt hat.
3. Sobald die Lernenden die Aufgabe beendet haben, kontrollieren Sie die Antworten gemeinsam im Plenum.

Sie können die Übung wiederholen, indem Sie die Tabelle variieren (klicken Sie hierfür noch einmal auf *Übungsmaterial erzeugen*) und die Lernenden noch einmal gegen die Uhr oder gegen sich selbst antreten lassen. Die Aktivität verbessert die unmittelbare, ganzheitliche Erkennung der Wörter und trägt so zu ihrer Verankerung im Gedächtnis bei. Vor allem Wörter, die vor der Lektüre noch unbekannt waren, profitieren von einer Wiederholung.

Tabelle. QuAX generiert außerdem eine Tabelle, die einen Überblick über eine Auswahl der verwendeten Variablen gibt. Die Tabelle enthält die Lemmas der Textwörter (Lemma), ihre Häufigkeit im Text (n), ihre Häufigkeit im Deutschen pro eine Million Wörtern (f), ihren Häufigkeitsrang (Rang), ihre Häufigkeitsklasse (Klasse) sowie eine Einordnung in die GER-Niveaustufen (GER). Sie können die Spalten der Tabelle sortieren, nach bestimmten Einträgen durchsuchen und die Tabelle als CSV-Datei herunterladen. Die Abkürzung *CSV* bedeutet *Comma-separated values* („Komma-getrennte Werte“). Sie können die Datei in ein beliebiges Tabellenkalkulationsprogramm importieren. Achten Sie dabei darauf, als Begrenzungszeichen das Komma auszuwählen.

3.4 Korpuslinguistische Ressourcen

Die Häufigkeitsverteilung der Textwörter im Deutschen ermittelt QuAX auf der Grundlage einer Textsammlung (Korpus) der Leipzig Corpora Collection (LCC) (Goldhahn et al., 2012). Die LCC ist eine für Forschung und Lehre frei nutzbare Sammlung maschinenlesbarer Texte verschiedener Sprachen. QuAX nutzt die deutschsprachigen Korpora *Mixed-typical* (Jahr 2011), *News* (Jahre 2010 bis 2015) und *Wikipedia* (Jahre 2007, 2010, 2014, 2016). Die Korpora enthalten zusammen etwa zehn Millionen Sätze aus online gecrawlten Texten deutschsprachiger Zeitungen, Internetseiten und Einträgen in dem online Lexikon Wikipedia.

Für die Lemmatisierung nutzt QuAX das Programm *TreeTagger* (H. Schmid, 1995). Bei der automatischen Lemmatisierung wird jedes Wort aus seiner flektierten Form in seine Grundform gebracht. Zum Beispiel werden die verschiedenen Formen definiter Artikel (*der, dem, den, des, das, die*) einer einheitlichen Grundform (*die*) zugeordnet. Auch Verben, Nomen und alle Wortarten, die in unterschiedlicher Flexionsform auftreten (zum Beispiel *trinken, trank, getrunken* bzw. *Glas, Gläser,*

Gläsern), werden mit ihrer Grundform versehen (*trinken* bzw. *Glas*). Die Grundform (Lemma) wird üblicherweise verwendet, wenn die Häufigkeit eines Wortes in einem Korpus bestimmt werden soll.

QuAX bildet die Häufigkeit der Lemmas im eingegebenen Text auf ihre Häufigkeit im Korpus ab. Um Vergleichbarkeit herzustellen, werden die absoluten Häufigkeiten vorher z-transformiert, d.h. auf den Achsen wird angezeigt, wie viele Standardabweichungen die Häufigkeit eines Wortes von der durchschnittlichen Häufigkeit aller Wörter im Text bzw. im Korpus entfernt liegt. Dieses Verfahren büßt umso mehr an Verlässlichkeit ein, je kürzer der analysierte Text ist. Aus der absoluten Häufigkeit der Lemmas im Korpus ermittelt QuAX ihre Häufigkeitsklasse. Die Häufigkeitsklasse eines Wortes ist der Koeffizient aus der absoluten Häufigkeit des häufigsten Lemmas im Korpus und der absoluten Häufigkeit des untersuchten Lemmas, logarithmiert zur Basis zwei und aufgerundet. Das häufigste Lemma im Korpus ist der definite Artikel, der 17.352.900 Mal vorkommt. Das Lemma *Obst* zum Beispiel kommt dagegen nur 1.426 Mal im Korpus vor. *Obst* fällt also in die Häufigkeitsklasse $\log2(^{17.352.900}/_{1.426})$, aufgerundet 14. Das Lemma *Begriff* taucht dagegen 18.966 Mal im Korpus auf. Seine Häufigkeitsklasse ist also $\log2(^{17.352.900}/_{18.966})$, aufgerundet 10. Je seltener ein Lemma, umso höher seine Häufigkeitsklasse. Die Häufigkeitsklasse ist ein in der Korpuslinguistik verbreitetes Maß, um die relative Häufigkeit verschiedener Lemmas robust über Korpora unterschiedlicher Größe und Inhalte hinweg einzuschätzen.

Für die Aufbereitung der Korpora, alle Berechnungen und die online Applikation nutzt QuAX die Programmiersprache *R* (R Core Team, 2013) und das Paket *Shiny* (Chang et al., 2017). Pseudowörter wurden mit dem online frei verfügbaren Programm *Wuggy* (Keuleers & Brysbaert, 2010) erzeugt.

Literatur

Chang, W., Cheng, J., Allaire, J., Xie, Y. & JonathanMcPherson. (2017). Shiny: Web Application Framework for R [https://CRAN.R-project.org/package=shiny].

Corrigan, R. (2004). The acquisition of word connotations: Asking 'What happened'. *Journal of Child Language*, *31*(2), 381–398. https://doi.org/10.1017/S0305000903005981

Diessel, H. (2016). Frequency and lexical specificity in grammar: A critical review. In H. Behrens & S. Pfänder (Hrsg.), *Experience Counts: Frequency Effects in Language* (S. 209–238). Berlin, Germany, De Gruyter.

Diessel, H. (2019). *The Grammar Network: How Linguistic Structure is Shaped by Language Use*. Cambridge, England, Cambridge University Press.

Ellis, N. C. (2015). Implicit AND Explicit Language Learning: Their dynamic interface and complexity. In P. Rebuschat (Hrsg.), *Implicit and explicit learning of languague*. Amsterdam, The Netherlands, John Benjamins.

Ellis, N. C., Römer, U. & O'Donnell, M. (2015). Second language constructions: Usage-based acquisition and transfer. In J. W. Schwieter (Hrsg.), *The Cambridge Handbook of Bilingual Processing* (S. 234–254). Cambridge, England, Cambridge University Press.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99. https://doi.org/10.1016/0010-0277(93)90058-4

Goldhahn, D., Eckart, T. & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. (N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis, Hrsg.) [https://wortschatz.uni-leipzig.de/de/download, heruntergeladen Februar 2020]. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (Hrsg.), *Proceedings of the 8th International Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, European Language Resources Association (ELRA). https://wortschatz.uni-leipzig.de/de/download, heruntergeladen Februar 2020.

Hoey, M. (2004). The textual priming of lexis. In G. Aston, S. Bernardini & D. Stewart (Hrsg.), *Corpora and Language Learners* (S. 21–41). Amsterdam, The Netherlands, John Benjamins.

Keuleers, E. & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator [http://crr.ugent.be/programs-data/wuggy, heruntergeladen am 13. Februar 2020]. *Behavior Research Methods*, *42*(3), 627–633. https://doi.org/10.3758/BRM.42.3.627

Madlener, K. (2015). *Frequency effects in instructed second language acquisition*. De Gruyter.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*, 59–82.

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria, R Foundation for Statistical Computing. https://www.R-project.org/

Roehr-Brackin, K. (2018). *Metalinguistic awareness and second language acquisition*. New York, NY, Routledge.

Römer, U. (2004). Comparing real and ideal language learner input: The use of an EFL textbook corpus in corpus linguistics and language teaching. In G. Aston, S. Bernardini & D. Stewart (Hrsg.), *Corpora and language learners* (S. 151–168). Amsterdam, The Netherlands, John Benjamins.

Schmid, H.-J. (2018). Unifying entrenched tokens and schematized types as routinized commonalities of linguistic experience, In *Yearbook of the German Cognitive Linguistics Association*. https://doi.org/10.1515/gcla-2018-0008

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Schmitt, N. (2012). Formulaic language and collocation. In C. A. Chapelle (Hrsg.), *The Encyclopedia of Applied Linguistics* (S. 1–10). Malden, MA, Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal0433

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford, England, Oxford University Press.

Tschirner, E. (2006). Häufigkeitsverteilungen im Deutschen und ihr Einfluss auf den Erwerb des Deutschen als Fremdsprache (E. Corina, C. Marello & C. Onesti, Hrsg.). In E. Corina, C. Marello & C. Onesti (Hrsg.), *Atti del XII Congresso Internationale di Lessicografia*, Alessandria, Italy, Edizioni dell'Orso.

Tschirner, E. (2019). Der rezeptive Wortschatzbedarf im Deutschen als Fremdsprache (T. Studer, I. Thonhauser & E. Peyer, Hrsg.). In T. Studer, I. Thonhauser & E. Peyer (Hrsg.), *Akten der XVI. Internationalen Deutschlehrertagung (IDT)*, Erich Schmidt.

Wagner, J. (2015). Designing for Language Learning in the Wild: Creating social infrastructures for second language learning. In S. W. Eskildsen & T. Cadierno (Hrsg.), *Usage-Based Perspectives on Second Language Learning* (S. 75–101). De Gruyter.

Watkins, P. (2017). *Teaching and Developing Reading Skills*. Cambridge, England, Cambridge University Press.