Erforschen testen – Tests erforschen

Mag. Benjamin Kremmel
Institut für Fachdidaktik, Bereich Didaktik der Sprachen
Universität Innsbruck

Agnes Frick
Studierende
Universität Innsbruck

Sandra Parhammer
Studierende
Universität Innsbruck

Stefanie Lutz
Studierende
Universität Innsbruck

1. Einleitung

Fremdsprachenlehrpersonen sind angewandte SprachwissenschaftlerInnen oder sollten es, wie Widdowson (1991) einmahnt, zumindest sein. Im Zuge der fachdidaktischen Ausbildung scheint es demnach wichtig, angehende SprachlehrerInnen mit den Methoden und Erkenntnissen der Spracherwerbs-, Sprachlehr- und Sprachtestforschung vertraut zu machen (Bartels 2005). Unter dem Leitsatz des forschenden Lernens (Schratz & Weiser 2002) jedoch genügt es kaum, dies auf theoretischer Ebene abzuhandeln, sondern verlangt vielmehr danach, Studierende selbst forschend tätig werden zu lassen, um ihnen für spätere, eigenständige Forschungsprojekte im Rahmen von Abschlussarbeiten oder Aktionsforschungs-

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik“ 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.
 ISBN 978-3-903122-51-2, DOI 10.15203/3122-51-2

vorhaben als praktizierende LehrerInnen im Klassenzimmer Instrumente und Erfahrungswerte in die Hand zu geben.

Dieser Beitrag berichtet vom Versuch, im Zuge einer universitären Lehrveranstaltung mit Englisch-Lehramtsstudierenden als Teil der Kursanforderungen ein kurzes Forschungsprojekt durchzuführen, bei dem die Aussagekraft verschiedener Vokabeltests erforscht werden sollte. Das Projekt hatte das Ziel, Studierende schrittweise in die angewandte Spracherwerbsforschung einzuführen und ihnen einen Einblick in grundlegende Datensammlung, -analyse und -interpretation zu ermöglichen, um sie sowohl für die kritische Interpretation von Testresultaten in ihrer zukünftigen Tätigkeit als auch für weitere Forschungsvorhaben zu rüsten.

Der Beitrag wird zunächst das Modell der Lehrveranstaltung und die Anforderungen an die Studierenden bzgl. der Forschungsarbeit darstellen, bevor die Ergebnisse der studentischen Forschungsarbeit präsentiert werden. Abschließend werden ausgewählte Erkenntnisse der umfassenden Evaluierung des Projekts vorgestellt, die das Meinungsbild der Studierenden zu dieser Bewertungsgrundlage erheben und Verbesserungsvorschläge ihrerseits erfassen. Der Beitrag wird aufzeigen, dass die Umsetzung eines derartigen Ansatzes mit der Involvierung von Studierenden in ein kurzes Forschungsprojekt durchaus zu fördern wäre, da es neben positiven Auswirkungen auf Studierende zusätzlich auch interessante Antworten auf innovative Forschungsfragen der Sprachtestforschung generieren kann.

2. Theoretische Perspektiven

Um LehrerInnen als ForscherInnen auszubilden und sie mit forschendem Lernen nicht nur als theoretischem Modell vertraut zu machen, sondern ihnen eigene Erfahrungen diesbezüglich mitzugeben, wird schon seit einigen Jahren gefordert „in der universitären Lehrerbildung deutlichere Akzente auf die Förderung forschungsmethodischer Kompetenzen zu legen“ (Ammann & Ostendorf 2007). Dazu ist es nötig, Situationen zu schaffen, in denen StudentInnen selbst forschen, Fragen stellen und sich auf den Weg nach Antworten begeben können (Schratz & Weiser 2002). Dies sollte idealerweise „in einem klaren organisatorischen Rah-

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik“ 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

men geschehen, in dem Lernende - in diesem Fall Studierende - selbstständig und selbsttätig an komplexen und interessanten Lernaufgaben“ (Schratz & Weiser 2002: 41) arbeiten und mit starkem Praxisbezug Erfahrungen im Umgang mit solchen Lernsituationen sammeln können (Bosse 2012). Im deutschsprachigen Raum noch vielfach ungenutzt, ist diese Form von Lehren und Lernen gerade in den USA schon einigermaßen etabliert. Als „Pädagogik des 21. Jahrhunderts“ (NCUR 2005, zitiert in McKayle 2011) angepriesen, wird hier unter dem Schlagwort *undergraduate research* vielerorts bereits ein Konzept implementiert, dessen Ziel es ist, Studierenden die Möglichkeit zu geben, akademisch zu wachsen und sie von der traditionell konzipierten Basis der Bloom'schen Lernzielpyramide (Bloom 1956) hin zur Spitze einer neuen Hierarchie (Anderson et al. 2001) zu begleiten, deren Ziel es nicht nur ist, Wissen zu verstehen, sondern darüber hinaus auch selbst neues Wissen zu schaffen (McKayle 2011). Erste Evaluierungen zeigen, dass diese Art des universitären Lernens zusätzliche Motivation schafft (Fechheimer et al. 2011) und auch dazu dient, Lerninhalte besser und nachhaltiger zu verinnerlichen (Nagda et al. 1988).

3. Praktische Umsetzung

Vor dem Hintergrund dieses theoretischen Konzepts wurde daher eine Implementierung in einer universitären Fachdidaktiklehrveranstaltung angedacht, die sich mit dem Lernen, Lehren und Überprüfen von Wortschatz im Englischunterricht befasste. Die einstündige Lehrveranstaltung war stark forschungsorientiert und konzentrierte sich auf die Diskussion von Forschungserkenntnissen und deren Relevanz bzw. deren Implikationen für den Fremdsprachenunterricht. Ziel der Lehrveranstaltung war es, angehende Englischlehrpersonen mit angewandter fachdidaktischer Spracherwerbsforschung vertraut zu machen und für eine kritische Evaluierung von Forschungsergebnissen zu sensibilisieren. Um diese Lerninhalte zu transportieren wurde als Kursanforderung daher ein Mikro-Forschungsprojekt von den Studierenden durchgeführt, das im Folgenden näher erläutert wird. Obwohl das Prinzip der *undergraduate research* im Idealfall mehr kreativen Spielraum für Studierende und unter anderem deren Involvierung bereits in

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik“ 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

der Findung der Forschungsfragen sowie der Konzeption und Planung des Forschungsvorhabens vorsieht oder vorsehen kann, wurde aus zeitlichen Gründen das Forschungsprojekt von der Lehrveranstaltungsleitung bereits vorab skizziert und die zu untersuchenden Forschungsfragen wurden folglich vorgegeben. Aus Gründen der Praktikabilität wurde dabei auf Fragen der Sprachtestforschung zurückgegriffen, da die Studierenden als angehende Lehrpersonen mit der Durchführung, Evaluierung und besonders der Interpretation von Wortschatztests bzw. Sprachtests im Allgemeinen im späteren Lehrberuf alltäglich konfrontiert sein werden und diese Fragen daher als sehr relevant für das spätere Berufsfeld der Studierenden eingestuft wurden.

4. Das Forschungsprojekt

Ziel des studentischen Forschungsprojekts war es, den *Test of Multi-Word Expressions* (Martinez 2011), der das phraseologische Wortschatzwissen von LernerInnen messen soll, auf seine Validität zu untersuchen. Unter detaillierten Anleitungen hatten die Studierenden des Kurses die Aufgabe, Daten von jeweils drei EnglischlernerInnen zu sammeln und anhand dieses Datensatzes folgende drei Forschungsfragen zu beantworten:

(1) Wie verhält sich das allgemeine **Wortschatzwissen** von LernerInnen im Sinne individueller Wörter im Vergleich zu deren phraseologischer Wortschatzkenntnis?
(2) Wie gut repräsentieren jeweils die 10 Items aus den ersten beiden Häufigkeitsniveaus die phraseologische Wortschatzkenntnis von KandidatInnen in diesen zwei Niveaus?
(3) Wie gut repräsentieren diese individuellen Testitems die eigentliche Wortschatzkenntnis, die KandidatInnen über diese Items haben?

Um diese Fragen zu untersuchen, erhoben die Studierenden von jedem/jeder TeilnehmerIn sowohl einen *Vocabulary Size Test* (Nation & Beglar 2007) als all-

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik“ 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

gemeines Wortschatzmessinstrument als auch einen *Test of Multi-Word Expressions* (*TMWE*) als phraseologisches Messinstrument. Im Anschluss daran führten die Studierenden individuell Interviews mit den TeilnehmerInnen durch, die die phraseologische Wortschatzkenntnis näher erheben sollten. Im Folgenden werden die Messinstrumente beschrieben.

4.1 Forschungsinstrumente

4.1.1 Vocabulary Size Test

Der *Vocabulary Size Test* (*VST*) ist ein Test, der die Größe des schriftlich rezeptiven Wortschatzes misst. Der Test besteht im Original aus insgesamt 140 Testitems, bei denen KandidatInnen die richtige Definition oder das richtige Synonym eines individuellen Wortes aus vier Optionen auswählen müssen. Die Testitems basieren dabei auf einer Häufigkeitsliste des *British National Corpus* (Nation 2004), die die häufigsten Wortfamilien des Englischen in eine Rangordnung bringt. Der Test ist in 14 Häufigkeitsniveaus (1K-14K) unterteilt, wobei aus jedem 1000er Niveau je 10 Items präsentiert werden. Jedes einzelne Item repräsentiert daher 100 Wortfamilien, was impliziert, dass die Anzahl der richtigen Antworten im Test mit 100 multipliziert werden kann, um die Wortschatzgröße der TeilnehmerInnen zu errechnen (Nation & Beglar 2007). Ein Beispielitem des Tests ist in Abb. 1 angeführt.

***STONE**: He sat on a stone.*

a. hard thing

b. kind of chair

c. soft thing on the floor

d. part of a tree

Abb. 1: Beispielitem aus dem Häufigkeitsniveau **2K** des **Vocabulary Size Tests** (Nation & Beglar 2007)

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik“ 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

Wie anhand des Beispiels ersichtlich, handelt es sich beim *VST* um ein *Multiple-Choice* Testformat. Jedes Item besteht aus einem Wort, das in einen nicht-definierenden Kontext eingebettet ist und für welches das korrekte Synonym oder die korrekte Beschreibung gefunden werden muss. Der Kontext verrät lediglich, um welche Wortart es sich handelt, die Bedeutung des Wortes kann aber nicht aus dem Kontext erschlossen werden. Dadurch soll sichergestellt werden, dass tatsächlich das Wortschatzwissen der TeilnehmerInnen und nicht etwa die Deduktionsfähigkeit der KandidatInnen überprüft wird. Da es sich um einen *Multiple-Choice* Test handelt, bei dem die Antwortmöglichkeiten bereits vorgegeben sind, überprüft dieser Test die rezeptiven Wortschatzkenntnisse der TeilnehmerInnen und zeigt nicht auf, ob die KandidatInnen die Wörter auch tatsächlich in ihrem Sprachgebrauch verwenden können. Ziel des *Vocabulary Size Tests* ist es, den Vokabelumfang von LernerInnen in den ersten 14000 häufigsten Wortfamilien des Englischen zu messen. In Anbetracht der Forschungsergebnisse aus Korpusstudien, die die lexikalischen Anforderungen verschiedener sprachlicher Aktivitäten untersuchten (Adolphs & Schmitt 2003; Nation 2006; Schmitt & Schmitt 2014; Webb & Rodgers 2009; van Zeeland & Schmitt 2013), kann aufgrund der Ergebnisse ein indirekter Rückschluss darauf gezogen werden, welche Aufgaben Lernende aufgrund ihres Wortschatzwissens ausführen beziehungsweise nicht ausführen können. Für den Zweck dieser Studie wurden nur die ersten vier Häufigkeitsniveaus, das heißt 40 Items zu den ersten 4000 Wortfamilien (*1K-4K*), verwendet, um zu vergleichen, wie groß das Wortschatzwissen individueller Wörter gegenüber jenem von Phrasen, bestehend aus mehreren Wörtern, die eine Bedeutungseinheit bilden, bei LernerInnen ist.

4.1.2 Test of Multi-Word Expressions

Um die phraseologische Wortschatzkenntnis zu erheben, wurde der *Test of Multi-Word Expressions (TMWE)* (Martinez 2011) durchgeführt. Dieser Test umfasst fünf Häufigkeitsniveaus (*1K-5K*) und prüft jeweils 10 Items pro Niveau. Die Testitems des *TMWE* wurden aus Martinez & Schmitts (2012) *Phrasal Expressions List* entnommen, einer Liste der 505 häufigsten *Multi-Word Expressions* des Englischen. Da die Liste mithilfe desselben Korpus generiert wurde wie der *VST*,

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik“ 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

ist die Einteilung der Phrasen in die Häufigkeitsniveaus mit jener, die dem *VST* zugrunde liegt, ident. Dies bedeutet jedoch, dass nicht jedes Häufigkeitsniveau gleich viele Phrasen enthält. Im ersten Häufigkeitsniveau finden sich beispielsweise nur 32 Phrasen. Im zweiten Niveau hingegen scheinen aufgrund ihrer relativen Häufigkeit 84 Phrasen auf. Dieses Ungleichgewicht kann potentiell problematisch für das *item sampling* sein.

Im Gegensatz zum *VST*, bei welchem die TeilnehmerInnen ein korrektes Synonym oder eine korrekte Erklärung für ein spezifisches Wort auszuwählen haben, müssen die KandidatInnen im *TMWE* ein Synonym oder eine Erklärung für eine lexikalische Einheit bzw. Phrase auswählen. Die sogenannten *Multi-Word Expressions* bestehen aus mehreren Wörtern oder Phrasen, die in Kombination miteinander eine andere Bedeutung annehmen als lediglich die Summe der wortwörtlichen Übersetzungen der einzelnen Bestandteile der Phrase. Der Test legt das Augenmerk auf lexikalische Einheiten beziehungsweise Phrasen, da diese einen großen Teil der gesprochenen Sprache ausmachen und somit für eine effektive Kommunikation unentbehrlich sind (Martinez & Schmitt 2012). Phraseologisches Wissen, gemessen mit diesem Instrument, zeigt sich außerdem als zentraler Bestandteil erfolgreichen Leseverstehens (Kremmel 2012). Das Instrument wurde bislang jedoch keiner extensiven Validierung unterzogen. Dies ist vielmehr Ansatzpunkt und genuiner Beitrag der hier vorgestellten Studie.

Beim Testformat handelt es sich, wie auch beim *VST,* um einen *Multiple-Choice* Test mit je vier Antwortmöglichkeiten, wovon nur eine korrekt ist. Im Rahmen des Tests muss für jedes Item die korrekte Erklärung respektive ein korrektes Synonym ausgewählt werden. Die Items sind wie am Beispiel erkennbar (Abb. 2) in einen sie nicht-definierenden Kontext eingebettet und jede der vier Antwortmöglichkeiten ergibt im Hinblick auf den Beispielsatz Sinn. Deshalb kann die Bedeutung der Phrasen nicht erraten werden. Da in diesem Test wieder die rezeptiven Fähigkeiten der TestteilnehmerInnen in Bezug auf lexikalische Einheiten abgeprüft werden, kann nicht darauf geschlossen werden, ob die KandidatInnen die betreffenden Phrasen wirklich in ihrem aktiven Sprachgebrauch verwenden. Eine Beispielfrage ist in Abb. 2 dargestellt.

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik“ 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

to do with: *It is to do with money.*

a. making

b. for

c. about

d. our

Abb. 2: Beispielitem aus dem Häufigkeitsniveau 2K des
Test of Multi-Word Expressions (Martinez 2011)

Für den Zweck dieser Studie wurde auch der *TMWE* verkürzt. Anstatt der 50 Items des ursprünglichen Tests wurden – wie auch beim *VST* – nur die 40 Items der ersten vier Häufigkeitsniveaus verwendet, um eine Vergleichbarkeit der beiden Tests zu ermöglichen.

Abschließend kann hier festgestellt werden, dass sich der *VST* (Nation & Beglar 2007) und der *TMWE* (Martinez 2011) in formaler Hinsicht auf den ersten Blick sehr ähnlich sind. Beide Tests sind schriftliche *Multiple-Choice* Tests, die rezeptives Sprachwissen in Bezug auf Wortschatz testen. Die Items sind in beiden Tests in einen sie nicht-definierenden Kontext eingebettet, folgen jedoch einem anderen Testkonstrukt, da im *VST* Vokabelwissen als Bedeutungswissen einzelner Wörter konzipiert ist und im *TMWE* Vokabelwissen aus dem Bedeutungswissen längerer lexikalischer Einheiten besteht.

4.1.3 Das Interview – Phrasal Expressions List

Als Kriterium, um zu überprüfen, wie gut ein Wortschatztest die eigentliche Wortschatzkenntnis von KandidatInnen repräsentiert, ist ein Interview, das nach dem Abrufen einer Wortbedeutung ohne zusätzliche Hilfe oder Auswahloptionen fragt, ein erprobtes Mittel (Nation & Webb 2011; Paul et al. 1990; Schmitt et al. 2001; Schmitt 2010). In der vorliegenden Studie wurden die Forschungsfragen 2 und 3 daher anhand eines Abgleichs der Testresultate mit den Antworten der KandidatInnen aus einem solchen Interview beantwortet. Für das

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik“ 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

Interview wurden 2 Häufigkeitsniveaus der *Phrasal Expressions List* von Martinez & Schmitt (2012) ausgewählt, von welcher der *TMWE* seine Items bezieht. Das *1K* Niveau der Liste besteht dabei insgesamt aus 32 Items, das *2K*-Niveau umfasst 84 Phrasen. Die TestteilnehmerInnen wurden nach Beantwortung der beiden oben beschriebenen Tests aufgefordert, im Dialog mit dem/der InterviewerIn die Bedeutung dieser 116 Items zu erklären. Auf diese Weise konnte nicht nur überprüft werden, wie adäquat der *TMWE* (bzw. die jeweiligen 10 Testitems) das phraseologische Wissen des jeweiligen Häufigkeitsniveaus als Ganzes widerspiegeln, sondern auch, ob die Messergebnisse des *TMWE* in diesen 10 Testitems eine angemessene Repräsentation des eigentlichen Wortschatzwissens dieser getesteten Phrasen darstellt. So konnte die Validität von auf *TMWE*-Testergebnissen basierenden Aussagen über das Wortschatzwissen der KandidatInnen untersucht werden, was angesichts der gut dokumentierten Neigung zu Rateverhalten bei KandidatInnen in der Beantwortung des *Multiple-Choice* Formats zentral ist (Gyllstad et al. 2015; Kremmel & Schmitt i.V.; Stewart 2014). Durch das Interview konnte also analysiert werden, ob richtige Antworten im Test wirklich auf Wortschatzwissen oder eher auf erfolgreiches Erraten einer der vier Optionen zurückzuführen sind.

Im Interview wurden den KandidatInnen die Items der *PHRASE List* inklusive des dort angegebenen Beispielsatzes vorgelegt. Die KandidatInnen waren dann aufgefordert, die Phrasen mündlich zu übersetzen und/oder zu erklären, entweder auf Englisch oder in ihrer Muttersprache. Das Interview wurde von den Studierenden digital aufgezeichnet, um im Auswertungsprozess bei Unklarheiten noch einmal auf die digitalen Audio-Dateien zurückgreifen zu können.

4.2 Untersuchungsdesign

Die Untersuchung wurde wie folgt durchgeführt: Jede/r Studierende rekrutierte drei Freiwillige, die über ein Minimum an Englischkenntnissen verfügen mussten. Den TeilnehmerInnen wurde von den Studierenden der Studienablauf erklärt. Sie wurden gebeten, eine Einverständniserklärung über ihre Teilnahme an der Studie zu unterzeichnen. Im Anschluss wurde zunächst der *VST* und daran anschließend der *TMWE*, jeweils in der oben beschriebenen gekürzten

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik“ 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

Variante mit je 40 Items, durchgeführt. Dann folgte das Einzelinterview, das in der Regel nicht länger als 40 Minuten dauerte und das Wissen in Bezug auf die ersten beiden Häufigkeitsniveaus der *PHRASE* List verifizierte. Während des Interviews entschieden die Studierenden für jede bearbeitete Phrase, ob der/die KandidatIn die Bedeutung der Phrase kannte und hielten diese im Bewertungsbogen fest. Außerdem wurden in einem kurzen Fragebogen am Ende demografische Daten sowie Angaben zu Lerndauer, Zeitpunkt des letzten schulischen englischen Sprachunterrichts, Auslandsaufenthalten und der Regelmäßigkeit der englischen Sprachverwendung erhoben. Diese Angaben, Testantworten und Entscheidungen aus dem Interview wurden dann von den Studierenden in eine vorprogrammierte EXCEL 2010-Vorlage zur Auswertung eingegeben.

4.3 ProbandInnen

Während die Studierenden für ihre Seminararbeit lediglich den selbst gesammelten Datensatz von drei TeilnehmerInnen auswerten und beschreiben mussten, wurden für die hier vorgelegte Analyse alle Daten der KursteilnehmerInnen zusammengeführt und analysiert. Insgesamt nahmen 48 KandidatInnen an der Studie teil (31 weiblich, 17 männlich). Das Durchschnittsalter der Befragten lag bei 25,6 Jahren (Standardabweichung = 9). Alle TeilnehmerInnen gaben Deutsch als ihre L1 an. 85% der KandidatInnen gaben an, dass sie die österreichische Reifeprüfung abgelegt hatten, was bezüglich ihrer Sprachkompetenz in Englisch auf ein Mindestniveau von B2 laut dem *Gemeinsamen Europäischen Referenzrahmen für Sprachen* (*GERS*) (Trim, North & Coste 2001) schließen lassen könnte. Da in vielen Fällen allerdings diese Reifeprüfung bereits einige Jahre zurücklag und damit in die Zeit vor der Einführung der GERS-basierten standardisierten Reifeprüfung fällt, ist diese Annahme lediglich tentativ. Im Durchschnitt hatten die KandidatInnen zehn Jahre Englisch gelernt. Elf KandidatInnen hatten einen längeren Auslandsaufenthalt (Durchschnitt 2,2 Monate) im englischsprachigen Raum hinter sich. Durchschnittlich lag die letzte (schulische) Englischklasse der TeilnehmerInnen zum Zeitpunkt der Testung ungefähr 5 Jahre (63 Monate) zurück.

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik“ 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

4.4 Ergebnisse[1]

4.4.1 Forschungsfrage 1: Vergleich Ergebnisse *VST* vs. *TMWE*

Die erste Forschungsfrage der Studie verweist darauf, inwiefern es Unterschiede zwischen den Ergebnissen des *VST* und jenen des *TMWE* gibt, die auf eine Divergenz zwischen dem Wortschatzwissen individueller Wörter und dem phraseologischer Elemente hinweisen. Die Hypothese besagt, dass, Martinez & Schmitt (2012) folgend, die Befragten ein besseres Ergebnis im *VST* erzielen würden, da in der schulischen Wortschatzvermittlung der Fokus vor allem auf Einzelwörter und nicht auf lexikalische Einheiten gelegt wird. Die Durchschnittswerte der beiden Tests bestätigen diese Annahme jedoch nicht. Wie in Abb. 3 ersichtlich, lag der Mittelwert für den *VST* bei 31,71 und für den *TMWE* bei 31,06 (jeweils von 40).

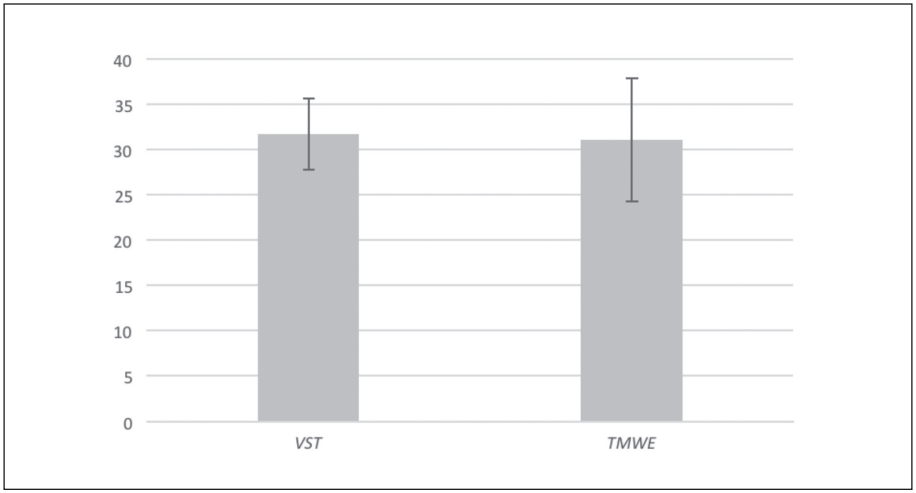

Abb. 3: Vergleich Durchschnittswerte **VST** und **TMWE**

1 Die im Folgenden vorgelegte Analyse wurde von Studierenden ohne jegliches statistisches Vorwissen durchgeführt. Es wurde für diese Publikation darauf verzichtet, diese zu erweitern, um sowohl darzustellen, zu welchen Analysen selbst unerfahrene Studierende in der Lage sind, als auch was im Zuge der Lehrveranstaltung erwartet wurde.

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik“ 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

Obwohl die Gruppe der getesteten Personen trotz ihrer relativen Heterogenität über ein ausgeglichenes Wortschatzwissen in Bezug auf Einzelwörter und hinsichtlich lexikalischer Einheiten verfügt und die Mittelwerte in beiden Tests sich nicht signifikant unterscheiden, ist in der Standardabweichung ersichtlich, dass die Resultate im *TMWE* breiter gestreut sind. Zieht man die Ergebnisse einzelner TestteilnehmerInnen heran, zeigt sich des Weiteren, dass die Resultate einzelner Personen in den beiden Tests zum Teil deutlich divergieren: So erzielten 42% der TeilnehmerInnen bessere Resultate im *VST*, 44% erzielten bessere Resultate im *TMWE*. Insgesamt ist also kein klarer Trend gegeben, der die angestellte Hypothese unterstützt.

Die Ausgeglichenheit in Bezug auf die Testergebnisse in beiden Tests könnte in erster Linie aufgrund der Heterogenität der TeilnehmerInnengruppe erklärt werden. Ein Grund für eine höhere Anzahl an richtigen Antworten im *VST* als im *TMWE* könnte sein, dass lexikalische Einheiten zumeist unterschätzt werden und auch im Sprachunterricht noch nicht ausreichend bzw. systematisch Einzug gefunden haben (Martinez & Schmitt 2012: 299).

Außerdem fällt auf, dass viele Testpersonen, die bessere *VST* Resultate aufwiesen, nach beziehungsweise außerhalb ihres Englischunterrichts in der Schule wenig bis gar nicht mit der Sprache in Berührung kamen. So gaben 60% der Personen mit einer höheren Punktezahl im *VST* als im *TMWE* an ein- bis zweimal im Monat oder seltener Englisch zu verwenden. Im Gegensatz dazu gaben 85% der KandidatInnen, welche ein besseres Ergebnis im *TMWE* erzielen konnten, an, täglich oder ein- bis zweimal pro Woche Englisch zu verwenden. Auch erzielten Personen, bei denen die letzte Sprachverwendung weniger lange zurücklag, tendenziell bessere Resultate im *TMWE*. Bei 86% der Personen mit besseren Resultaten im *TMWE* als im *VST* lag die letzte Sprachverwendung nicht länger als eine Woche zurück. Nur bei 65% der Personen mit besseren Resultaten im *VST* als im *TMWE* war dies ebenso der Fall. Dies könnte darauf hindeuten, dass phraseologische Wortschatzkenntnis mit einer stärkeren Einbettung in den englischen Sprachgebrauch einhergeht und dass ohne diese Einbettung das Erlernen bzw. Behalten von Phrasen schwieriger ist als jenes von Einzelwörtern.

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik“ 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

4.4.2 Forschungsfrage 2: Repräsentativität der 10 Testitems für das jeweilige Häufigkeitsniveau

Die zweite Forschungsfrage untersuchte, inwiefern sich die Ergebnisse des *TMWE* von jenen des Interviews unterscheiden. Dadurch sollte evaluiert werden, ob die limitierte Anzahl an Test*items* des *TMWE* repräsentativ das Phrasenwissen der jeweiligen Häufigkeitsniveaus abbildet und damit das *Sampling* des *TMWE* zuverlässig ist. Wie bereits erläutert, werden beim *TMWE* je 10 Items für das *1K* wie für das *2K* Niveau abgefragt. Die dem Test zugrundeliegende *PHRASE List* beinhaltet jedoch 32 Items für das *1K* und 84 Items für das *2K* Niveau. Aufgrund dieser Unterschiede wurden von den Studierenden die prozentuellen Mittelwerte der Testresultate (10 Items) und der Interviewresultate (32 bzw. 84 Items) miteinander verglichen. Die Grafik in Abb. 4 zeigt, dass sich die Durchschnittswerte der beiden Messinstrumente sowohl auf dem *1K* als auch dem *2K* Niveau lediglich minimal unterscheiden. Im *TMWE* wurden im Durchschnitt 84% der *1K* Fragen richtig beantwortet, was das im Interview verifizierte Wortschatzwissen des gesamten *1K* Niveaus (86%) gut abzubilden scheint (vgl. Tab. 1). Das *2K* Niveau zeigt ähnliche Ergebnisse mit 80% richtigen Antworten im *TMWE* und 81% im Interview. Somit könnte die Annahme getroffen werden, dass das *item sampling* des *TMWE* zufriedenstellend repräsentativ ist.

Tab. 1: Vergleich Durchschnittswerte **TMWE** und **PHRASE List** (Interview)

	TMWE	***PHRASE List* (Interview)**
1K	84%	86%
2K	80%	81%

Eine detaillierte Analyse der individuellen Test- und Interviewergebnisse zeigt jedoch, dass in einigen Fällen Aussagen über das Phrasenwissen der einzelnen Niveaus basierend auf dem Testresultat nicht ausreichend aussagekräftig sind. Das Testergebnis, das jeweils auf nur 10 Items beruht, über- oder unterschätzt die phraseologische Wortschatzkenntnis eines gesamten Niveaus drastisch.

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik“ 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

Abb. 4 zeigt die Differenz zwischen Test- und Interviewergebnissen für das *1K* Niveau sortiert nach Größe der Differenz für die einzelnen KandidatInnen. Es zeigt sich, dass auf dem *1K* Niveau der Test für einzelne KandidatInnen das Wortschatzwissen für dieses gesamte Niveau bis zu 34% unterschätzt (KandidatIn 1) bzw. bis zu 28% überschätzt (KandidatIn 48). Auf dem *2K* Niveau zeichnet sich ein ähnliches Bild ab. Abb. 4 zeigt die Differenz zwischen Test- und Interviewergebnissen für das *2K* Niveau, wiederum sortiert nach Größe der Differenz für die einzelnen KandidatInnen (Die KandidatInnenreihung in Abb. 5 ist unabhängig von der Reihung in Abb. 4). Der Test führt auf diesem Niveau in einzelnen Fällen zu einer Unterschätzung von 33% bzw. einer Überschätzung von 21% des Wortschatzwissens des gesamten *2K* Niveaus. Dies stellt die oben getroffene Annahme des zufriedenstellenden *item samplings* zumindest teilweise infrage.

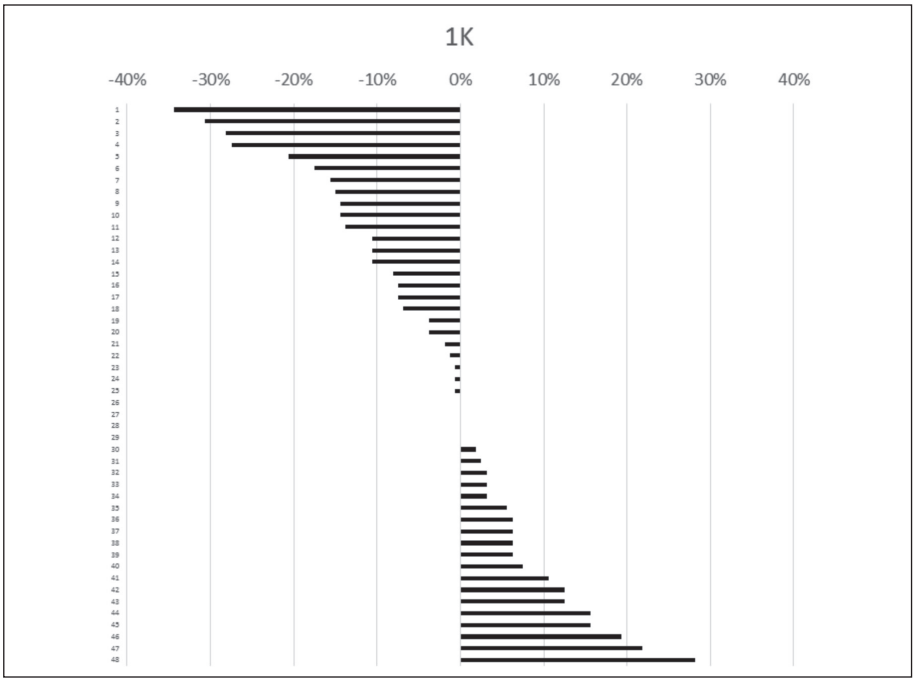

Abb. 4: Über-/Unterschätzung des Wortschatzwissens des gesamten **1K** Niveaus

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik“ 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

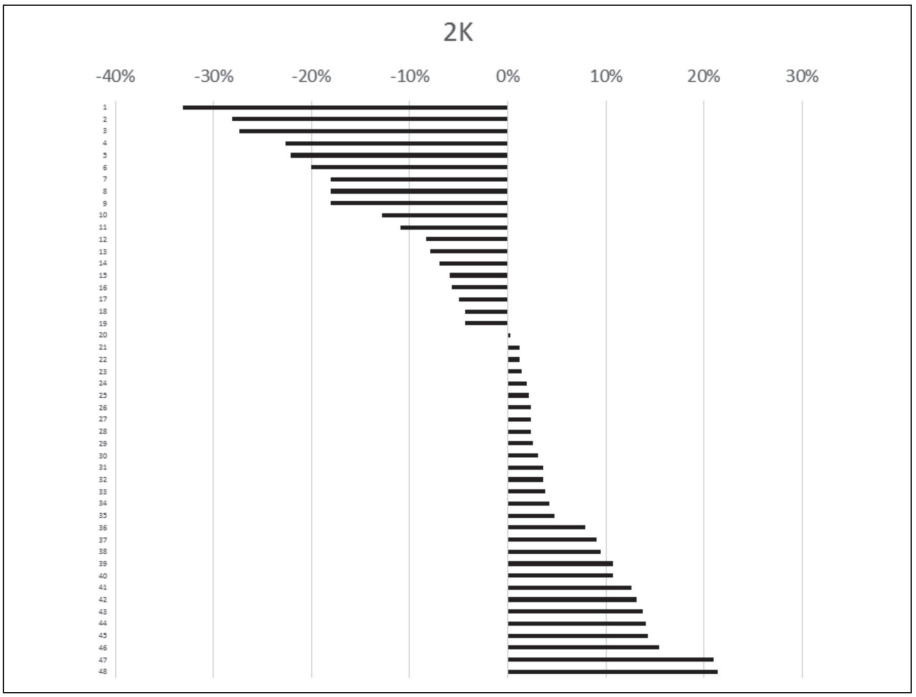

Abb. 5: Über-/Unterschätzung des Wortschatzwissens des gesamten **2K** Niveaus

4.4.3 Forschungsfrage 3: Repräsentativität der einzelnen Items

Auf der Basis von Forschungsfrage 2 untersuchten die Studierenden nicht nur die Repräsentativität der Testresultate für das Wissen aller Phrasen eines Häufigkeitsniveaus, sondern auch die Repräsentativität der Testresultate der individuellen Testitems für die respektiven Phrasen. Auf diese Weise sollte ermittelt werden, ob der *TMWE* das Wissen der KandidatInnen in den einzelnen Items überschätzt, unterschätzt oder ob die Interpretation zulässig ist, dass eine richtige Beantwortung eines Testitems auch wirklich auf ein zugrundeliegendes Wissen um die Bedeutung der getesteten Phrase zurückzuführen ist. Die Analyse dieser Forschungsfrage war besonders wichtig, denn geschlossene Testformate wie *Multiple-Choice* laufen immer Gefahr, das eigentliche sprachliche Wissen verzerrt

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik“ 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

darzustellen, da Faktoren wie Distraktorenqualität oder Ratewahrscheinlichkeit einen Einfluss auf das Testergebnis haben können. In der Analyse des Datensatzes für die Beantwortung dieser Frage sind dabei vier Szenarien möglich, die in Abb. 6 dargestellt sind.

		Testitem	
		richtig	**falsch**
Interview	**richtig**	Übereinstimmung (A)	Unterschätzung (B)
	falsch	Überschätzung (C)	Übereinstimmung (D)

Abb. 6: Übereinstimmung, Überschätzung oder Unterschätzung

Im Idealfall wäre zu erwarten, dass ein/e KandidatIn, der/die ein Testitem richtig beantwortet, auch in der Lage ist, die Bedeutung dieser Phrase im Interview zu erklären (A), sofern die Testantwort auf Phrasenwissen und nicht auf anderen Faktoren basiert. Im Umkehrschluss wäre auch zu hoffen, dass ein/e KandidatIn, der/die das Testitem falsch beantwortet, nicht in der Lage ist, die Bedeutung der getesteten Phrase im Interview abzurufen (D). Test- und Interviewergebnis sollten bei einem qualitativ hochwertigen Wortschatztest daher möglichst übereinstimmen. Erzielt ein/e KandidatIn im Test eine richtige Antwort, ohne Bedeutungswissen der Phrase nachweisen zu können, überschätzt der Test das eigentliche Wortschatzwissen des/der KandidatIn (C). Vice versa unterschätzt ein Test das Wortschatzwissen dann, wenn KandidatInnen trotz nachweislichen Phrasenwissens die entsprechenden Items im Test falsch beantworten (B).

Im Zuge der Studie wurde eine durchschnittliche Übereinstimmung von ca. 80% zwischen Testresultat und eigentlicher Wortschatzkenntnis der individuellen Items festgestellt, während eine jeweils 10%ige Über- respektive Unterschätzung gefunden wurde. Dieses Ergebnis ist einigermaßen überraschend, da andere Studien (Kremmel & Schmitt, i.V.) von einem starken Trend zur Überschätzung bei *Multiple-Choice* Formaten berichten. Insgesamt entspricht der Wert der Übereinstimmung jedoch bisherigen Forschungsergebnissen. Betrachtet man die Ergebnisse der einzelnen Personen, konnten 30 von 48 Getesteten auf dem

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik“ 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

1K Niveau zwischen 80% und 100% Übereinstimmung erzielen. Ein ähnliches Ergebnis zeigen die einzelnen Testpersonen bezogen auf das *2K* Niveau, wo 29 von 48 Befragten zwischen 80% und 100% Übereinstimmung erzielen konnten (vgl. Tab. 2).

Tab. 2: Analyse der individuellen Items: Fälle der Übereinstimmung, Unterschätzung und Überschätzung

	1K	***2K***
Übereinstimmung (A+D)	83%	79%
Unterschätzung (B)	8%	12%
Überschätzung (C)	9%	9%

Die Resultate weisen tentativ darauf hin, dass der *TMWE* als Messinstrument für Phrasenwissen einigermaßen gut funktioniert und bei vorsichtiger Interpretation überwiegend valide Schlüsse zulässt. Die von den Studierenden durchgeführte Studie zeigt jedoch trotz vielversprechender Ansätze, dass in Hinblick auf die Repräsentativität der Items sowohl für einzelne Häufigkeitsniveaus (FF2) als auch für das Bedeutungswissen individueller Phrasen (FF3) weitere rigorose Validierungsforschung zu diesem Test betrieben werden muss. Dies ist besonders angesichts der Limitationen dieser Studie zu unterstreichen. Obwohl die Stichprobe der vorliegenden Studie durch die organisatorischen Rahmenbedingungen mit nur 48 TeilnehmerInnen relativ gering war und die Interviews und Interviewurteile von 16 verschiedenen BeurteilerInnen vorgenommen wurden, was potentiell die Reliabilität der Ergebnisse gefährdet, lassen sich dennoch vorläufige Trends aus den Resultaten ablesen, die wertvolle Erkenntnisse für das Forschungsfeld aufzeigen.

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik“ 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

5. Evaluation des Forschungsprojekts: Rückmeldung der Studierenden

Neben der Generierung relevanter Forschungsergebnisse ist vor allem die studentische Wahrnehmung als Evaluationskriterium für die Art der Lehrveranstaltung in Erwägung zu ziehen. Um die Rückmeldungen der Studierenden zu diesem Pilotprojekt zu erfassen, wurde ein umfangreicher Feedbackfragebogen entwickelt, der neben allgemeinen Angaben zur Lehrveranstaltung, zu deren Inhalten sowie zur Lehrveranstaltungsleitung auch konkrete Kritik an der Durchführung der Forschungsarbeit im Zuge der Kursanforderungen erfasste. Dafür wurden elf Aussagen über die Lehrveranstaltung verschriftlicht, zu denen sich die LehrveranstaltungsteilnehmerInnen auf einer fünfstufigen Likert-Skala positionieren konnten. Die Skala verlief von „1 – stimme völlig zu", bis zu „5 – stimme überhaupt nicht zu". Im Folgenden (Tab. 3, 4, 5) sind die Mittelwerte dieser Befragung dargestellt. Je niedriger der angezeigte Mittelwert, desto mehr Zustimmung fand die jeweilige Aussage bei den Studierenden. Die Standardabweichung zum jeweiligen Mittelwert wird in der Tabelle ebenfalls angegeben. Diese ist jedoch aufgrund der Größe der Stichprobe nur bedingt aussagekräftig.

Tab. 3: Studierendenmeinungen zum Forschungsprojekt

Aussage	**Mittel**	**Std.Abw.**
Ich habe zum ersten Mal eine solche Forschungsarbeit im Zuge einer LV durchgeführt.	1,86	1,03
Die Forschungsfragen der Arbeitsaufgabe waren relevant für meine späteren Aufgaben als Lehrperson.	2,00	1,35
Die Forschungsfragen der Arbeitsaufgabe waren interessant.	1,79	1,31
Die Forschungsarbeit mit Daten war interessanter als eine gewöhnliche Seminararbeit zu verfassen.	2,08	1,12
An einer realen Forschungsfrage zu arbeiten hat mich motiviert.	2,00	1,24
Durch die Forschungsarbeit habe ich nun eine bessere Vorstellung davon, was es heißt, fachdidaktische Forschung zu betreiben.	1,43	0,76
Durch die Forschungsarbeit werden mir die Inhalte der LV länger in Erinnerung bleiben.	1,93	0,83

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik" 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

Tab. 3 zeigt die Resultate der Erhebung der Studierendenmeinungen zur Arbeitsaufgabe „Forschungsprojekt". Die Mittelwerte bewegen sich dabei zwischen 1,43 und 2,08. Die Antworten der Studierenden zeigen, dass sie durch die Forschungsarbeit einen besseren Einblick in die fachdidaktische Forschung erhielten. Die Studierenden empfanden die Aufgabenstellung als interessant und gaben an, dass die Forschungsarbeit im Zuge der Lehrveranstaltung eine neue und innovative Erfahrung für sie war. Sie gaben des Weiteren an, dass sich das Projekt positiv auf das langfristige Behalten der Lehrveranstaltungsinhalte auswirken wird. Ebenso beurteilten sie die Arbeitsaufgabe, an einer realen, noch unbeantworteten Forschungsfrage zu arbeiten, als motivierend und bewerteten das Projekt als überwiegend relevant für ihre spätere Lehrtätigkeit. Außerdem gaben die Studierenden an, dass sie diese Art datengenerierender Forschungsarbeit interessanter fanden als eine herkömmliche, rein auf Fachliteratur basierende Seminararbeit zu verfassen.

Tab. 4: Studierendenmeinungen zu zukünftigen Forschungstätigkeiten

Aussage	**Mittel**	**Std.Abw.**
Ich werde nach dieser LV wahrscheinlich wieder einmal ein Forschungsprojekt im Bereich der Fachdidaktik durchführen (z.B. mit meiner Klasse als Lehrperson).	2,33	1,23
Die Durchführung der Arbeitsaufgabe hat mein Interesse an der fachdidaktischen Forschung geweckt.	2,46	1,13

Tab. 4 stellt die Studierendenmeinung zu Aussagen über mögliche zukünftige Forschungstätigkeiten dar. In Hinblick auf zukünftige Forschungsprojekte schien eine Durchführung eines solchen Mikro-Projektes ebenfalls positive Impulse zu setzen und Interesse an weiterer Forschung zu instigieren. Angesichts der Mittelwerte von 2,33 bzw. 2,46 (jeweils von 5) sind die Rückmeldungen der Studierenden als durchaus erfreulich zu bewerten. Vier TeilnehmerInnen (29%) an der Lehrveranstaltung stimmten der Aussage völlig zu, sich vorstellen zu können, wieder einmal forschend aktiv zu werden. Zwei weitere (14%) stimmten derselben Aussage eher zu. Nur ein/e StudentIn stimmte der Aussage überhaupt nicht zu.

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik" 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

Tab. 5: Studierendenmeinungen zu akademischer Tätigkeit

Aussage	**Mittel**	**Std.Abw.**
Die LV hat mich dazu bewogen, eine Diplomarbeit im Bereich der Fachdidaktik in Erwägung zu ziehen.	3,25	1,22
Die Ergebnisse der Arbeit bei einer Tagung präsentieren zu können war ein zusätzlicher Anreiz für mich.	3,64	1,50

Tab. 5 veranschaulicht die Studierendenmeinung zum Effekt des Forschungsprojekts auf die Motivation hinsichtlich weiterer akademischer Tätigkeiten. Trotz des teilweise initiierten Interesses an der Forschung gaben nur einzelne TeilnehmerInnen an, diese Forschungsbemühungen in weiterführende akademische Tätigkeiten wie eine Tagungspräsentation oder eine Diplomarbeit in der Fachdidaktik umsetzen zu wollen. Nur ein/e StudentIn stimmte der Aussage völlig zu, dass die Lehrveranstaltung sie dazu bewogen hat, eine fremdsprachendidaktische Diplomarbeit in Erwägung zu ziehen. 50% der StudentInnen bewerteten diese Aussage neutral. Für sechs StudentInnen (43%) war die Möglichkeit, ihre Arbeit bei einer wissenschaftlichen Tagung präsentieren zu können, überhaupt kein zusätzlicher Anreiz. Die Ergebnisse in Tab. 5 sind daher wohl mehr als Arbeitsauftrag und Feedback für die LV-Leitung zu verstehen denn als Kritikpunkt am hier beschriebenen Forschungsprojekt. Diese Vermutung müsste jedoch durch über die schriftliche Befragung hinausgehende bzw. auf dieser basierende Interviews verfestigt oder ggf. entkräftet werden.

6. Schlussfolgerungen

Trotz des limitierten Datensatzes kann abschließend zusammengefasst werden, dass die ersten Erfahrungen mit der Implementierung eines Mikro-Forschungsprojekts in den Kursanforderungen einer universitären Fremdsprachendidaktiklehrveranstaltung durchaus positiv und vielversprechend zu bewerten sind. Angesichts dessen, dass eine derartige Forschungsarbeit nicht nur für die Studie-

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik“ 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

renden motivierend, interessant und relevant erscheint, sondern dies auch einen genuinen Wissenszuwachs über das Forschungsfeld zur Folge hat, kann ein weiterführender Ausbau der dargestellten Bewertungsmodalitäten nur empfohlen werden.

Literatur

Adolphs, S. & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics, 24* (4), 425-438.

Ammann, M. & Ostendorf, A. (2007). Forschendes Lernen – über die Verbindung forschungsmethodischer und fachlich-inhaltlicher Kompetenzentwicklung in der universitären Lehrerbildung. In C. Kraler & M. Schratz (Hrsg.), *Ausbildungsqualität im Lehrerberuf* (123-139). Wien: LIT-Verlag.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R. et al. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Bartels, N. (2005). *Applied Linguistics and Language Teacher Education.* New York: Springer.

Bloom, B. S. (1956). *Taxonomy of Educational Objectives, the classification of educational goals – Handbook I: Cognitive Domain*. New York: McKay.

Bosse, D. (2012). Vom Unterrichtsbeamten zum autonomen Schulreformer – Schulentwicklung als essenzieller Bestandteil universitärer Lehrer/innenbildung. In C. Kraler, H. Schnabel-Schüle, M. Schratz & B. Weyand (Hrsg.), *Kulturen der Lehrerbildung – Professionalisierung eines Berufsstands im Wandel* (89-103). Münster: Waxmann.

Fechheimer, M., Webber, K. & Kleiber, P. (2011). How Well Do Undergraduate Research Programs Promote Engagement and Success of Students? *CBE Life Sciences Education, 10,* 156-163.

Gyllstad, H., Vilkaité, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL International Journal of Applied Linguistics, 166,* 276-303.

Kremmel, B. (2012). *Explaining variance in reading test performance through linguistic knowledge: the relative significance of vocabulary, syntactic and phraseological knowledge in predicting second language reading comprehension*. Masterarbeit, Lancaster University. Lancaster.

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik" 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

Kremmel, B. & Schmitt, N. (i.V.). *Interpreting Vocabulary Test Scores: What Do Various Item Formats Tell Us about Learners' Ability to Employ Words?*

Martinez, R. (2011). *The development of a corpus-informed list of formulaic sequences for language pedagogy*. Dissertation, University of Nottingham. Nottingham, UK.

Martinez, R. & Schmitt, N. (2012). A Phrasal Expressions List. *Applied Linguistics, 33* (3), 299-320.

McKayle, C. A. (2011). *Involving Undergraduates in Research*. Verfügbar unter: http://www.qem.org/PDM Presentations folder/McKayleInvolving Undergraduates in Research.pdf [29.03.2016].

Nagda, B. A., Gregerman, S. R., von Hippel, W. & Lerner, J. S. (1988). Undergraduate Student-Faculty Research Partnerships Affect Student Retention. *The Review of Higher Education, 22* (1), 55-57.

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Hrsg.), *Vocabulary in a second language: Selection, acquisition, and testing* (3-13). Amsterdam: John Benjamins.

Nation, I. S. P. (2006). How Large a Vocabulary Is Needed For Reading and Listening? *Canadian Modern Language Review, 63* (1), 59-82.

Nation, I. S. P. & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31* (7), 9-13.

Nation, I. S. P. & Webb, S. (2011). *Researching Vocabulary*. Boston, MA: Heinle-Cengage ELT.

Paul, P. V., Stallman, A. & O'Rourke, J. P. (1990). *Using three test formats to assess good and poor reader's word knowledge*. Technical Report No. 509. Urbana-Champaign: University of Illinois, Center for the Study of Reading.

Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. Basingstoke: Palgrave Macmillan.

Schmitt, N. & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47* (4), 484-503.

Schmitt, N., Schmitt, D. & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18* (1), 55-88.

Schratz, M. & Weiser, B. (2002). Dimensionen für die Entwicklung der Qualität von Unterricht. *Journal für Schulentwicklung, 6* (4), 36-47.

Stewart, J. (2014). Do Multiple-Choice Options Inflate Estimates of Vocabulary Size on the VST? *Language Assessment Quarterly, 11* (3), 271-282.

Trim, J., North, B. & Coste, D. (2001): *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Webb, S. & Rodgers, M. P. H. (2009). Vocabulary demands of television programs. *Language Learning, 59* (2), 335-366.

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik" 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.

Widdowson, H. G. (1991). *Aspects of language teaching*. Oxford: Oxford University Press.

Van Zeeland, H. & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics, 34,* 457-479.

Barbara Hinger (Hg.), Zweite „Tagung der Fachdidaktik" 2015.
Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht.
 ISBN 978-3-903122-51-2, DOI 10.15203/3122-51-2