Leonard Pon
Vladimir Karabalić
Sanja Cimer
(eds./Hrsg.)

Applied Linguistics Today:
Research and Perspectives

Angewandte Linguistik heute:
Forschung und Perspektiven

Proceedings from the CALS conference 2011
Beiträge von der KGAL-Konferenz 2011

PETER LANG
Internationaler Verlag der Wissenschaften

Originaler Forschungsbeitrag

Automatische Bewertung von Lernertexten
Ein Konzept und seine Anwendungen

Hans Jürgen Heringer

Zusammenfassung

Objektive Bewertungen schriftlicher Lernerproduktionen sind ein notorisches Problem von Sprachprüfungen. Menschliche Rater korrelieren hier allgemein nur mit 0.62.

Dennoch folgt die automatische Analyse im ersten Schritt der Idee: Nachbildung menschlicher Rater. So hat sie ein erstes Außenkriterium für die Eichung.

Das vorzustellende Konzept bezieht ausschließlich textuelle Parameter ein. Es unterscheidet:

- lokale oder textinterne Parameter und
- externe Parameter, die auf Datenbanken basieren, die aus großen Korpora gewonnen sind.

Die einzelnen Parameter werden definitorisch vorgestellt.

Praktische Anwendungen existieren für verschiedene Deutschprüfungen (Zertifikat Deutsch, DSH, TestDaF).

Hier wird das Konzept für Englisch (Niveau B1) vorgeführt.

Am Ende bleibt die Frage: Wie weit kann es gelingen, die automatische Bewertung als Bewertung sui generis zu etablieren?

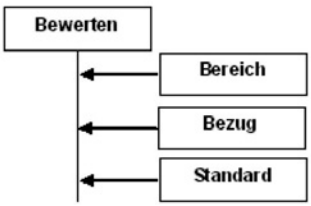
Schlüsselwörter: Automatische Bewertung; Lernertext; Bewertungskriterien; textuelle Parameter.

0.

In diesem Beitrag geht es um die Bewertung von Lernertexten im Bereich diverser Sprachprüfungen. Hierbei wird einer schriftlichen Produktion – in der Regel durch zwei Prüfer – ein Punktwert zugeordnet, der in die Punktwertung der übrigen drei Bereiche Leseverstehen, Hörverstehen und mündliche Kommunikation eingerechnet wird. Die Gewichtung der Teilbereiche mag dabei unterschiedlich sein. Diese Zuordnung eines Punktwerts wird auch mit Scoring bezeichnet und darf nicht verwechselt werden mit einer didaktischen Korrektur, die etwa an den Lernerautor adressiert ist und ihm Möglichkeiten zur Verbesserung seiner Kompetenz aufweist, oder Korrekturen, die dem Lehrer als Grundlage für didaktische Maßnahmen dienen.

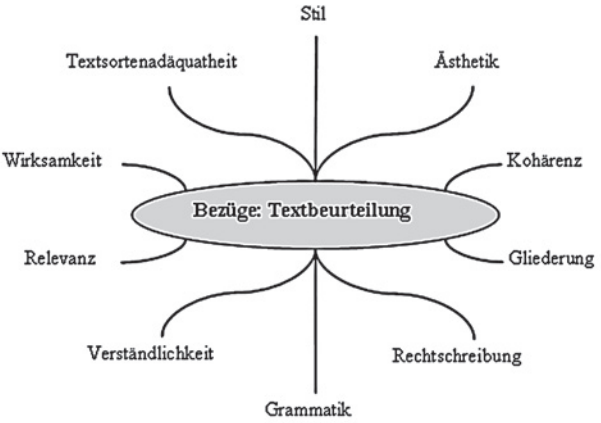
1. Bewerten

Etwas zu bewerten ist eine Art Allerweltsakt, den wir täglich und häufig still vollziehen. Wir können unsere Wertungen auch in bewertenden Sprechakten öffentlich machen, so wenn wir etwa sagen „Dieses Omelett schmeckt mir“. In unserem Zusammenhang und allgemein ist aber wichtig zu unterscheiden zwischen solchen gustatorischen Bewertungen, gegen die schwerlich etwas vorzubringen ist, und eher allgemeinen oder allgemein gültigen Bewertungen, die deskriptiv fundiert sein müssen. Um solcher Art Bewertungen wird es hier gehen. Der Akt des Bewertens hat einen inneren Aufbau, eine Binnengliederung, die wichtige Eigenschaften aufzeigt. Wenn wir von etwas sagen „Er ist schnell“, so ist es wichtig, von welcher Art das Bewertete ist. Ein Auto muss schon schneller sein als ein Mensch oder eine Schnecke, damit die Wertung gerechtfertigt wäre. Eine Bewertung bezieht sich also auf einen Bereich, der in der nominalen Formulierung oft expliziert wird: Ein schnelles Auto ist anders zu fundieren als ein schneller Wurm. Dies teilen solche wertenden Adjektive mit anderen relativen Adjektiven wie *groß*. So ist eine große Schlange natürlich größer als ein großer Wurm. Es kommt eben auf den Bereich an. Wenn wir etwas bewerten, so tun wir das immer in Bezug auf einen bestimmten Aspekt oder bestimmte Aspekte. Ich bewerte zum Beispiel ein Auto in Bezug auf sparsamen Verbrauch, in Bezug auf Preis, in Bezug auf Höchstgeschwindigkeit usw. Wenn ich etwa sage „Dieses Auto ist schnell“, bewerte ich es in Bezug auf Geschwindigkeit. Wenn ich sage „Dieses Auto ist geil“, so bleibt der Bezug implizit und vage, kann aber in der Sprechsituation durchaus gegeben sein. Eine dritte Komponente des Bewertens ist der Standard. Um eine Wertung wie „Dieses Auto ist schnell“ zu fundieren, muss man sozusagen wissen, über welchem Standard es liegen müsste. Der Standard kann natürlich relativ zum Bereich gesetzt sein. Er wird für einen Kleinwagen anders liegen als für einen Sportwagen, für Autos Baujahr 1910 anders als für solche 2010. Schematisch sähe die Aktstruktur so aus:



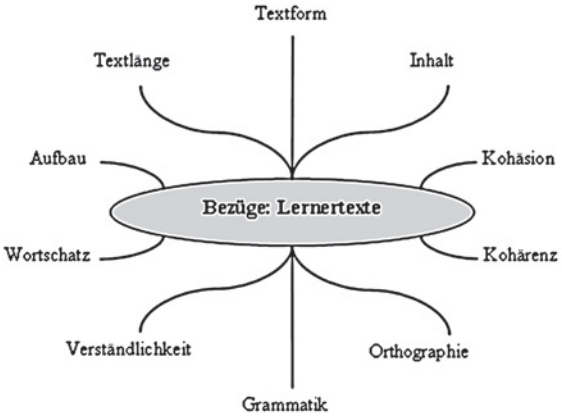
Grafik 1

Texte werden nun in Bezug auf vielerlei Aspekte bewertet. Gängige Bezüge zeigt die Grafik:



Grafik 2

Die Standards sind dabei in den wenigsten Fällen explizit oder explizierbar. Das macht Textbewertungen so heikel und das gilt auch für die Bewertung von Lernertexten, die allerdings unter anderen Aspekten beurteilt werden. Die Grafik zeigt einige gängige Bezüge für Lernertexte:



Grafik 3

Eine neuere Bewertungsideologie proklamiert, dass es hier weniger auf die Fehler ankommen sollte, die ein Lerner macht, als vielmehr auf das, was er schon kann. Dieser Ansatz ist wohl auch den Can-Do-Formulierungen des Common European Frameworks geschuldet, ist aber didaktisch bestenfalls motivationell zu begründen. Logisch ist der Ansatz verfehlt, weil aus einer korrekten Verwendung wissenschaftslogisch nie auf die Beherrschung einer Regel geschlossen werden kann. In dem kurzen Exkurs am Ende des Beitrags exemplifiziere ich das am Beispiel.

2. Sprachtests

Ein Sprachtest besteht wie jeder Test aus zwei Komponenten: einer Menge T von Testitems und einer Menge F von zu testenden Fähigkeiten:

$T = \{t_1, t_2, ... t_n\}$

$F = \{f_1, f_2, ... f_m\}$

Die einzelnen t_i müssen keine direkte Beziehung zu den f_i haben und in Intelligenztests kann man solcher Art Beziehung auch nicht erkennen. Für Sprachtests wäre das allerdings sehr unplausibel. Man wird ja kaum annehmen, dass man abtesten könne, ob jemand ein Idiom korrekt verwenden kann, indem man ihn geometrische Figuren zeichnen lässt. Und dennoch könnten bestimmte sprachliche Fähigkeiten (etwa orthographische) als Indikatoren für andere (etwa mündliche Kommunikation) taugen.[1]

Als Teil einer Prüfung muss das Scoring den Anforderungen an empirisch fundierte Tests allgemein genügen:

- Objektivität
- Reliabilität
- Validität

Objektiv ist ein Test, der nicht durch den Prüfer beeinflusst werden kann. Reliabel ist ein Test, der in der Wiederholung zu gleichen Ergebnissen führt. Valide ist ein Sprachtest, der genau die Fähigkeiten misst, die er zu messen beansprucht.

1 Für TestDaF wurde tatsächlich argumentiert, mündliche Kommunikationsfähigkeit sei abbildbar über C-Tests (Grotjahn 2007). Allerdings scheint mir die Korrelation von 0.64 doch eher bescheiden.

Mit allen drei Kriterien ist es nach meiner Meinung in bestehenden Sprachprüfungen nicht weit her. Am problematischsten ist unter linguistischen Gesichtspunkten die Validität, die aber für einen Test an erster Stelle zu stehen hätte. Die Standards der APA sind die weltweit anerkannten Standards für psychologisches und pädagogisches Testen. Sie fordern:

> When the validation rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified in reference to the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should also be clearly explained and justified.
>
> Comment: For example, test developers might provide a logical structure that maps the items on the test to the content domain, illustrating the relevance of each item and the adequacy with which the set of items represents the content domain. Areas of the content domain that are not included among the test items could be indicated as well. (APA-Standard 1.6)

Eine hinreichend detaillierte Beschreibung sprachlicher Kompetenz gibt es nicht – vielleicht kann es sie ob der Komplexität und der fundamentalen Eigenschaften menschlicher Sprachen nicht geben. Von einer detaillierten Beschreibung der Fähigkeiten F, die in Textproduktionen abgetestet werden sollen, sind die gängigen Prüfungen weit entfernt. Wenn in einer Sprachprüfung getestet werden soll, ob ein Kandidat einen persönlichen Brief schreiben kann, dann ist eine derartige Beschreibung einer Fähigkeit linguistisch hoffnungslos unterbestimmt. Denn welche sprachlichen Fähigkeiten braucht der Kandidat hierfür? Welche grammatischen, semantisch-lexikalischen oder stilistischen Kenntnisse? Analoges gilt für die Abbildung der einzelnen t_i auf einzelne f_i.

Eine automatische Bewertung wird ohne Weiteres objektiv und reliabel sein. Einzig die Validität und das sog. Außenkriterium werden hier zum Problem. Dieses Problem wird im vorgestellten Konzept umgangen, weil es nur beansprucht, die gleichen Scores zu generieren wie menschliche Korrektoren. Das Ziel ist also eine Art Simulation menschlicher Korrektoren. Somit ist für das Konzept absolut irrelevant, wie Korrekturen zu ihren Scores kommen, und ebenso, wie dem automatischen Algorithmus dies gelingt. Testtheoretisch und linguistisch ist beides aber von höchstem Interesse.

3. Das Konzept

In dem hier vorgestellten Konzept werden ausschließlich textuelle Parameter verwendet. Dies ist kein Manko der automatischen Analyse. Es ist vielmehr wis-

senschaftliche Hygiene, auf alle nicht operationalisierbaren Kriterien wie Inhalt, Idiomatik usw. zu verzichten. Nicht nur, weil sie nicht operationalisierbar wären, sondern vor allem, weil wir einer unhaltbaren sprachtheoretischen Ideologie aufsitzen würden, nach der der Inhalt irgendwo anders sein sollte als im Text. Um einen Eindruck zu geben, stelle ich einige der ursprünglich 25 Parameter vor und kommentiere sie.[2] Lokale Parameter sind rein bezogen auf den Lernertext:

- Textlänge
  Anzahl aller Wortformenvorkommen des Textes
  Die Textlänge spielt in den meisten Tests eine Rolle, da auf dem jeweiligen Niveau ein unterschiedlich langer Text vom Probanden verlangt wird. Prinzipiell ist unbestreitbar, dass der Text eine gewisse Länge haben muss, damit aus ihm auf die Kompetenz geschlossen werden kann. Wie lang der Text allerdings sein müsste, darüber ist empirisch wenig bekannt. Bemerkenswert ist, dass menschliche Rater längere Texte höher bewerten (eigene Untersuchungen, auch Attali/Burstein 2006: 5). Die Angabe der Textlänge in Wortformenokkurrenzen ist üblich, denkbar wäre auch die Zeichenzahl. Es ist plausibel, dass eine maschinelle Auszählung der Textwörter nach Lösung kleinerer technischer Probleme verlässlicher ist als das Zählen menschlicher Rater.
- Wiederholrate
  Quotient aus Anzahl der Vorkommen aller Wortformen und Anzahl aller Wortformentypes (= token/type ratio)
  Dieses Kriterium wird intuitiv von menschlichen Ratern genutzt und scheint auch in Bewertungskatalogen anerkannt. Wer sozusagen die gleichen Wortformen öfter verwendet, zeigt natürlich weniger von seiner Kompetenz. Aber menschliche Rater werden den Parameter wohl kaum verlässlich auswerten können. Da ist die Maschine wesentlich verlässlicher.

Externe Parameter nehmen Bezug auf Datenbanken der Basissprache:

- Fehler
  Zahl inkorrekter Wortformen(vorkommen)
  Ein wichtiger Parameter ist die Ermittlung von Fehlern. Fehler sind für sich ein weites Feld. Schon die Definition, was ein Fehler ist, bleibt umstritten. Eine flexible und angemessene Fehlertypologie ist nach wie vor ein Desiderat.

2 Für die weiteren Parameter Frey/Heringer 2007.

- Morphologische Tiefe
  Durchschnittlicher Rang der 30 ranghöchsten korrekten Wortformen
  Der Parameter gleicht alle Wortformen des Textes ab mit einer Tabelle, in der Wortformen nach Frequenz geordnet und invers mit Rangwerten versehen sind. Es geht also darum, zu bewerten: Wie selten sind die Wortformen, die der Proband benützt (und also benützen kann)? Die Beschränkung auf die 30 seltensten ist willkürlich, könnte aber zur empirischen Anpassung variiert werden.
- K_rang
  Relative Häufigkeit von Kollokationspaaren (bi-gramme[3])
  Der Probandentext wird fortlaufend in Paare der Wortformenokkurrenzen zerlegt. Die Paare werden abgeglichen mit einer Tabelle solcher Pärchen, die mit Frequenzen aus einem großen Korpus (Textwörter $n = 10^8$) versehen sind. Der Parameter kann unterschiedlich ausgewertet werden: Kommt das Pärchen in der Tabelle überhaupt vor? Wie frequent ist es?
- TR_rang
  Relative Häufigkeit von Kollokationstripeln (tri-gramme)
  Der Probandentext wird fortlaufend in Tripel der Wortformenokkurrenzen zerlegt. Die Tripel werden abgeglichen mit einer Tabelle solcher Tripel, die mit Frequenzen versehen sind. Der Parameter kann unterschiedlich ausgewertet werden: Kommt das tri-gramm in der Tabelle überhaupt vor? Wie frequent ist es?

Gegenüber pauschalen Bewertungen des Vokabulars, wie es von menschlichen Ratern verlangt wird, kann das elektronische Programm wesentlich differenzierter erheben:

- Lexikalische Tiefe
  Durchschnittlicher Rang der 20 ranghöchsten Lemmas
  Der Parameter gleicht nach der Lemmatisierung alle Lemmas des Textes ab mit einer Tabelle, in der Wörter der Basissprache in der Grundform nach Frequenz geordnet und umgekehrt mit Rangwerten versehen sind. Es geht also darum, zu bewerten: Wie selten sind die Wörter, die der Proband benützt (und also benützen kann)? Die Beschränkung auf die 20 seltensten ist willkürlich, könnte aber zur empirischen Anpassung variiert werden.

3 n-gramme werden in verschiedenen Zusammenhängen verwendet. Eine frühe Anwendung von Cavnar/Trenkle: http://citeseer.ist.psu.edu/cache/papers/cs/810/ und http:zSzzSzwww.info.unicaen.frzSz~giguetzSzclassifzSzcavnar_trenkle_ngram.pdf/n-gram-based-text.pdf, 20.11.2006.

- Lexikalische Elaboriertheit
  Durchschnitt aller Lemma-Ränge
  Auch dieser Parameter gleicht nach der Lemmatisierung alle Lemmas des Textes ab mit einer Tabelle, in der Wörter in der Grundform nach Frequenz geordnet und umgekehrt mit Rangwerten versehen sind. Es geht also darum, zu bewerten: Wie selten sind die Wörter, die der Proband benützt (und also benützen kann)? Die Wahl des Durchschnitts präferiert ein bestimmtes Rechenverfahren. Hier zeigt sich, dass über die verschiedenen Parameter empirisch wenig bekannt ist.
- Lexikalische Varianz
  Quotient aus Anzahl Lemmas der Inhaltswörter und Anzahl Lemmas der Inhaltswörterokkurrenzen
  Mit diesem Parameter soll überprüft werden, wie oft der Proband das gleiche Wort in unterschiedlichen Formen verwendet, landläufig gesprochen wiederholt.
- Lexikalische Ladung
  Quotient aus Anzahl Lemmas der Inhaltswörterwortformenokkurrenzen (sorry!) und Anzahl der Funktionswörterokkurrenzen
  Hiermit soll ermittelt werden, in welchem Verhältnis sozusagen inhaltsarme und somit auch hochfrequente Funktionswörter im Vergleich zu eher dicken Inhaltswörtern stehen. Dem liegt eine sicherlich diskutable Kategorisierung zu Grunde.

Thesen der folgenden Art bilden den Ausgangspunkt:

- Je größer die Wiederholrate, umso schlechter der Text
- Je mehr Fehler, umso schlechter der Text
- Je größer Lexikalische Tiefe, umso elaborierter der Text
- Je größer K_rang, umso elaborierter der Text
- Je größer TR_rang, umso elaborierter der Text

Letztlich geht es aber nur darum, die beste Korrelation mit den Scores der menschlichen Rater zu finden.

Automatische Essay-Scorer werden in den USA schon seit Jahren entwickelt, allerdings auf ganz unterschiedlicher Basis.[4]

4 Eine ganz frühe Arbeit bildet Page 1966. Bekannt ist beispielsweise der E-Rater. Einen Überblick gibt Dikli (2006).