

**Evelyn Frey / Hans Jürgen Heringer**

## **Prüfungstraining am Computer**

**Ein Projekt des Goethe-Instituts in Zusammenarbeit mit der Universität Augsburg und dem Hueber-Verlag**

Im Jahr 2002 entschloss man sich in der Prüfungszentrale des Goethe-Instituts, ein Projekt aufzulegen, das es den Prüfungskandidaten weltweit erlauben sollte, ihre Prüfungen am Computer abzulegen und diese auch auf elektronischem Wege zu trainieren. Es sollte sich bei dem angestrebten Produkt allerdings nicht nur um ein interaktives Training solcher Testformate handeln, die in einer schriftlichen Echtprüfung dem Kandidaten vorgelegt werden, sondern das Programm sollte wesentlich mehr können: es sollte freie schriftliche Lernerproduktionen (also den Prüfungsteil „Aufsatz“) automatisch bewerten können, so, als ob eine Lehrperson diese Arbeit bepunktet hätte.

Diese Zielsetzung war ein absolutes Novum. Alle bisherigen Versuche einer automatischen Textkorrektur bei anderen Prüfungsanbietern waren nicht erfolgreich. Sollte dieses Ziel erreicht werden, würde es sich also um ein absolutes Alleinstellungsmerkmal bei den Prüfungen des Goethe-Instituts handeln.

Um es vorweg zu nehmen: Wir haben unser Ziel erreicht! Die automatische Textkorrektur funktioniert und ist inzwischen in Form eines Prüfungstrainings auf CD-ROM auf dem Markt. Der Weg dahin war allerdings lang und erforderte intensive interinstitutionelle Forschungszusammenarbeit. Im vorliegenden Artikel wollen wir den Weg zur automatischen Textkorrektur sowie die Besonderheiten der Korrektursoftware beschreiben.

Das Goethe-Institut bietet derzeit 13 zentrale Prüfungen an, von denen einige zusammen mit anderen Partnerinstitutionen entwickelt wurden (zum Beispiel das „Zertifikat Deutsch“ mit Prüfungspartnern aus Österreich (Österreichisches Sprachdiplom Deutsch), der Schweiz (Lern- und Fortbildungszentrum für Sprachen der Universität Fribourg) und Deutschland (Weiterbildungs-Testsysteme GmbH) oder die „Prüfung Wirtschaftsdeutsch“ mit dem Deutschen Industrie- und Handelskammertag und den Carl-Duisberg-Centren). Damit ist das Goethe-Institut die einzige DaF-Mittlerorganisation, die Prüfungen auf allen Stufen des GERS (Gemeinsamer Europäischer Referenzrahmen für Sprachen) anbietet. Die meistverkaufte Prüfung ist das „Zertifikat Deutsch“ (ZD), eine Prüfung auf der Niveaustufe B1 des GERS. Aufgrund dieser hohen Reichweite der Prüfung wurde das ZD als Pilotprojekt für die Entwicklung eines elektronischen Prüfungstrainings und insbesondere der automatischen Textkorrektur ausgewählt. Andere Prüfungsniveaus sollen folgen.

Das ZD besteht aus einem schriftlichen und mündlichen Prüfungsteil. Der schriftliche Prüfungsteil wiederum, der in elektronischer Version angeboten werden sollte, besteht aus den Prüfungsteilen Leseverstehen, Hörverstehen und Schriftlicher Ausdruck. Jeder dieser Prüfungsteile besteht wiederum aus bis zu drei kleineren Prüfungsteilen, die jeweils unterschiedliche Fertigkeiten abtesten (zum Beispiel beim Leseverstehen: globales Verstehen vs. detailliertes Verstehen u.a.).

Im Prüfungsteil „Schriftlicher Ausdruck“ muss der Prüfungskandidat einen kleinen Aufsatz zu einem vorgegebenen Thema schreiben. Der/Die Kandidat/in erhält einen Inputtext (meist ein Brief eines Freundes oder einer Freundin), auf den er/sie schriftlich reagieren soll. Für den Antwortbrief, den der Kandidat schreiben soll, werden vier Inhaltspunkte vorgegeben, die vom Kandidaten in eine sinnvolle Reihenfolge gebracht werden müssen und die er inhaltlich mit jeweils ein bis zwei Sätzen abarbeiten muss. Hier ein Beispiel:<sup>1</sup>

Ihre 16-jährige Bekannte aus Deutschland schreibt Ihnen folgenden Brief:

*Münster, den ...*

*Liebe(r)...,*

*stell dir vor, ich habe eine ganz tolle Neuigkeit: es klappt nun doch, dass ich dich in deiner Heimat besuchen kann. Meine Tante will mir diese Reise bezahlen, weil ich in der Schule in diesem Jahr so gute Noten habe!*

*Ich habe aber noch so viele Fragen an dich: Wie ist das Wetter? Wo kann ich wohnen? Welche Kleidung brauche ich? Was können wir alles machen? Und und und...*

*Bitte gib mir bald eine Antwort auf meine Fragen und schreib mir alles, was ich vor meiner Abfahrt wissen muss!*

*Ganz herzliche Grüße aus Münster  
von deiner  
Martha*

Schreiben Sie Ihrer Bekannten einen Antwortbrief. Sie haben dazu 30 Minuten Zeit. Denken Sie daran: Schreiben Sie die richtige Anrede und einen passenden Schluss. Schreiben Sie auch das Datum. Schreiben Sie zu jedem der vier Leitpunkte etwas und überlegen Sie sich eine sinnvolle Reihenfolge:

- Was Sie zusammen unternehmen werden.
- Vorschlag zum Termin und ein Grund für diesen Termin.
- Wohin Ihre Bekannte kommen soll und bei wem sie wohnen wird.
- Was Ihre Bekannte unbedingt noch über Ihr Heimatland wissen sollte.

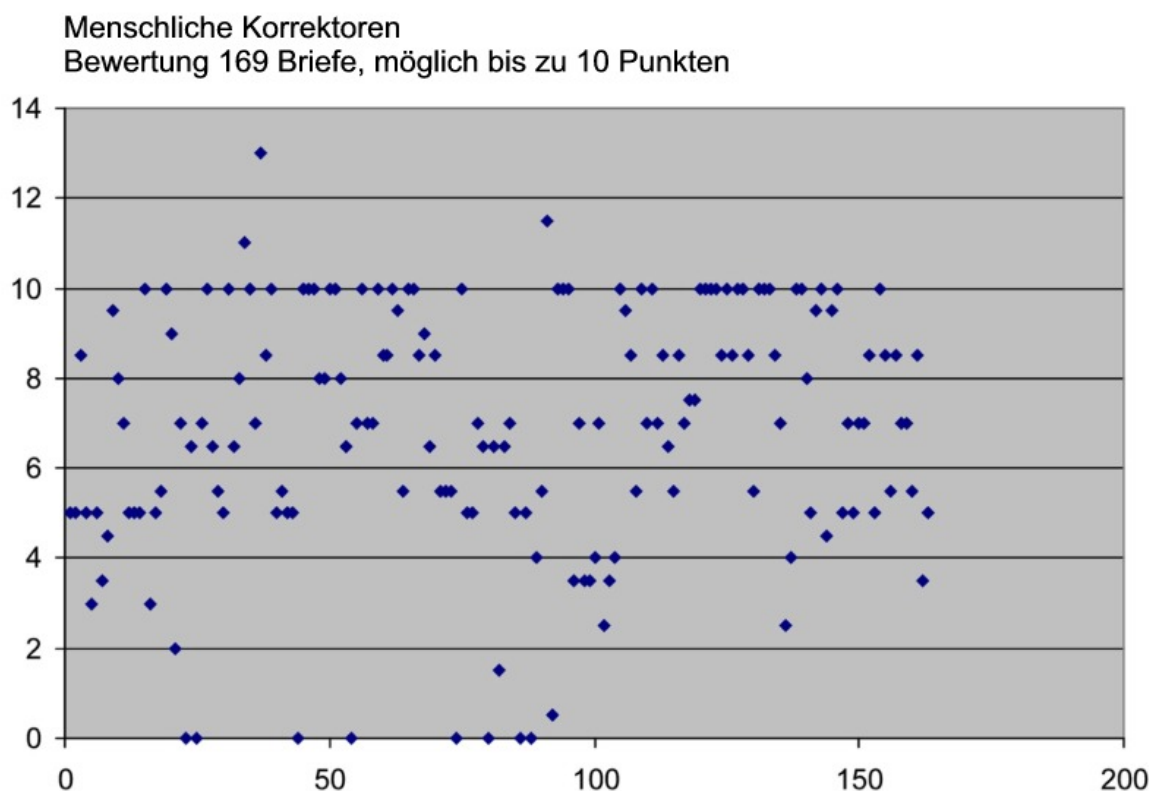
Der Aufsatz wird anschließend von zwei Korrektor(inn)en unabhängig voneinander bewertet. Im Zweifelsfall kann ein Drittkorrektor hinzugezogen werden. Die Korrektor(inn)en gehen bei der Bewertung nach exakt festgelegten Bewertungsmaßstäben vor. Diese Maßstäbe orientieren sich an der Erfüllung der Aufgabenstellung (Sind alle vier Inhaltspunkte in ausreichender Länge behandelt?), der grammatischen Korrektheit, der orthografischen Korrektheit und werten insbesondere auch solche Kriterien wie Kohärenz und Flüssigkeit (Wie gut sind die Sätze miteinander verbunden? Textaufbau? Satzaufbau?). Je nachdem, wie gut oder schlecht ein Kandidat/eine Kandidatin diese Kriterien erfüllt, kann er/sie zwischen 0 und 10 Punkten für den Aufsatz erhalten.

Ursprünglich hatten wir<sup>2</sup> gedacht, dass wir die Ergebnisse einer solchen Lehrerkorrektur als Außenkriterium für die maschinelle Korrektur ebenfalls anwenden könnten.

<sup>1</sup> Beispiel aus: Frey, Evelyn und Roland Dittrich. 2000. *Training Zertifikat Deutsch*. Ismaning: Hueber.

<sup>2</sup> „Wir“ das sind Priv. Doz. Dr. Evelyn Frey, Leiterin der Prüfungszentrale des Goethe-Instituts, und Prof. Dr. Hans Jürgen Heringer, Lehrstuhl für Deutsch als Fremd- und Zweitsprache und deutsche Philologie an der Universität Augsburg.

„Außenkriterium“ heißt, dass das Korrekturprogramm dann „gut“ wäre, wenn die maschinellen Punktwerte in signifikanter Weise mit den Lehrerpunkten übereinstimmen würden. Doch so einfach war die Sache nicht, denn in allen Vergleichskorrekturen, die wir vornahmen, war absolut keine Konsistenz zwischen den Ergebnissen der menschlichen Korrektor(inn)en festzustellen, die ausgereicht hätte, um daraus einen Eichmaßstab zu machen! Trotz intensiver Korrektorenschulungen und einheitlich vorgegebener Bewertungsmaßstäbe divergierten die Punktwerte bei ein- und derselben Prüfungsarbeit in ganz erheblichem Maß. Dies ging sogar so weit, dass statt der maximalen Punktzahl von 10 durchaus von mehreren Korrektor(inn)en bis zu 14 Punkten vergeben wurden. Hierzu eine kleine grafische Illustration:



Auf der x-Achse befinden sich die Probandenaufsätze, also von Nr. 1 bis 160. Auf der y-Achse sind die Punktwerte angegeben, die von den verschiedenen Korrektor(inn)en für ein- und dieselben Arbeiten vergeben wurden. Die Verteilungen scheinen nahezu beliebig, in keiner Weise sind irgendwelche Gaußschen Verteilungswerte zu erkennen. Dieser Befund war ziemlich ernüchternd und wir entschlossen uns, ein anderes Außenkriterium anzusetzen, den in Linguistenschulen altbewährten „native speaker“, der sich bald als hinreichendes Außenkriterium erwies.

Besonders schwierig war es, die Maschine dazu zu befähigen, Bewertungsgrößen wie „inhaltliche Korrektheit“, „Textkohärenz“ oder „Flüssigkeit“ in irgendeiner Weise adäquat abzubilden, so dass eine Maschinenbewertung überhaupt erfolgen konnte. Orthografische Regeln waren relativ leicht zu hinterlegen, aber bei grammatischer Korrektheit wird alles schon relativ kompliziert, denn *schönen Tag* mag zwar richtig sein, sofern der unbestimmte Artikel oder kein Artikel vorangeht, aber *\*der schönen Tag* ist unkorrekt. Wie soll das aber programmiert werden? Hier war die Augsburger Datenbank (Lehrstuhl Heringer) von entscheidender Bedeutung.

Die Heringer-Datenbank umfasst zzt. etwa 80 Millionen Textwörter, denen Type- und Tokenwerte zugeordnet sind, d.h. wir können an einem Wort via Datenbankzugriff in Sekundenschnelle erfahren, wie häufig es verwendet wird, ob es sich um eine grammatisch korrekte Form handelt usw. Damit war der erste Schritt zur Feststellung der grammatischen Korrektheit getan: Die Software wurde so programmiert, dass Dupel und Tripel (also Syntagmen von zwei oder drei zusammen gehörenden Wörtern) gleichzeitig ausgewertet werden. So kann eindeutig festgestellt werden, ob eine Endung grammatisch richtig oder falsch ist (also: *\*der schönen Tag vs der schöne Tag*) und ein entsprechender Punktwert kann der Auswertung zugeordnet werden.

Ein wesentlicher Teil der Forschungsarbeit bestand darin, Parameter zu finden, durch die die inhaltlichen Kriterien maschinell bewertet werden konnten. Es musste ja zum Beispiel auch eine Frage wie „Thema verfehlt oder nicht?“ von der Maschine richtig beantwortet und bepunktet werden. Auch dies war durch Zugriffe auf die Datenbank möglich. Zu jedem Prüfungsthema kann ein Kanon von Wörtern und Syntagmen erstellt werden (eine Art Lösungsschlüssel), die bei der Behandlung eines Themas obligatorisch vorkommen müssen. Darüber hinaus ist aufgrund der Verwendungshäufigkeit einer Probandenäußerung festzustellen, auf welchem sprachlichen Niveau sich seine Äußerung befindet. Wir konnten also die Software so programmieren, dass das Korrekturprogramm nicht nur grammatische und inhaltliche Korrektheit feststellen kann, sondern auch, auf welchem sprachlichen Niveau sich die Äußerungen befinden (hier: Ist der Text dem sprachlichen Niveau B1 angemessen?).

Ziemlich am Anfang der Forschungsarbeiten hatten wir einen Kanon von zwischenzeitlich 35 Parametern erarbeitet, nach denen ein Probandentext untersucht und bewertet werden kann. Hier einige Beispiele aus dem Gesamtkanon:

- **Textlänge:** Die Länge eines Textes spielt beim Prüfungsteil „Aufsatz“ natürlich eine erhebliche Rolle und es wird vom Kandidaten – je nach Niveaustufe – auch eine mehr oder weniger umfangreiche Textlänge verlangt (im ZD sollten es optimal etwa 120 Wörter sein). In der Software ist eine Mindestanzahl von Wörtern als Kill-Schwelle programmiert (40 Wörter); Texte unter dieser Gesamtlänge werden nicht bewertet.
- **Wiederholrate:** Das Korrekturprogramm erkennt, ob ein Kandidat immer nur dieselben Wörter verwendet oder ob er eine gewisse Varianz in der Lexik aufweist. Dies ist ein Indiz für sein Sprachniveau.
- **Fehler (inkorrekte Wortformen):** Bei der Ermittlung falscher Wortformen arbeitet das Programm mit einer Mischung aus einer umfangreichen Tabelle deutscher Wortformen und einer Basis an Grundvokabular, aus dem mögliche Wortformen erzeugt werden. Die orthografische Fehlersuche richtet sich nach der neuen Rechtschreibung.
- **W\_Komplexität:** Unter diesem Parameter fassen wir die Komplexität von Wörtern auf. In der Verständlichkeitsforschung wird er immer wieder als sicherer und stabiler Parameter verwendet. Das Programm stellt die Komplexität (den Umfang) eines Wortes fest und stellt so eine Korrelation zum sprachlichen Niveau des Kandidaten her.
- **Lexikalische und morphologische Tiefe:** Die Lemmata des Textes werden in Bezug auf Morphologie und Lexik mit einer Frequenztafel abgeglichen. So kann festgestellt werden, ob ein Kandidat auch mal seltenere Wortformen benützt und ob er auch über einem seiner Niveaustufe angemessenen Wortschatz verfügt und nicht nur die einfachsten und am häufigsten gebrauchten Wörter benützt.

- **Subordinationstiefe:** Mit diesem Parameter sind Aussagen über die syntaktische Komplexität möglich. Allerdings steht kein Parser zur Verfügung, der ein Maß für die syntaktische Komplexität liefern könnte, und deshalb messen wir ersatzweise die Häufigkeit der Subjunktionen (sowie deren morphologische und lexikalische Tiefe). Der Parameter stellt sich als recht zuverlässig heraus.

Hinzu kommen Parameter, die Kollokationspaare (binomisch) und Kollokationstripel (trinomisch) messen sowie auf Gliederung und Kohäsion und solche, die auf lexikalische Breite und Varianz bezogen sind.

Es stellte sich rasch heraus, dass eine ausreichend hohe Treffsicherheit bei den maschinellen Bewertungsergebnissen auf dem B1-Niveau bereits beim Durchlauf durch fünf bis sechs Kriterien zu erreichen war. Wir glichen die verschiedenen Parameter immer wieder miteinander ab, bis wir den Korrekturkanon für die B1 schließlich vorliegen hatten.

Nun bestand eine sehr hohe Schwierigkeit noch darin, diese Parameter so zu gewichten, dass die Maschinenkorrektur das erwartete Punkteergebnis erbrachte. Es sollten ja etwa – wie in der Menschenbewertung der Papier- und Bleistiftprüfungen – Kommafehler nicht so stark gewichtet werden wie ein Endungsfehler und dieser wiederum sollte nicht so stark gewichtet werden wie ein inhaltlicher Fehler. Dieser Aufgabe widmete sich Heringer in mehreren Monaten intensivster Forschungsarbeit, die mir unzähligen maschinellen Auswertungsversuchen einherging. Schließlich stand auch diese Formel, die eine exakte Gewichtung und zuverlässige Bepunktung zulässt und das ganze Programm konnte in die ersten Erprobungen gehen.

Diese verliefen – bis auf einige technische Probleme – recht erfolgreich und so konnten schließlich drei komplette Übungsprüfungen für das ZD programmiert werden. Diese liegen inzwischen auf CD-ROM vor und werden vom Hueber-Verlag in Kürze vertrieben. Auf der CD befinden sich drei Übungssätze CD mit Leseverstehen, Hörverstehen und Schriftlichem Ausdruck. Die Übungsprüfungen sind so programmiert, dass sie in derselben Zeit, die in der Echtprüfung zur Verfügung steht, gemacht werden sollten. Natürlich kann sich der Übende aber auf Wunsch auch so viel Zeit nehmen, wie er gerne möchte, was gerade bei den ersten Übungsdurchgängen von Vorteil ist, bis man mit dem technischen Ablauf und der Prüfung selbst vertraut ist. Auch den Prüfungsteil Hörverstehen hört der Teilnehmende genau sooft und in genau derselben Zeit wie in der Echtprüfung. Aber auch hier sind individuelle Trainingsdurchgänge möglich.

In den Prüfungsteilen Lese- und Hörverstehen erhält der Teilnehmende umgehend sein Ergebnis. Er kann sehen, welche Prüfungsaufgaben er richtig (oder falsch) gemacht hat und erfährt, welche Punktzahl er in einer richtigen Prüfung für seine Leistungen bekommen hätte. Beim Prüfungsteil „Schriftlicher Ausdruck“ kann sich der Übende nach Belieben – allerdings kostenpflichtig – Bewertungsdurchläufe einkaufen. Er kann also seinen Aufsatz beliebig oft schreiben und selbst korrigieren, bevor er ihn offiziell bewerten lässt. Dazu schickt er seinen Aufsatz elektronisch zum Bewertungssystem und erhält innerhalb weniger Sekunden sein Punkteergebnis zurück. Diesen Service kann er beliebig oft in Anspruch nehmen. Aufgrund dieser automatischen Textkorrektur ist es einem Probanden also möglich, sein sprachliches Niveau in Vorbereitung auf eine Prüfung auch im Prüfungsteil freier schriftlicher Ausdruck feststellen zu lassen. Das war bisher nicht möglich. Zur Bewertung war immer der Lehrer/die Lehrerin oder zumindest ein Muttersprachler nötig (der allerdings meist nicht über die Punktwerte Bescheid wusste).

Eine Folge-CD-ROM mit weiteren drei Übungsprüfungen zum ZD ist bereits in der Herstellung und wird in einigen Monaten auf den Markt kommen. Relativ weit fortgeschritten sind die Arbeiten für die Niveaustufe A2 (Prüfung „Start Deutsch 2“). Es ist geplant, auch für diese Prüfung möglichst bald eine Übungs-CD, ebenfalls wieder mit der Möglichkeit zur automatischen Textkorrektur, auf den Markt zu bringen. Mittelfristig ist der Einsatz des Programms auch online bei den Echtprüfungen geplant.