

DaF: Freie Arbeiten im Detail beurteilen I

Rudolf Reinelt

Inhaltsverzeichnis

Einleitung: Drei Ansätze

1. Vorstellung der Problematik

1.1. Daten

1.2. Hintergrund: Der Unterricht bis zum Test

Unterrichtsablauf

Optimierung

Gesamtkursergebnis

Gelerntes und Form des Tests: Bedingungen

Ein neuer Test: „Freie Arbeiten“

Beurteilung der Arbeiten

1.3. Einige Probleme des optimierten Unterrichtsansatzes und der durchgeführten Prüfung

2. Kriterien für detaillierte Beurteilungen

2.1. Wie geht die Beurteilung?

2.1.1. Beurteilungen und Maßstäbe

2.1.2. Detailbeurteilungen

2.2. Das Beurteilungsraster

2.3. Besonderheiten

2.3.1. Reduzierte (dekremental additive) Punktevergabe

3. Zwei Bewertungsansätze

3.1. Vorbemerkung

3.2. “Freie Arbeiten” holistisch: Eindruck und Initialbewertungen

3.2.1. Holistische Schnell- und Minimalbeurteilung

3.3. Textbeispiele für Einzelanalysen

Einleitung: Drei Ansätze

In letzter Zeit können im Fremdsprachenunterricht in Japan drei Ansätze unterschieden werden:

- a) Der traditionelle Ansatz, dessen Hauptbestandteil immer noch die Grammatik-Übersetzungsmethode ist, bei der, extrem vereinfacht gesagt, der Lehrer oder die Lerner etwas in die Muttersprache übersetzen und der Lehrer Grammatik erklärt und Lerner in Tests übersetzen oder Multiple-Choice-Fragen beantworten. Produktivität in der Fremdsprache ist nicht vorgesehen. Trotz aller Bemühungen des Kultusministeriums (MEXT (n.d.)) dürfte sich, abgesehen von einigen wenigen Schulen, im Bereich des Englischunterrichts nicht allzu viel geändert haben, Eine informelle Umfrage unter 74 Studienanfängern für Deutsch an der Ehime Universität ergab im April 2018: Selbstgefühl nur 30% Aktivitäten der Lerner und nur 26% der Zeit in Englisch im Englischunterricht an der Oberschule. Dies betrifft auch noch weitgehend das Unterrichten von zweiten Fremdsprachen an Universitäten, obwohl hierzu keine neuen Zahlen vorliegen.
- b) In der letzten Zeit werden in vielen Fällen unter dem Namen "Active Learning", auch mit Hilfe von Lehrbüchern und on-line-Medien, Lerneraktivitäten integriert und die Lerner so am Unterricht beteiligt. Oft werden sogar Rollenspiele geübt und z.T. sogar vorgeführt. Es ist aber auch bei dem augenblicklichen Vorgehen fraglich, ob viel von der Fremdsprache über das reine "Lernen" hinaus verwendbar erworben wird.

Beiden Ansätzen ist gemein, dass die Lerner unmittelbar nach der zumeist von Anfang an schriftlichen Einführung und Behandlung eines inhaltlichen Teils nach dem Schließen des Lehrbuches, oder dem Ende der Kursstunde bzw. dem Abschluss des Semesters sehr wenig behalten. Wenig erfolgreiche Erfahrungen damit, wieviel solche Lerner z.B. frei in der Zielsprache, also nach sechs Jahren Englisch bzw. einem Jahr Deutsch sprechen können, sollten nicht allzugroße Erwartungen wecken.

- c) Vor diesem Hintergrund schreiben viele Neuanfänger in ihrem „Wunsch“ und als „Zufriedenheitskriterium“ für den zweiten Fremdsprachenkurs an der Universität "kaiwa", d.h. Sprechen: 2018 gab es z.B. insgesamt 75 Nennungen von 74 Studenten bei den beiden Items. Der Verfasser hat daraufhin einen Deutschkurs entwickelt, in dem die Studenten im Unterricht selbst sehr viel sprechen. Übungen zu den anderen, ebenfalls obligatorischen, Fertigkeiten werden weitgehend ausgelagert, da sie auch im Internet erledigt werden können.

In diesem optimierten Kurs lernen die Kursteilnehmer nach dem „Ökonomischen Prinzip“ von Humboldt: „*Von begrenzten Mitteln unbegrenzten Gebrauch machen*“, hier Unbegrenztheits-

prinzip (UBP) genannt, im eigentlichen Sinn von “Sprachen (auch) sprechen lernen” z.B. sehr viele Kommunikationssituationen zu bestehen.

Aufgrund der Spontanitätsanforderung erfolgt allerdings das Überprüfen durchgehend noch im Unterricht bzw. in der Prüfung am Semester- bzw. Jahresende. Für die Fertigkeit Schreiben wurden dafür “Freie Arbeiten” als ein neuer Testtyp entwickelt.

Das Ziel dieses Beitrags ist, ansatzweise einen Apparat bereitzustellen, wie die mit dieser Testform produzierten “Freien Arbeiten” detailliert analysiert und für die Lerner und Kursleiter für die Vorbereitung weiterer Kurse nutzbar gemacht werden können.

Übersicht

Im ersten Teil stellen wir die Daten, den Hintergrund und einige Problematiken des optimierten Unterrichtsansatzes und der durchgeführten Prüfung vor. Teil zwei stellt einige Kriterien vor für einen Ansatz zur detaillierteren Beurteilung und was für Probleme bei der Benutzung vorkommen können. Teil drei stellt zwei Ansätze, die holistische Gesamtbewertung und die detaillierte Punktebewertung anhand von Textbeispielen gegenüber. Teil vier zeigt die Details von Kriterien in der Anwendung auf zwei Testtexte. Teil fünf diskutiert Weiterungen und Teil sechs präsentiert als Referenz und Übungsanhang einen der längsten Texte, der je von StudentInnen in den 80 Minuten geschrieben wurden.

Alle Beispiele in diesem Beitrag sind so anonymisiert, dass Übereinstimmungen mit und Schlüsse auf beteiligte Personen bzw. die Lerner rein zufällig, aber aufgrund der beschränkten Anzahl an Namen in Japan natürlich möglich sind.

In diesem Beitrag sind *Kursteilnehmer*, *Student* und *Lerner*, sowie *Verfasser*, *Lehrer* und *Kursleiter* jeweils gleichbedeutend verwendet, außerdem wird repräsentativ für beide Geschlechter die männliche Form gewählt.

Vorarbeiten des Verfassers

Zu den Bewertungen der abschließenden Prüfungen liegen viele Arbeiten des Verfassers vor, in einigen wurden schon Bewertungen der schriftlichen Produkte untersucht, um zu schnellen Beurteilungsverfahren und Ergebnissen zu kommen (Reinelt 2014, 2017a, 2018) bzw. diese durch wiederholte Tests gegenzuchecken (Reinelt 2017b). Allerdings mussten dabei detaillierte Untersuchungen der Produktionen hintangestellt werden, und diese sollen hier Untersuchungsgegenstand sein.

1. Vorstellung der Problematik

1.1. Daten

Tafel 1 enthält die wichtigsten Daten zu den in dieser Arbeit verwendeten Produktionen.

Tafel 1 Wichtige Daten

Unterrichtsstunden in Deutsch für Neuanfänger beim Verfasser:
2x 90 Min (1x davon beim Verfasser) pro Woche, 15 Unterrichtsstunden pro Semester (die letzten beiden für mündliche und schriftliche Tests) x2 Semester (SS, WS), Kursabsolventen 2018 etwa 70, hier betrachtet: Arbeiten von etwa 20 Lernern aus einer Klasse im Wintersemester (WS).

1.2. Hintergrund: Der Unterricht bis zum Test

Unterrichtsablauf

Der Unterricht beim Verfasser selbst findet in Kleleinheiten von je 5 bis 20 Minuten statt, wonach möglichst zu einer anderen Fertigkeit, zumindest aber zu einer anderen Aktivität bzw. einem anderen Thema gewechselt wird, so dass die Lerner am Ende des Semesters sowohl einerseits die Fremdsprache in ihrem gelernten Rahmen sprechen, sie aber andererseits auch lesen und schreiben können.

Nach der Unterrichtsstunde trägt der Kursleiter die nicht-neuen Inhalte in einen im Moodle der Ehime-Universität bereitgestellten und von den Kursteilnehmern ein bis zwei mal pro Woche einzusehenden sogenannten "Unterrichtsfile" ein. Neue Inhalte werden nach der zweiten oder dritten - idealerweise erst nach der vierten - Behandlung eingetragen, um die unvermeidliche Löschung des mündlich Gelesenen durch zu frühe schriftliche Repräsentation zu verhindern.

Optimierung

Die Kursteilnehmer lernen im Laufe des Kurses beim Sprechen, und nachher bei Übungen zum Schreiben dazu, möglichst viele verschiedene Bereiche und Verwendungen der Zielsprache kennen, ebenso wie ein weites Vokabular, das allerdings nicht bei allen Kursteilnehmern gleich ist, da - abgesehen von unterschiedlichen Lernanstrengungen - einige Aufgaben mit der Realität der Lerner zu tun haben und diese von Person zu Person verschieden ist.

Gesamtkursergebnis

Insgesamt haben die Lerner am Ende des Kurses wohl A2 erreicht, was in der mündlichen Prüfung bewiesen wird. In anderen Bereichen, wie z.B. in den ab hier betrachteten schriftlichen Tests dürften die Unterschiede allerdings sehr groß sein.

Gelerntes und Form des Tests: Bedingungen

Für den nach einem solchen Unterricht zu erstellenden Test gibt es dann einige Bedingungen:

- Aufgrund der vielfältigen Verwendungsmöglichkeiten einzelner sprachlicher Elemente kann man nicht mehr nur 1:1 Relationen zwischen sprachlichen Einheiten und bestimmten Bedeutungen und/oder Funktionen annehmen (außer bei Sprechakten wie *Danken* und *Entschuldigungen*). Die Lerner im Unterricht beim Verfasser lernen und üben auch, dass es mehrere Möglichkeiten des sprachlichen Ausdrucks geben kann und diese auch mehrere Verwendungen haben können. Dann reichen streng formale Tests, bei denen diese Relation weitgehend vorausgesetzt ist, nicht mehr aus.
- Der Bezug zum Gelernten ist in sofern wichtig, als nicht irgendetwas Neues verlangt werden kann. Wenn Lerner aber etwas ganz Neues, noch dazu in perfektem Deutsch, schreiben, muss zumindest der Verdacht des Abschreibens/Auswendiglernens angenommen werden.
- Die Kursteilnehmer haben so viele verschiedene Bereiche kennengelernt, dass ein „Abtesten“ jedes einzelnen Bereichs eine eigene Testzeit von mehreren Unterrichtsstunden in Anspruch nehmen würde.
- Die Tests müssen so offen sein, dass die Lerner so viel wie möglich präsentieren, darstellen können. Im Sinne eines *formative assessment* (Wicking 2016, 2017) hilft eine solche Art der Bewertung den Lernern dann auch und gibt ihnen Vertrauen in ihr Gelerntes (vgl. die Studentenmeinungen in Reinelt 2017c).
- Die Kurse des Verfassers haben am Ende also einerseits keinen fest standardisierten Test. Andererseits sind sie auch nicht ganz frei, oder auf Abtesten von Kreativität, wie in gelenkter Form im *Creative Writing* vorliegend, ausgerichtet.
- vorher standardisierte Tests sind also als Abschluss ausgeschlossen (obwohl sie im Unterricht selbst vorkommen), da sie immer ungerecht wären: Entweder wird der Bereich getroffen, den ein Lerner besonders gut beherrscht oder eben nicht, und es müssten Bereiche unberücksichtigt bleiben, die z.B. wiederum ein anderer Lerner beherrscht.
- Trotz des Unbegrenztheitsprinzips umfasst das Gelernte aber sicher zu wenig (einheitliches) Vokabular, um einen Bereich ausführlich darzustellen.
- Die Lerner haben nicht alle Teile eines Bereiches oder Themas so gut gelernt, das sie einen eigens daraufhinzielenden Test gut bestehen könnten.
- Bei dem Verfassen handelt es sich sicherlich – bei aller guten Absicht - nicht um durchgehendes Konstruieren. Dies ist wahrscheinlich in der Zielsprache sehr schwierig und würde wohl schiefgehen!
- Andererseits erfolgt die Texterstellung nicht so automatisch wie im muttersprachlichen

Alltag.

- Aufgrund der notwendigen Offenheit erfordern die Tests viel Kreativität von den Lernern, entweder schon in der Vorbereitung, oder an bzw. auf der Stelle nach der Nennung des Testthemas.

Ein neuer Test: „Freie Arbeiten“

Im Test am Ende des Semesters bzw. Jahres sollen die Lerner zeigen können, wieviel sie gelernt haben. Um dies zu ermöglichen, musste ein neuer Typ bzw. ein neues Format eines schriftlichen Tests entwickelt werden, das wir “freie Arbeiten” genannt haben. Dies wurde auch ansatzweise im Unterricht vorher geübt.

Ähnliche Versuche sind schon bekannt. So hat z.B. Melse (1990) die folgende Arbeitsdefinition:

„Writing skills refer to the skills of producing a written, cohesive, and well rounded text in which formulations can be chosen freely.“ Genauere Beschreibungen und Inhalte finden sich auch in Fasoglio et al. (2015:11), wobei allerdings die Anforderungen von zusammenhängenden Texten und guten Formulierungen in Japan normalerweise nicht übernommen werden können.

Jeweils am Ende des ersten und zweiten Semesters machen die Lerner in den Kursen beim Verfasser eine mündliche und eine schriftliche Prüfung, die aus technischen Gründen gleichzeitig durchgeführt werden und sich normalerweise über zwei Mal/Wochen erstrecken.

A) Beim ersten Prüfungstermin werden alle Lerner einzeln in zufälliger Reihenfolge aus dem Klassenzimmer in ein anderes Zimmer gerufen und dort sprechen sie ca. 2 bis 3 Minuten mit einem deutschen Muttersprachler oder einem habituellen Muttersprachler (Reinelt 2015, 2017d), danach gehen sie zurück ins Klassenzimmer und schreiben weiter. Während der ganzen Zeit, also etwa 80 von 90 Minuten Unterrichtszeit, schreiben alle Studenten gleichzeitig gänzlich ohne Materialien, Lehrbuch usw. eine “freie Arbeit”:

- im Sommersemester zum Thema “Mein Deutsch” mit der Anweisung, alles zu schreiben, was sie bis dahin gelernt (und an dem Tag erinnert) haben;
- im Wintersemester mit dem Thema “Ich” mit der Anweisung, sich und alles was sie auf Deutsch können so darzubieten, dass ein zukünftiger Firmenchef aufgrund dieser Arbeit sie sofort einstellen würde.

Weitere Hinweise am Anfang des schriftlichen Tests:

- Es geht natürlich darum, soviel wie möglich zu schreiben. Dabei sind Wiederholungen des sprachlichen Materials, allerdings in unterschiedlicher Funktion möglich.
- Die Darstellung als Gespräch ist einfacher und überzeugender als das Verfassen eines längeren Textes: Dialogisch kann man mehr schreiben und damit mehr Punkte bekommen und

auch sich selbst während des Schreibprozesses mehrere Antworten finden lassen (sich an eine Wiederholung aus dem Unterricht erinnern), und der Beurteiler kann ein Gespräch leichter nachvollziehen. Dies dient auch der Vermeidung von Lücken oder inhaltlich unverbundenen Teilen (Inseln) oder konversationell unsinnigen Reihungen (sonst schreibt jemand alle Zahlen von 0 bis 1000);

– Beim zweiten Test “Ich” sollen die Lerner auch darauf achten, dass genügend Inhalte aus dem Wintersemester (haben/es gibt, Kleidung, Wohnung, Beschreibung von Gegenständen, Modalverben, Studentenfest (Perfekt), Weihnachten, Präpositionen, Lagen, usw) vorkommen. B) Beim zweiten Mal etwa eine Woche später werden die Studenten, die beim ersten Mal nicht mündlich geprüft werden konnten, wie die Studenten beim ersten Mal einzeln zur mündlichen Prüfung gerufen. Als Check, ob das erste Mal nicht einfach nur Auswendiggelerntes geboten wurde, schreiben alle Kursteilnehmer beim zweiten Mal frei zum gleichen Thema und mit den gleichen Bedingungen und Anweisungen, diesmal aber nur 60 Minuten, in denen die meisten auch fertig werden.

Diese zweite schriftliche Prüfung wurde erst vor zwei Jahren eingeführt, um auswendig Gelerntes und echt Gelerntes, das auch nach dem Abliefern des Auswendiggelernten noch zur Verfügung steht, zu unterscheiden (Reinelt 2017b). Eine detaillierte Untersuchung zu den Unterschieden steht noch aus, ein Vergleich wurde in Reinelt (2018) unternommen.

Natürlich sind die Inhalte, die die Studenten wählen, ganz frei und die Stimmungslage der Studenten ist sicher anders als vorher: Beim ersten Mal ist sie sicher hoch, dann kommt wahrscheinlich der Abschwung, wenigstens für den Bereich des Schreibens. Dann stellt dieses zweite Schreiben aber eher den Mindeststand des Fremdsprachenwissens bzw. -könnens dar, auf dem dann aber aufgebaut werden kann.

Beurteilung der Arbeiten

In diesem Beitrag beschränken wir uns in Teil zwei bis vier nur auf die schriftlichen Produktionen aus diesem Test aus einer Klasse im WS 2018.

In einem ersten Durchgang werden alle Arbeiten schnell beurteilt, um in dem kurzen gegebenen Zeitrahmen an der Universität zu einer Note zu kommen. Dabei werden die Arbeiten nach Punkten und auch holistisch, aber nicht im Detail, benotet, wobei beide Typen von Benotungen adäquat zu sein scheinen (Harsch & Martin 2012). Allerdings ist für ein sinnvolles Studentenfeedback, wenn dies gegeben werden kann, und für Kursleiter als Quelle von Verbesserungen im folgenden Jahr eine detaillierte Untersuchung nötig. Den Apparat dazu versuchen wir im folgenden zu entwickeln.

1.3. Einige Probleme des optimierten Unterrichtsansatzes und der durchgeführten Prüfung

Der in 1.2. vorgestellte Ansatz versucht den Spagat zwischen möglichst viele Möglichkeiten zum Zeigen geben einerseits und andererseits das Gezeigte dann genauer beurteilen. Bevor wir auf Einzelheiten eingehen, seien hier noch einige Vorbemerkungen gemacht:

A) Bei der Beurteilung sind folgende Punkte wichtig:

In dem hier vorgestellten Verfahren ist die Beweislast fast umgekehrt: Während in standardisierten Tests die Ausführung für die Lerner schwer ist, sind die Produktionen leicht durch die Beurteiler zu beurteilen, da, idealerweise, schon alle Kriterien vorher abgeschätzt sind. In den beim Verfasser gegebenen Tests ist es genau umgekehrt: Die Lerner haben schon etwas, das sie präsentieren - und es ist bei vielen Lernern beileibe nicht das gleiche - und nun ist es an den Beurteilern zu beurteilen, ob das gut oder schlecht ist, bzw. wie viele Punkte man dem zuteilen kann.

Ein wichtiger Unterschied zu anderen Verfahren ist auch, dass bei der detaillierten Beurteilung nicht zu eng anzusetzen ist. Vielmehr geht es darum, mehrere Möglichkeiten in Betracht zu ziehen.

B) Das offene Format der freien Arbeiten geht über normale Testformate hinaus, und dabei müssen die Lerner eine bestimmte Stufe von Kreativität erreichen.

Die Verortung des Tests zeigt eine Mittelstellung zwischen festem Test und Creative Writing. Diese Situation geht z.B. gegen Mozaffari (2013): So ist z.B. Kollegenfeedback aufgrund der Verschiedenheit der Kurse aus praktischen Gründen nur schwer möglich. Ebenso ist ein Vergleich der Produktionen der Lerner in diesen Kursen mit solchen aus anderen Kursen nicht möglich: Welcher Lehrer (nicht nur in Japan) würde dies mit seinen Studenten erlauben?

In Wirklichkeit brauchen die Lerner in diesem Test viele Besonderheiten einer kreativen Persönlichkeit (s. die Kreativitätsdilemmata bei Szentmihalyi (in Cimermanova 2013: 20)). Kreativität ist im Unterricht sinnvoll und wird auch von den Studenten angenommen.

Andererseits sind deshalb die vier Kriterien für Creative Writing im Beitrag von Mozaffari (2013: 2216) nicht ohne weiteres anwendbar: "image, voice, characterization, story und validated against expert".

"Image", "voice" und "characterization" lassen sich bei dem geringen Vokabular, das den Lernern zur Verfügung steht, nur wenig modulieren, obwohl schon einige ihre Produktion als eine Art Geschichte, eine Art "story" aufbauen.

Anders als bei Creative Writing sollen die Ergebnisse gerade nicht öffentlich gemacht werden: Genau dies wollen japanische Studenten nicht. Auch die korrigierte Rückgabe ist nur bedingt

sinnvoll.

Was den Produktionen noch am nächsten kommen könnte, sind Beschreibungen im GER, sowohl die in den Can Do -Listen, wie auch die für Schreiben in den Stufen-Beschreibungen (CoE (n.d.) und als Versuch zum Schreiben: Harsch & Rupp (2011)). Um den letztgenannten Punkten nachzugehen, ist es nötig, auf die Details der Arbeiten einzugehen. Dies erfolgt im nachfolgenden Teil mit einem Kriterienraster.

2. Kriterien für detaillierte Beurteilungen

Teil zwei stellt einige Kriterien für einen Ansatz zur detaillierteren Beurteilung vor und was für Probleme bei der Benutzung vorkommen können.

2.1. Wie geht die Beurteilung?

2.1.1. Beurteilungen und Maßstäbe

Bevor wir überhaupt eine Beurteilung abgeben können, ist es wichtig, festzulegen, was der Maßstab sein könnte.

Mehrere Stufen der Beurteilung relativ zu einem Maßstab bieten sich an, die auch mit mehreren Analysemitteln einhergehen. Hier ist eine unvollständige, hierarchisierte Aufstellung:

– Eigentlich gibt es bei “freien Arbeiten” keine absolute Beurteilung, da der Gebrauchsbereich der Sprache unendlich ist, und somit alles, was nicht gerade falsch oder unverständlich ist, als gute Leistung anzusehen ist. Es bleibt also die Frage: Relativ zu was?

– Relativ zum Deutschen als Ganzen: Sicher nicht sinnvoll, da es sich um Anfänger handelt;
– Intern relativ zum ganzen Deutschlernen: Dies wäre sicher sinnvoll, damit die Lerner ihren Fortschritt sehen können. Dies wird auch bei der mündlichen Prüfung angedeutet;

– Relativ zum im Unterricht Behandelten? Wäre sinnvoll, würde aber eine – leider nicht machbare - komplette Aufzeichnung des Unterrichtsgeschehens voraussetzen;

– oder relativ zum Unterrichtsfile (Grammatik, viele Themenbereiche, Chunks, usw.). Dies wäre die einfachste Methode. Dazu könnte man den Unterrichtsfile einscannen und dann die von den Studenten gebrauchten Wörter, Phrasen und Ausdrücke mit einem Concordancer (z. B. Antconc (n.d.)) gegenrechnen lassen. Leider ist die Einlesetechnik noch nicht so weit!

– Relativ zu bestimmten Stufen des GER. Dies ist sicher die beste Vorgehensweise, Sie ist allerdings schon teilweise dadurch abgesichert, dass die Kurse von vornherein im GER erwähnte Elemente enthalten. Eine genauere Untersuchung wäre ein eigenes Forschungsprojekt.

– Im gegenseitigen Vergleich: “Mein Deutsch” im SS und “Ich” im WS. Damit könnte man

den Fortschritt gegenüber dem ersten Semester herausfiltern, soweit er sich in den Produktionen der Lerner niedergeschlagen hat. Teilweise ist dies schon durch die Anweisung an die Lerner garantiert, selbst dafür zu sorgen, dass das WS in den freien Arbeiten genügend repräsentiert ist. – Schließlich könnte man die Produktionen daraufhin überprüfen, was sie jeweils an Neuem bieten. Da aber diese Art von Kreativität, bis auf einige schon im Unterricht geübte Ausnahmen (interessante Selbst- bzw. Familienvorstellung, Urlaubsbericht, Deutschlandreisebericht), normalerweise in Japan auch in der Muttersprache nicht vorgesehen ist, entfällt dieses Kriterium aus Gründen der Fairness. Besonders interessante Teile werden natürlich zusätzlich positiv bewertet.

2.1.2. Detailbeurteilungen

Jedes Beurteilungssystem soll zu gerechten und berechenbaren Beurteilungen führen. Dies ist bei freiem Schreiben wie in den hier diskutierten Beispielen nur sehr schwer zu erreichen. Gemessen an den in 2.1.1. angegebenen Maßstäben sind die Beurteilungen sehr schwierig, z.T. eben auch wegen dem UBP, d.h. der Beurteiler kann erst beim genaueren Ansehen der “freien Arbeit” feststellen, um was es geht und wie gut die Arbeit ist.

Um aber die Arbeiten auch für eine detaillierte Beurteilung und als Schatz für Probleme, die in späteren Unterrichten zu beheben sind, zu nutzen, stellen wir im folgenden detailliertere Ansätze vor.

Was sich machen lässt, ist ein Analysemittel, z.B. in der Form eines aus den bisherigen Produktionen gewonnen Rasters, bereitzustellen und damit zu versuchen, diejenigen Produktionen zu erfassen, die von den Studenten vorliegen. Dieses Raster und einige Diskussionspunkte werden im folgenden vorgestellt. Am Ende sollte, wie bei Beurteilungen, bei denen einzelne Items isolierbar sind, ein Punkteergebnis stehen. Es wird also darum gehen, wieviel Punkte jeweils für eine bestimmte Produktion vergeben werden.

Um eine Punktevergabe zu ermöglichen, ist es nötig, die zu beurteilende Einheit sehr klein zu machen. Menschliches Sprechen und dessen schriftliche Darstellung geschieht normalerweise in *Sätzen*, die selbst wieder aus *Wörtern* bestehen. Diese beiden Kategorien sind dann die Gegenstände der Analyse.

Die Studenten schreiben unaufgefordert meist nur einen Satz oder eine Äußerung in eine Zeile. Dies gibt uns einen Ansatz zur Trennung. Besonders im WS erscheinen aber auch längere Konversationen, die dann auch als Ganzes bewertet werden können.

Dabei wird *jede Zeile einzeln betrachtet*, und wenn nötig *einzelne Ausdrücke, z.B. jedes Wort einzeln*.

Ein besonders Problem stellen Wortkombinationen, d.h. Chunks (Krause-Ono 2018), dar: Wie sind Redewendungen und andere Phrasen zu bewerten? Als ganzes oder einzeln? Wir werden an jeder Stelle angeben, welche Beurteilung jeweils sinnvoll erscheint. Teilweise geben wir auch Argumente dafür und dagegen.

2.2. Das Beurteilungsraster

In diesem Teil stellen wir einige Kriterien anhand von Bewertungsbeispielen für “freie Arbeiten” dar. Die Auswahl der Kriterien hier ist nur ein Vorschlag, es gibt keine Notwendigkeit, dass es genau diese sein müssten, und es können auch weniger oder sogar mehr sein, wenn dies notwendig würde.

Tafel 2: Die Kriterienliste und Beispiele

Titel:

Phänomen in Deutsch	[der Originaltext aus der “freien Arbeit”]
Strukturtyp	vorher? nachher?
Variationen	
Beschreibung	
Begründung für die Einbeziehung in die Bewertung	
Punkte relativ zu SS bzw. WS oder später	x Punkte (=p), u.U. mit Bedingungen
Anmerkungen	z.B. Fehler

Beispiel: Freizeitfrage

Phänomen in Deutsch	Was machen Sie in der Freizeit?
Strukturtyp	Feste Phrase, setzt Anfangsgespräch voraus, gefolgt von freien Inhalten, mit <i>ich</i> im Subjekt möglich.
Variationen	Unterschied zu Hobby! Das im Deutschen zu fragen ist nicht üblich!
Beschreibung	Allgemeine Frage nach Aktivitäten außerhalb des Arbeitszusammenhangs
Begründung für Einbeziehung in die Bewertung	Wichtige mögliche Fortsetzung nach dem Anfangsgespräch
Punkte relativ zu SS bzw. WS	Wegen Verwechslungsmöglichkeiten mit anderen Sprechakten bei Fehlern (Was macht Sie? (Beruf) /das (bezahlen)?) :

	Im SS als feste Phrase ein Punkt pro Wort > 6 Punkte, im WS auch als zusammengesetzte Phrase: Was machen Sie+ in der Freizeit 2 Punkte oder als Freizeitfrage 1 Punkt möglich, je nach Ziel der Prüfung.
Anmerkungen	Im WS u.U. nur ein Satzpunkt

Und die Antwort auf die Freizeitaktivitätenfrage

Titel: Frageantwort im Satz

Phänomen in Deutsch	Ich lese Romane
Strukturtyp	hier: Frage, oder Überlegungsmarker o.ä.; gleiche weiter, Detaillierungsrückfrage oder Ende.
Variationen	Alle Sätze und alle Infinitive legal
Beschreibung	Antwort auf Frage, oder unabhängige Aussage über sich selbst, deshalb Satz mit Ich am Anfang, normale Satzstellung.
Begründung für Einbeziehung in die Bewertung	Ich-Sätze, ermöglicht Rückfragedialog oder Reihung
Punkte relativ zu SS bzw. WS	Ich=1, v=1, andere Teile pro richtigem Wort = 1, dann reduzierend; bei weiteren Beispielen pro richtigem Verb=1 + Rest des Satzes pro neuem Wort =1
Anmerkungen	Bekannt aus Englischlehrbuch, aber Unrealität unbekannt.

Kurze Erklärungen:

Der **Titel** leitet die Beurteilung ein und bietet eine erste Kategorisierung.

Phänomen in Deutsch ist der zu analysierende Teil, wortwörtlich aus der freien Arbeit des Lernalers übernommen.

Strukturtyp gibt an, welche Stelle der Satz/Teil in einem möglichen Gespräch bzw. anderer Konversation hat, und hat zwei Teile: Das was vorausgeht (gehen kann), und das was nachfolgen kann.

Beschreibung versucht eine grammatische oder anderweitige Einordnung in das System der Sprache/ des Deutschen.

Begründung für Einbeziehung in die Bewertung. Nicht immer ist klar, warum etwas einbezogen werden sollte bzw. zu welchen Übungen es dienen sollte. Dies wird an dieser Stelle erklärt.

Punkte relativ zu SS bzw. WS oder später: X Punkte (=p), u.U. mit

Bedingungen, z.B. 3p im SS, aber nur 1p im WS, weil dann die ganze Phrase als richtig/ bekannt vorausgesetzt werden kann.

Für jedes Wort wird ein Punkt gegeben, außer wenn es mehr als einen Schreibfehler enthält (dazu Reinelt 2017e). Richtig geschriebene Phrasen erhalten beim ersten Mal einen Punkt pro Wort, nachher nur für neue Teile (s. den nächsten Abschnitt). Für ganze Sätze gibt es immer dann einen Punkt, wenn es für die Wörter keine Punkte mehr geben kann (z.B. für eine Wiederholung).

Nicht immer ist der Wert eines Satzes der gleiche: Am Anfang des Unterrichts ist er sicher sehr hoch, weil z.B. neu. Der Wert kann aber mit der Zeit weniger Punkte bringen, z.B. am Ende des WS, wenn der Inhalt schon als bekannt vorausgesetzt werden kann. Auf jeden Fall wird aber mindestens ein Punkt für den Satz gegeben (und dann Satzpunkt genannt).

Anmerkungen erfolgen bei Besonderheiten des Ausdrucks oder seiner Verwendung. Hier sind auch Fehler angemerkt, wenn sie punktmindernd sein können.

2.3. Besonderheiten

2.3.1. Reduzierte (dekremental additive) Punktevergabe

Chunks sind Wortabfolgen, die sehr häufig in gleicher oder stark kontrolliert veränderter Weise benutzt werden. Unser menschlicher Sprachgebrauch besteht zu einem sehr großen Teil daraus, und ihre Beherrschung erleichtert den zweckgerichteten Umgang mit der Zielsprache, weil nicht immer alles neu von einem Wort zum anderen aufgebaut werden muss, was in einer Konversation auch zeitlich gar nicht möglich wäre. Im Bereich Deutsch als Fremdsprache in Japan war die erste durchgehende Verwendung von Chunks im Lehrbuch Marui/Reinelt (1987), und ihr Einsatz ist erklärt bei Krause-Ono (2018).

Sehr oft ist es sinnvoll, z.B. zu einer Frage mehrere Antworten zu geben. Die Lerner können so zeigen, dass sie einen großen Wortschatz (d.h. viel gelernt und diesen sich z.B. in Eigeninitiative angeeignet) haben. Außerdem kann der gleiche Ausdruck mit wenigen Veränderungen (z.B. grammatische Person) z.B. für die Beschreibung anderer Personen eingesetzt werden.

Die Antworten sind einander dann ähnlich (aber nicht gleich, dann zählen sie nicht) und zeigen gerade die Variation(en). Wie kann man in solchen Fällen Punkte vergeben?

Eine Möglichkeit besteht in der *dekremental additiven* Punktevergabe (s. Tafel 3). Dabei wird strukturähnlichen Elementen bei unterschiedlichen Vorkommen, z.B. beim ersten Mal gegenüber dem dritten Mal jeweils ein anderer Wert zugeschrieben.

Die folgende Abstufung in Tafel 4 demonstriert dies.

Tafel 3: Dekremental additive (abgestufte) Punktevergabe

1. Vorkommen: Pro Wort 1 Punkt
 2. Vorkommen: Pro Phrase 1 Punkt
 3. und weitere Vorkommen: Pro neuem Wort/neuer Phrase 1 Punkt.
- Bei wiederholten Vorkommen in einem anderen funktionalen Zusammenhang: 1 Satzpunkt

Betrachten wir dazu das folgende Beispiel:

Tafel 4: Beispiel für dekremental additive Punktevergabe

Ausgangsfrage::

Was machen Sie heute Abend denn so alles?

8 Punkte für die Wendung in der Frage

- morgen ---?

1 Punkt

Antworten, jeweils mit > und der Punktzahl nach dem Wort:

Von>1 18>1 Uhr>1 bis>1 19>1 Uhr mache>1 ich>1 das>1 Abendessen>1 T=9

Von 19 Uhr>1 bis 20Uhr>1 esse>1 ich>1 zu Abend>1 T=5

Von 20 Uhr bis 23 Uhr>1 lerne>1 ich Englisch>1 T=3

Von 24 Uhr bis 1 Uhr>1 lerne ich Deutsch>1 T=2

Dann>1 bade>1 ich. T=2

Um>1 21 Uhr schlafe>1 ich (eigentlich: gehe ich schlafen). T=3

insgesamt für die Antworten hier: T=24

3. Zwei Bewertungsansätze

Dieser Teil stellt zwei Ansätze kurz vor:

- a) die holistische Gesamtbewertung aller Produktionen in einer Klasse, und
- b) zwei für die detaillierte Punktebewertung vorbereitete Beispieltexte, die dann im vierten Teil ausführlich erklärt wird.

3.1. Vorbemerkung

Alle Studenten schreiben unabhängig voneinander, nicht nach einer bestimmten Lehrer- oder Partnernvorgabe, und nur das, was sie denken schreiben zu müssen, bzw. schreiben können. Wenn sich viele Lerner nur an bestimmte gleiche Teile im Unterricht erinnern, erscheint natürlich bei vielen der gleiche oder ein ähnlicher Inhalt. Dann kann es im Sinne eines schnelleren Beurteilungsablaufs reichen, nur die wichtigsten Unterschiede zu berücksichtigen. Bleibt mehr

Zeit, erlaubt andererseits ein detaillierterer Ansatz bei entsprechender Mehrarbeit vielleicht eine gerechtere Beurteilung. Im ersten Fall reichen allgemeine Kommentare und die schnellen Punktezählungen (3.2.), im zweiten Fall muss man die Texte vorbereiten, so dass sie für eine detailliertere Betrachtung benutzbar sind (3.3).

Alle Beispiele sind aus Ich 2 im WS 2017, d.h. die Studenten haben schon einmal, beim ersten Test "Ich" viel mehr geschrieben, und dieses Mal geht es für die beurteilenden Kursleiter darum, herauszufinden, wieviel auch etwa eine Woche nach dem Abschwung zu Ende des ersten langen Tests noch hängengeblieben ist bzw. erinnert wird. Wir geben dafür zwei Typen von Bewertungsbeispielen für freies Schreiben.

3.2. "Freie Arbeiten" holistisch: Eindruck und Initialbewertungen

Das folgende sind Initialbewertungen, die jeweils aus einer schnellen Durchsicht und nach einer ebenso schnellen Punktezählung und den sich aus den beiden ergebenden Kommentaren bestehen.

3.2.1. Holistische Schnell- und Minimalbeurteilung

Tafel 5 Schnellbeurteilung

Reihenfolge jeweils: Name, Beurteilung, Punkte

Ik wenig	68
Ai viel, wenig Fehler, manchmal ohne Verb	249
Ar manchmal schief	138
Ta problematisch	105
To ohne Dialog, sehr wenig	41
Sh1 English beigemischt, wenig Variation	72
Ko interessanter Dialog	153
In wenig, fast nur SS	78
Wo sehr wenig trotz viel Hilfe!	72 erschreckend!
Oc kurz und sehr gut, WS viel	138
Ki1 kurz und viel, gut!	184
Sh2 Fast nur SS	155
Ki2 kurz, kein Perfekt!	128
Ur etwas schwach, wenig Fehler!	85
Ish1 schnell, nicht so sehr viel, aber fast alles Gemachte vorbildlich!	Schnell viele Punkte: 275
Ta etwas schiefe Fehler, beispielhaft für diesen Typ	110

Ok hat eigentlich schon 4 Stufe gemacht 131
Ka viel und gut! 252
An gut und viel 187
Ish2 als Vorbild? Einige Teile nicht! 191
Ki3 etwas schief 220

Bei den Kommentaren lässt sich feststellen, dass es oft um die Länge bzw. Kürze geht und wieviel in Relation dazu gut oder weniger gut geboten wurde.

Auffällig sind auch Beispiele mit Stärke- und Schwäche- Kommentaren, und ob das WS berücksichtigt wurde. Schon an dieser Stelle wird klar, dass es weite Unterschiede in den Punktzahlen gibt.

3.3. Textbeispiele für Einzelanalysen

Um den Unterschied zwischen Schnellgesamtbeurteilung und Detailbeurteilung darzustellen und nachvollziehbar machen zu können, geben wir zuerst zwei Texte (Ik und Ur) aus dem unteren Punktebereich der Texte in 3.2, hier vorerst noch unanalysiert. Dieses Vorgehen kann in der Wirklichkeit aus Zeitgründen leider erst nach dem Semesterabschluss erfolgen.

Beide Texte sind vom handschriftlichen Original unkorrigiert übertragen, lediglich zur Indizierung verschiedener Sprecher, wie sie auch im Original vorhanden sind, wurden Einrückungen gemacht.

3.3.1. Text A: Ik 68

Was haben Sie?
Ich habe einen Tischein Bed, und ein Uhr.
Aber keinen Bücher.
Was machen Sie in der Freizeit.
Ich lese Romane
Was machen Sie in Merch?
Ich gehe nach Y mit meiner Freundin
Und denn gehe vor "C".
Wozu?
Ich möchte lecker noodeln essen.
Wer Romane?
M T.
Wie alt sin Sie?
Ahtzehn, aber morgen ich bin neunzehen.
Was machen Sie?

Ich studiere Humanwissenschaft

Was haben Sie heute Abend?

Ich studiere Japanisch.

Notenpunkte: 68

3.3.2. Text B: Ur

0- Ich heiÙe C

Woher kommen Sie?

O-Aus T

Wo?

O-Mit K.

Entschuldigung du haben weiÙ pullover. Das ist gut.

O-Danke, ich habe das in T gekauft.

Was cost ist das?

200 Euro

O-Was will du in April mahrt?

ich will N fahren

O- Was will du maht?

ich will S burger essen

O-Sher gut!

Was haben sie Winter gemacht?

O-Anfang Janur habe ich mein Freund getroffen.

sie whonst in T. Wir haben KÙchen gegessen mit Zug.

O-Was KÙchen?

Chokolade und banane

0 Sher gut!!

Ich mÙchte Deuchland gehen. Und Sie?

0 Ich mÙchte H in Japan gehen?

Was mÙcht Sie machen?

0 Ich mÙchte Ski fahren. Du hast einmal Ski gemacht?

0 Noin. Ubur mochte ich Ski fahren.

Notenpunkte: 85

Wie Harsch & Martin (2012) gezeigt haben, sind holistische und kriteriengeleitete Beurteilungen nicht allzu unterschiedlich in ihren Ergebnissen für die Teilnehmer und können sogar teilweise als äquivalent angesehen werden. Der folgenden Teil vier eruiert, ob mit detaillierteren Untersuchungen nicht noch genauere Beurteilungen möglich sind.