11.4 Datenauswertung

Für die Datenauswertung wurden wissenschaftliche Standards aus der quantitativen Korpusanalyse berücksichtigt (u.a. Mezger et al., 2016; Perkuhn et al., 2012). Das Zählen von Einheiten ist die Grundlage quantitativer korpusanalytischer Methoden (Meißner et al., 2016, S. 309f.). Die Abläufe in der Datenauswertung waren daher mit dem Ziel verbunden, die Transkripte so aufzubereiten, dass Listen über häufige Lemma-Types (Nennform des Lexems, z.B. Nominativ Singular bei Substantiven) und Wortkombinationen generiert werden konnten. Anhand der Listen konnten verschiedene formalsprachliche Aspekte (u.a. frequente und weniger frequente Einheiten, Meißner et al., 2016, S. 310) des natürlichen mündlichen Sprachgebrauchs gewonnen werden (Tab. 32).

Tab. 32: formalsprachliche Aspekte der Korpusanalyse

Bezeichnung	**Kernvokabular**
Token	Anzahl der insgesamt verwendeten Wörter (Lexeme)
Lemma-Types	Anzahl der verschiedenen Wörter (Lexeme)
Häufigkeit	Absolute Häufigkeit pro Wort
Rang	Wörter werden im Hinblick auf die Häufigkeit absteigend sortiert (größter Wert = Rang 1).
Streuung (S)	Anzahl aller Fälle pro Lemma-Types
Bezeichnung	**feste Wortkombinationen**
Token	Anzahl der insgesamt verwendeten Wortkombinationen
Types-Kombinationen	Anzahl der verschiedenen Wortkombinationen
Häufigkeit	Absolute Häufigkeit pro Wortkombination
Rang	Wortkombinationen werden im Hinblick auf die Häufigkeit absteigend sortiert (größter Wert = Rang 1).
Streuung (S)	Anzahl aller Fälle pro Types-Kombination

Im Folgenden werden die Schritte der Datenauswertung für die Analyse des Kernvokabulars sowie für die festen Wortkombinationen beschrieben.

Kernvokabular

Das Auswertungsschema für die Analyse des Kernvokabulars war stark angelehnt an die Untersuchung von Boenisch (2014b), um die gewonnenen Ergebnisse mit dem Referenzkorpus aus Boenisch (2013, Kl. 2 und 4) vergleichen zu können (Kap. 10.1). Abweichend zum Analyseschema von Boenisch (2014b) wurde in der vorliegenden Arbeit *machen* nicht zu den Hilfsverben gezählt, sondern als Vollverb behandelt (S. 14ff.). Ebenso wurde *werden* als Hilfsverb gezählt.

Mit dem Erweiterungsprogramm MAXDictio konnten die Transkripte nach den oben genannten formalsprachlichen Aspekten (Tab. 34) quantitativ untersucht werden. Das Leerzeichen nach einem Token bildete das Grenzsymbol für ein Wort. Über die Funktion »Worthäufigkeiten« konnte pro Transkript eine Wortschatzliste generiert werden, die u.a. Auskunft über die verwendeten Häufigkeiten der einzelnen Wörter gab. Die Voreinstellung »Stopp-Liste« ermöglichte es, die Transkripte nach dem Forschungsanaliegen zu bereinigen (z.B. Ausschluss von arabischen Zahlen) (Abb. 18). Über die Voreinstellung »Wörter lemmatisieren/German« konnten die genutzten Wortformen in Lemma-Types umgewandelt werden.

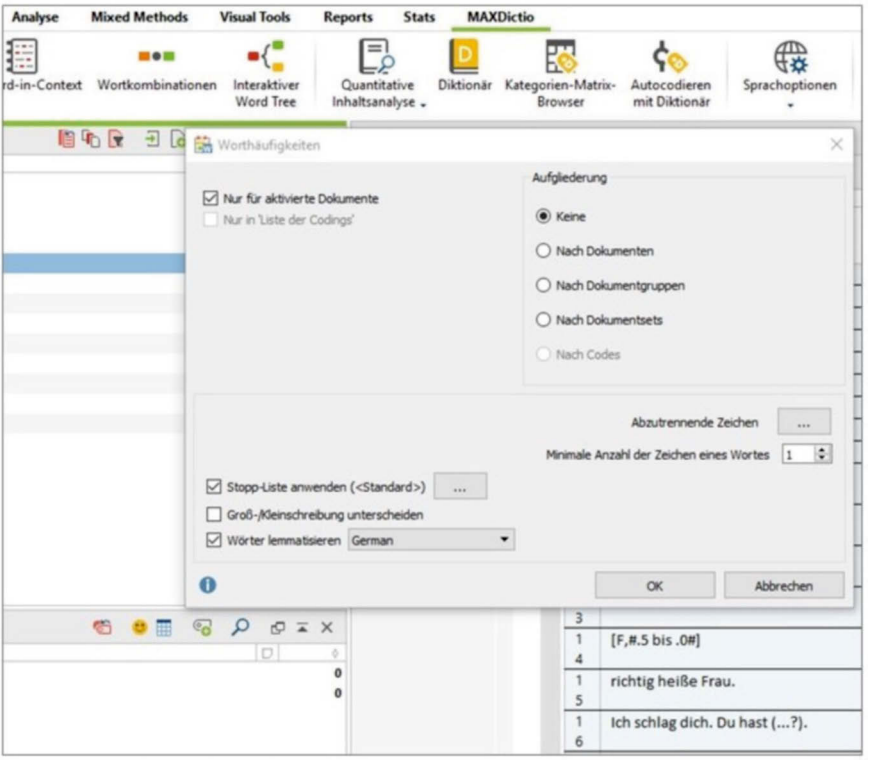

Abb. 18: Funktion Worthäufigkeiten in MaxDictio

Die aus dem Primärkorpus generierte Wortliste in MAXDictio konnte nicht uneingeschränkt für die Datenauswertung genutzt werden und musste daher in Excel exportiert und manuell weiter überarbeitet werden. Die Überarbeitung erfolgte über die Funktion »Liste der Fundstellen« in MAXDictio. So war es möglich, für jedes lemmatisierte Wort die tatsächliche Verwendung im Transkript nachzuvollziehen. Am Beispiel des Lemmas *weit* (Abb. 19) wird ersichtlich, dass nicht allein über die Lemmatisierung deutlich wird, ob es sich um das Adjektiv *weit* oder um das Adverb *weiter* handelte.

Die Kontextanalyse über »Liste der Fundstellen« ermöglichte eine genauere Wiedergabe der verwendeten Lemma-Types. Das Lemma *weit* wurde entsprechend des Kontextes als Adverb annotiert und zu *weiter* umbenannt. Zusammengesetzte Verben, die in einer Äußerung separiert auftraten (*Mach die Tür auf.*), wurden in einer manuellen Auswertung identifiziert und in der Grundform aufgeführt (*aufmachen*, Boenisch, 2014b).

Die notwendigen Überarbeitungsschritte werden in der folgenden Tabelle begründet aufgeführt (Tab. 33).

Die manuelle Überarbeitung von computergestützten Wortschatzlisten ist nicht unüblich. Beispielsweise machen Fandrych, Meißner und Wallner (2018) in einer Veröffentlichung anhand des Korpus »Gesprochene Wissenschaftssprache kontrastiv« (GeWiss) exemplarisch deutlich, dass »die aktuell verfügbaren für schriftsprachliche Korpora entwickelten Verfahren zur automatischen Aufbereitung […] für mündliche Korpora noch keine zufriedenstellenden Ergebnisse [liefern]« (Fandrych et al., 2018, S. 8; auch Ädel & Erman, 2012, S. 84; Perkuhn et al., 2012, S. 77; Ziem & Lasch, 2013, S. 72). In ihren Ausführungen beziehen sich die Autor:innen auf die orthographische Normalisierung, die Wortartenannotation (POS-Tagging), die Lemmatisierung und die Annotierung pragmatischer Kategorien (Fandrych et al., 2018, S. 8).

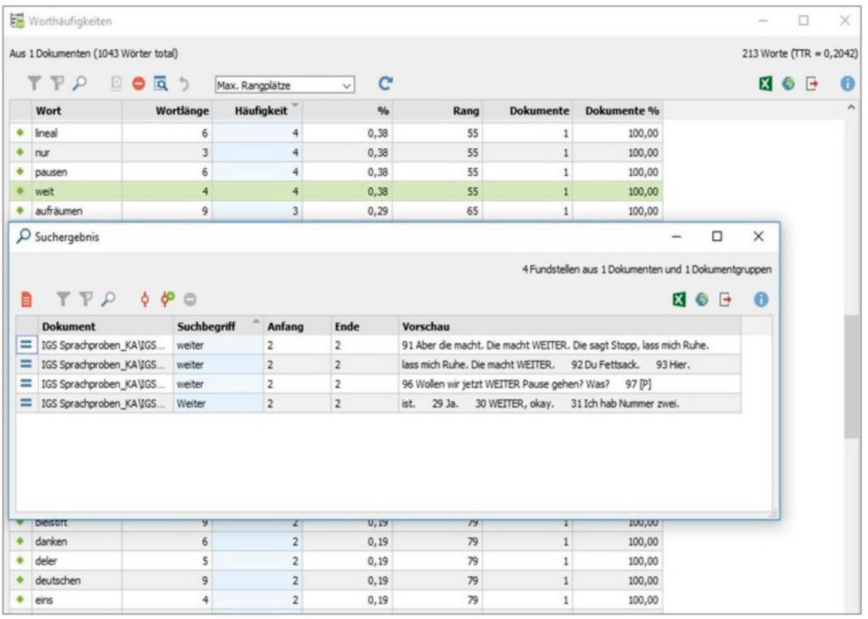

Abb. 19: Liste der Fundstellen am Beispiel des Lexems »weiter«

Tab. 33: Manuelle Überarbeitungsschritte für die Analyse des Kernvokabulars (Kv)

Manuelle Überarbeitungsschritte (Analyse Kernvokabular)	**Begründung**
Anpassung der lemmatisierten Wörter	Um den verwendeten Wortschatz der Untersuchungsgruppe im Hinblick auf das Kernvokabular vergleichen zu können, wurde auf einen Teildatensatz aus den Forschungsdaten von Boenisch (2014b; 2013) zurückgegriffen. Dieser bildet den Gebrauchswortschatz von 13 Schüler:innen aus der 2. Klasse der Grundschule sowie 15 Schüler:innen aus der 4. Klasse der Grundschule ab. Boenisch (2013) nutzte in seiner Analyse bei den Verben die Grundform (*habe*, *hast* > *haben*) als minimale Einheit. Darüber hinaus zeigt die Wortschatzliste, dass bei den Pronomen und Adjektiven der Nominativ Singular Feminin (*jeder*, *jedes* > *jede*) oder Neutrum (*dieser*, *diese* > *dies*) verwendet wurde. Damit der Vergleichsdatensatz von Boenisch (2013) effizient genutzt werden konnte, wurden die lemmatisierten Formen aus MAXDictio entsprechend angepasst (z.B. *jed* > *jede*; *ander* > *andere*; *spiel* > *spielen*). Die Lemmatisierung in MAXDictio erscheint nicht durchgehend nachvollziehbar und fehlerfrei. Zum Beispiel befinden sich hinter dem Lemma *weg* die Wortformen *weg* und *wegen* oder das Lemma *malen* steht für *mal*. Nach E-Mail Kontakt mit dem Kundenservice vom 26.03.2018 wurde darauf hingewiesen, dass MAXQDA externe Lemmata-Listen verwendet, die von dem/der Anwender:in frei editierbar sind. Für das Deutsche wird die Liste »Deutsche-Morphologie-Daten« von Daniel Naber (www.danielnaber.de/morphologie/) genutzt. Die Entwickler von MAXQDA haben keinen unmittelbaren Zugriff auf die Gestaltung der Liste und können Fehlermeldungen lediglich an die Autor:innen der Liste weiterleiten. Für die Wortschatzanalyse wurde die Lemmata-Liste in Bezug auf die Pronomen, die bestimmten Artikel und das Hilfsverb *sein* modifiziert. Abweichungen vom Transkriptionssystem (z.B. *gibts* > *gibt es*) und Rechtschreibfehler mussten manuell korrigiert werden. Zusammengesetzte Verben, die getrennt auftraten, wurden in die Grundform umgewandelt (z.B. *Mach zu* > *zumachen*).
Manuelle Auswertung der Pronomen und Artikel	Die Pronomen und Artikel wurden unter übergreifenden Lemma-Types subsummiert. So befanden sich alle bestimmten Artikel unter dem Lemma *der*, alle Personalpronomen unter *ich*, Possessivpronomen befanden sich teilweise unter *mein* usw. Diese Art der Lemmatisierung musste im Hinblick auf die Analyse von Kernvokabular zwingend aufgebrochen werden, um die verwendeten Wörter möglichst genau abzubilden (Modifizierung der Lemmata-Liste).

Manuelle Überarbeitungsschritte (Analyse Kernvokabular)	**Begründung**
Anpassung von Groß- und Kleinschreibung	Im Editor von MAXDictio zur Bestimmung der Worthäufigkeiten war es möglich, zwischen Groß- und Kleinschreibung zu unterscheiden. Darauf wurde zunächst verzichtet, um Wortdopplungen (z.B. *Ich*/*ich*, *Spiel*/*spiel*, *Was*/*was*) zu vermeiden. Die Großschreibung wurde nachträglich für die verwendeten Nomen angepasst.
Zuordnung der Wortarten (POS-Tagging)	Eine automatische Zuordnung der Lemma-Types zu den entsprechenden Wortarten ist über MAXDictio nicht möglich. Die Zuordnung erfolgte manuell. Zur Festlegung der Wortart wurde die Kontextanalyse in MAXDictio als Entscheidungskriterium genutzt und über »Duden online« sowie Duden (2009) verifiziert. Als Wortarten wurden folgende Lexemklassen annotiert: Verb, Substantiv, Adjektiv, Pronomen, Adverb, Partikel, Präposition. Die Interjektion wurde als spezifische Form der Partikel (Ausdruckspartikel, Duden, 2009, S. 597) zusätzlich annotiert. Die Eigennamen wurden in Anlehnung an Duden (2009) den Substantiven zugeordnet (S. 1249). In der Zusammenführung der Einzellisten zu einer Gesamtliste wurde deutlich, dass einigen Wörtern mehrere Wortarten je nach Verwendungsweise pro Fall zugeordnet wurden. In dieser Situation wurde die Wortartenannotierung an die Zuordnung der Referenzliste angelehnt. Ausnahme bildeten folgende Wörter und Wortartenzuordnungen: – *wie*: Pronomen (Primärkorpus); Konjunktion (Referenzkorpus) – *wo*: Pronomen (Primärkorpus); Konjunktion (Referenzkorpus) – *gleich*: Adjektiv (Primärkorpus); Präposition (Referenzkorpus) – *Morgen*: Substantiv (Primärkorpus); Adverb (Referenzkorpus) – *der*, *die*, *das*: Artikel (Primärkorpus); Pronomen (Referenzkorpus)
Bereinigung der Einzellisten	Die gewonnenen Wortschatzlisten pro Fall wurden qualitativ überprüft, um Abweichungen vom Transkriptionssystem bereinigen zu können.

Pro Stichprobenteilnehmer:in konnte unter Beachtung des skizzierten Verfahrens eine Kernvokabularliste gewonnen werden. Das Kernvokabular wurde aus den 22 Wortlisten mit der 80 %-Marke nach Boenisch (2014b) herausgefiltert. Für die Erstellung einer Gesamtliste wurden die Einzellisten in Excel über die Funktion S-Verweis zusammengeführt und das Kernvokabular auf der Grundlage der 80 %-Marke ermittelt. Über eine zusätzliche Spalte konnte das Streuungskriterium ≥50 % einbezogen und reflektiert werden.

Feste Wortkombinationen

Vorbereitend wurde für die Auswertung der festen Wortkombinationen sichergestellt, dass pro Untersuchungsteilnehmer:in maximal ein Transkript vorlag. Wurden zwei Transkripte zum jeweiligen Messzeitpunkt erstellt, wurden die Transkripte zu einer Datei zusammengeführt. Somit konnten 22 Transkripte im Hinblick auf Dreiwortkombinationen analysiert werden. Für die Gewinnung einer Gesamtliste über die genutzten

Wortkombinationen ($Liste_{WK}$) wurden in MAXDictio alle Transkripte der Untersuchungsgruppe markiert und über die Funktion »Wortkombinationen« ausgelesen (Abb. 20).

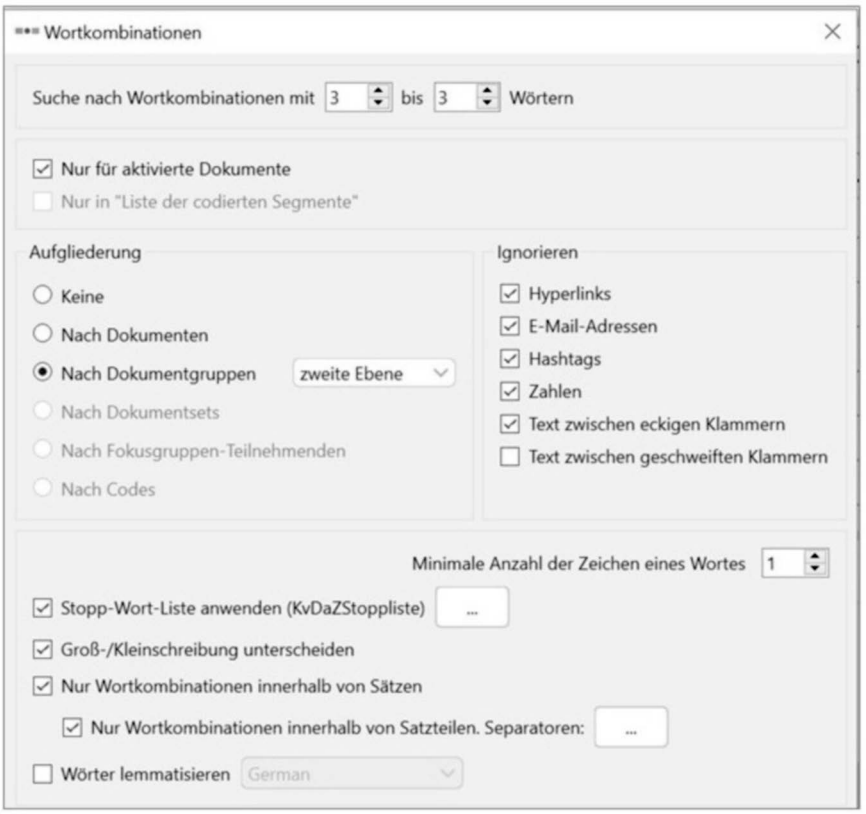

Abb. 20: Funktion »Wortkombinationen« in MaxDictio

In der benutzerdefinierbaren Einstellung wurde die Beachtung der Groß- und Kleinschreibung ausgewählt. Darüber hinaus wurde der zu untersuchende Kontext durch Satzgrenzen und Separatoren (z.B. (), ;, :) begrenzt. Die in MaxDictio erhobene Liste über alle Dreiwortkombinationen wurde in Excel exportiert und manuell nachbearbeitet. Über die Excel-Funktion »Konsolidieren« wurden alle mehrfach aufgeführten Dreiwortkombinationen mit den entsprechenden Häufigkeiten zusammengerechnet. Bei der Konsolidierung wurden Dreiwortkombinationen mit Groß- und Kleinschreibung zusammengefasst und nach dem häufigsten Gebrauch ausgegeben (z.B. *alles gut alles*; *Alles gut alles* > *alles gut alles*, *Wie heißt du*; *wie heißt du* > *Wie heißt du*). Die gewonnene $Liste_{Wk}$ wurde nach Häufigkeiten absteigend und alphabetisch sortiert. Danach erfolgte die Rangberechnung.

Über die Analyse der Häufigkeit-Rang-Beziehung und dem Verlauf der Zipf-Kurve wurde ermittelt, welche Wortkombinationen am häufigsten verwendet wurden und als fest eingestuft werden konnten (heuristischer Zugang). Als cut-off point wurde das rapide Abflachen der Verlaufskurve sowie die Rangverteilung verwendet, um einen möglichen Übergang von festen Wortkombinationen hin zu flexibel gebildeten Wortkombinationen zu identifizieren. Für die gewonnenen und als ganzheitlich eingestuften Dreiwortkombinationen wurde die Streuung berechnet (Wie viele Kinder nutzten die Dreiwortkombination?). Dafür wurde im Editor von MaxDictio zur Berechnung der Wortkombinationen die *Aufgliederung nach Dokumenten* ausgewählt, sodass nachvollzogen werden konnte, in welchem Transkript die Wortkombination genutzt wurde. Auf dieser Grundlage konnte die Anzahl der Streuung ermittelt werden.

Vergleichsanalyse

Bei der Analyse von Lernersprache ist die Hinzunahme eines Referenzkorpus von Personen mit Deutsch als Erstsprache erforderlich (Kap. 10.1). Daher wurde der Referenzdatensatz »Grundschule« aus dem Datensatz von Boenisch (2014b; 2013) herangezogen (Referenzkorpus). Das zur Verfügung gestellte Material bestand aus einem Worddokument mit Transkriptionen von $N = 28$ Schüler:innen sowie einer Excel-Liste mit dem aufbereiteten Wortschatz (Lemma-Types) nach Häufigkeit, Frequenz und Rängen. Die Kernvokabularliste aus dem Primärkorpus wurde mit dem Referenzkorpus über die Funktion S-Verweis in Excel verglichen. Gemeinsamkeiten und Unterschiede im Kern- und Randvokabular konnten auf diesem Weg deskriptiv herausgearbeitet werden. In weiteren Analysen wurden die am häufigsten verwendeten Wörter (TOP-Wörter) verglichen sowie die Wortartenverteilungen gegenübergestellt.

Damit der Referenzdatensatz auch für die Vergleichsanalyse der festen Wortkombinationen verwendet werden konnte, wurden die Transkriptionen nach dem oben beschriebenen Schema in MaxDictio ausgewertet. In der Vergleichsanalyse wurden die festen Wortkombinationen aus Primär- und Referenzkorpus in Excel gegenübergestellt und Gemeinsamkeiten und Unterschiede herausgearbeitet.

11.5 Prüfung der Reliabilität, Objektivität und Validität

Die Interrater-Reliabilität für die Gewinnung der Kernvokabularliste wurde über den Übereinstimmungskoeffizienten (Anzahl der übereinstimmenden Fälle geteilt durch Gesamtheit der analysierten Fälle) berechnet (Hirschmann, 2019, S. 98). Dafür wurden 10 % ($n = 2$) der Sprachaufnahmen aus dem Datensatz ($N = 22$) randomisiert ausgewählt, erneut transkribiert und als Wortlisten aufbereitet. Der Übereinstimmungskoeffizient der Wortschatzlisten entsprach bei den Token $rü = 0.99$ (99 %) und bei den Lemma-Types $rü = 1$ (100 %). Damit ließ sich die Reliabilität in Bezug auf die formalsprachlichen Parameter als sehr hoch einstufen. Die Reliabilitätsprüfung sagt jedoch noch nichts über die Übereinstimmung auf Wortebene (qualitative Ebene) aus. Um die Reliabilität auf der qualitativen Ebene einschätzen zu können, wurde die Abweichung der Transkripte im Hinblick