NAMEN UND WÖRTER

Freundschaftsgabe für
Josef Felixberger
zum 65. Geburtstag

Herausgegeben von

Gerald Bernhard, Dieter Kattenbusch und Peter Stein

Haus des Buches.
Verlag Christine Lindner
Regensburg 2003

Alfred Holl

Datenanalyseverfahren der Informatik (Data Mining) als Grundlage einer didaktischen Darstellung der französischen Verbalmorphologie

Abstract

Die Informatik kennt unter dem Namen Data Mining viele verschiedene Vorgehensweisen zur Datenanalyse, darunter Clusteranalyse-Methoden, die hier mit der Entwicklung eines neuen Algorithmus erstmals explizit für die Analyse morphologischer Systeme in der Linguistik eingesetzt werden. Voraussetzung für die Anwendung des vorgestellten automatisierbaren Verfahrens ist jedoch eine sehr gute linguistische Erschließung der untersuchten flektierenden Wortart der betrachteten Sprache, ansonsten ist es sprachunabhängig. Ergebnis ist eine von subjektiven Einflüssen weitgehend freie Liste morphologisch homogener Cluster, die rückläufig ähnliche, morphologisch analoge Vertreter der betreffenden Wortart enthalten. Es kann als Grundlage für eine didaktische Aufbereitung morphologischer Systeme dienen. Die Durchführung des Verfahrens wird an einem Ausschnitt der französischen Verbalmorphologie gezeigt.

1. Data Mining und Clusteranalyse

1.1 Datenanalyse morphologischer Systeme

„Data Mining bedeutet buchstäblich Schürfen oder Graben in Daten ... Die Ergebnisse lassen Muster in Daten erkennen, weswegen Data Mining auch als Datenmustererkennung übersetzt wird.“ (Alpar/Niedereichholz 2000: 3) Man sucht nach bisher nicht bekannten, versteckten Zusammenhängen und Ähnlichkeiten, deren Kenntnis einen wirtschaftlichen oder wissenschaftlichen Nutzen verspricht. Für die Disziplin der Datenanalyse finden sich statt des Ausdrucks „Data Mining“ – entsprechend der bekannten Instabilität der Terminologie in der Informatik – weitere, wie „Information Mining“ oder „Knowledge Discovery in Databases (KDD)“, ohne dass eine genaue definitorische Abgrenzung möglich wäre. Der gesamte Bereich umfasst heute eine Vielzahl unterschiedlicher, teils statistischer Methoden, die verbreitet zur Analyse großer Datenbestände eingesetzt werden, etwa im Marketing, um Regelmäßigkeiten im Kundenverhalten festzustellen oder um aus potentiellen Kunden besonders vielversprechende herauszufiltern.

Derartige Verfahren sind nicht neu; neu sind ihre Automatisierbarkeit und die Bezeichnung „Data Mining“. Jeder Naturwissenschaftler, der lange vor der Existenz moderner Computer aus seinen Beobachtungsdaten induktiv auf Zusammenhänge schloss und diese in mathematische Formeln fasste, betrieb Datenanalyse; ein vielfach mühseliges Unterfangen, wenn man etwa daran denkt, wie die berühmten Astronomen des frühen 17. Jahrhunderts von Hand aus umfangreichen Beobachtungstabellen die Bahngleichungen von Planeten herleiteten. Dasselbe tat jeder Linguist, der in Texten nach sich wiederholenden Mustern suchte und diese mit grammatischen Regeln beschrieb, und muss es weiterhin tun, denn nicht jede Art von Datenanalyse ist gut formalisierbar

und damit automatisierbar. Der spezifisch menschliche kreative Geistesblitz beim induktiven Erkennen von Zusammenhängen wird auf absehbare Zeit nicht ersetzbar sein. Das gilt besonders in der Linguistik, aber auch in den Naturwissenschaften: So antwortete der Physiker Friedrich Hund (1895–1997) in hohem Alter anlässlich eines Gastvortrages an der Universität Regensburg um das Jahr 1980 auf die Frage, wie ihm seine Hundsche Regel über Elektronenkonfigurationen in nicht-abgeschlossenen Schalen eingefallen sei, „durch Anstarren der Spektren“. (Holl 1997: 175) Wie wollte man diesen Vorgang automatisieren?

In diesem Beitrag erläutere ich gut automatisierbare Vorgehensweisen zur linguistischen Analyse flexionsmorphologischer Systeme und stelle sie in den Dienst der didaktischen Darstellung des französischen Verbalsystems. Ich beschränke mich von Anfang an auf eine einfache graphemische und synchrone Sichtweise, da von einem durchschnittlichen Sprachstudierenden keine eingehende Auseinandersetzung mit Phonologie und Sprachgeschichte erwartet werden kann. Was ich im folgenden für das Verb zeige, lässt sich ebenso auf alle anderen flektierenden Wortarten anwenden. In gleicher Weise sind meine Überlegungen auf andere mit Buchstabenschriften darstellbare Sprachen übertragbar.

Die vorgestellten Verfahren gehören zu der variantenreichen Gruppe der Clusteranalyse-Algorithmen. Unter Cluster versteht man in diesem Zusammenhang ganz einfach Mengen von Datensätzen. Ein Datensatz ist ein Tupel (eine Aneinanderreihung) zusammengehöriger Einzeldaten, im vorliegenden Zusammenhang etwa die Stammreihe eines Verbs oder ein Infinitiv mit dem numerischen Verweis auf ein Musterparadigma, ggf. mit Angabe der Bedeutung, falls der Konjugationstyp von ihr abhängt (z.B. bei *ressortir*). Ziel einer Clusteranalyse ist immer die Ermittlung von Clustern mit möglichst ähnlichen Elementen (Datensätzen), wobei Ähnlichkeit den jeweiligen Anforderungen entsprechend zu definieren ist.

Auch bisherige ausführliche Darstellungen von Verbalsystemen sind Produkte „manueller“ Clusteranalyse. Wird etwa zu jedem Musterverb eine Liste aller Verben mit den gleichen Unregelmäßigkeiten angegeben oder in einem Gesamtverzeichnis aller Verben zu jedem die Nummer des zugehörigen Musterverbs, so werden – im Sprachgebrauch des Data Mining – Cluster aus jeweils morphologisch analogen Verben gebildet. Morphologische Analogie bedeutet das Vorhandensein der gleichen morphologischen Eigenschaften, bei Verben also der gleichen Konjugationsmerkmale (Stammalternanzen und Personalendungen). Doch was nützen derartige Cluster unter einem sprachdidaktischen Blickwinkel? Sie sind so umfangreich und unstrukturiert, dass es unmöglich ist, sie auswendig zu lernen.

Sprachstudierende gehen in ihren Lernprozessen daher andere Wege. Sie wollen und müssen ihren Lernaufwand minimieren und würden am liebsten am Infinitiv Präsens Aktiv (im folgenden kurz Infinitiv) eines Verbs auch dessen Unregelmäßigkeiten ablesen können, so wie man tatsächlich den Konjugationstyp regelmäßiger Verben in romanischen Sprachen an der Infinitivendung erkennt. Ist nun aber mühselig ein Paradigma gelernt, so versucht der Lernende in einer ersten Stufe aus diesem Wissen größtmögliches Kapital zu schlagen und es auf weitere „ähnliche“ Verben anzuwenden, die er dann zu den gleichen morphologischen Eigenschaften „verurteilt“. Als tertium comparationis zur „Feststellung“ der Ähnlichkeit wird die rückläufige gewählt.

Die rückläufige Ähnlichkeit (Ausgangsgleichheit) manifestiert sich in einem gemeinsamen Ausgang im Infinitiv, wobei der Terminus „Ausgang“ hier nicht als morphologische Kategorie im Sinne von Endung verstanden wird, sondern ganz einfach als Graphemsequenz am Ende eines

Wortes, deren Länge jeweils pragmatisch festgelegt wird. Ein *n*-stelliger Ausgang sei definiert als Ausgang der Länge *n*, d.h. als die letzten (schließenden) *n* Buchstaben eines Infinitivs.

Die Strategie stimmt bei jedem Simplex und seinen Komposita (mit wenigen Ausnahmen), bei den regelmäßigen Verben und tatsächlich auch bei einem Teil der unregelmäßigen (im Frz. bei etwa einem Viertel (Holl 2002: 162)), aber eben nicht durchgängig, was eine erhebliche Fehlerquelle bedeutet. Beispielsweise haben *prendre* und *rendre* einen 6-stelligen gemeinsamen Ausgang, ohne aber morphologisch analog zu sein. Diese Strategie kann also hilfreich oder irreführend sein. Jedoch werden oft erst dann, wenn sie scheitert, in einer zweiten Stufe weitere Verben mit ihren Konjugationsbesonderheiten intensiv gelernt.

Man trifft hier auf zwei konkurrierende Formen von Ähnlichkeit, die keineswegs deckungsgleich sind: die rückläufige Ähnlichkeit des Infinitivs als lexikalischer Grundform und die morphologische Analogie. Sie implizieren sich gegenseitig nicht, insbesondere nicht die erste die zweite! Nun ist aber das der genannten Strategie zugrunde liegende analogische Denken (der Analogieschluss von partieller auf totale Ähnlichkeit) ein allgemeines, wesentliches – häufig unbewusstes – Grundprinzip menschlichen Lernens und Denkens. Es kann also nicht einfach ausgeschaltet werden, sondern man muss bewusst damit umgehen, der Sprachlernende ebenso wie der Sprachlehrende. Es ist am besten, den Sprachlernenden von vorn herein zu zeigen, in welchen Fällen rückläufige Ähnlichkeit morphologische Analogie impliziert und in welchen nicht. Das habe ich in Holl 2002: 151–158 ausführlich gezeigt.

Es ist daher der bisherigen Form der Clusteranalyse morphologischer Systeme, die für Nachschlagewerke weiterhin ihren Sinn behält, eine zweite hinzuzufügen, die ganz gezielt den Spezifika analogischen Denkens auf der Basis rückläufiger Ähnlichkeit didaktisch Rechnung trägt. Ziel ist die automatisierte explizite Ermittlung homogener Cluster ausgangsgleicher Verben (1.2). Ein Cluster heiße (flexions-)morphologisch homogen (im folgenden kurz homogen), wenn alle seine Verben morphologisch analog sind. Darauf aufbauend erfolgt eine „manuelle“ Nachbearbeitung (2.).

Für diese Form der Clusteranalyse wird die Menge aller Verben – in Gestalt ihrer Infinitive – nicht ungeordnet oder in der üblichen alphabetischen Sortierung von links betrachtet, sondern als rückläufig, also von rechts alphabetisch sortierte Menge. Dadurch werden ausgangsgleiche Infinitive benachbart angeordnet. Geordnete Grundmengen liegen in mir aus der Informatik bekannten Clusteranalysen nicht vor, so dass dort übliche Algorithmen nicht verwendet werden können.

Bei der daher notwendigen Neuentwicklung von Clusteranalyse-Verfahren ist generell zwischen zwei Grundtypen zu unterscheiden: „Bei den meisten Varianten wird so verfahren, dass entweder jedes zu gruppierende Objekt als ein Anfangscluster oder alle Objekte als ein Cluster gewählt werden. Danach werden die Anfangscluster zusammengefasst (agglomerativ) oder das alle Objekte umfassende Cluster aufgespalten (divisiv). In beiden Fällen geschieht das so, dass die Abstände zwischen den Elementen eines Clusters möglichst gering werden.“ (Alpar / Niedereichholz 2000: 11). Im vorliegenden Fall ist diese Abstandsbedingung sehr einfach; sie besagt, dass die letztlich ermittelten Cluster morphologisch analoge, ausgangsgleiche Verben enthalten. Die Strukturierung des allumfassenden Anfangsclusters wird als Top-down-Clusteranalyse bezeichnet (1.2), die Zusammenfassung ähnlicher einelementiger Anfangscluster als Bottom-up-Clusteranalyse (1.3). Die Ergebnisse beider Verfahren sind in ihren wesentlichen Teilen gleich, so dass die Nachbearbeitung in 2. auf einer weitgehend objektivierbaren Basis aufbauen kann.

1.2 Top-down-Clusteranalyse morphologischer Systeme

Diese Methode empfiehlt sich, wenn eine Sprache linguistisch bereits sehr gut erschlossen ist, d.h. wenn es ein rückläufiges Wörterbuch mit morphologischen Angaben gibt (selten, z.B. für das Russische Zaliznjak 1977; Juilland 1965 für das Französische ist phonetisch orientiert und daher für eine graphemische Auswertung nur sehr umständlich zu gebrauchen) oder alphabetische Verblisten (etwa in Langendorf 1984, Larousse 2000, Rousseau 2002, Willers 1984). In beiden Fällen muss vor der eigentlichen Datenanalyse eine Vorverarbeitung und Auswahl durchgeführt werden (cf. Alpar/Niedereichholz 2000: 6 f.). Aus rückläufigen Wörterbüchern sind die Verben zu extrahieren; lexikalische Grundformen mit gleichem Ausgang sind nämlich häufig nicht wortartenrein, d.h. am Ausgang ist die Wortart nicht zu erkennen, cf. Substantive wie z.B. *charpentier, glacier, mètre, navire*. Verblisten sind in rückläufige Sortierung zu überführen.

In der anschließenden eigentlichen Datenanalyse betrachtet man kurz gesagt Cluster von ausgangsgleichen Verben und untersucht sie auf morphologische Analogie. Ist ein Cluster morphologisch analog, ist man fertig. Andernfalls verlängert man den Ausgang um einen Buchstaben nach links, splittet das Cluster auf und verfährt wie oben weiter.

Formaler und ausführlicher lässt sich die Datenanalyse wie folgt beschreiben:

Erster Schritt: Es wird die gesamte vorgegebene Verbliste betrachtet. Alle Verben mit dem gleichen einstelligen Ausgang, d.h. mit dem gleichen letzten Buchstaben im Infinitiv, bilden jeweils ein Cluster. Die so konstruierten Cluster werden auf Homogenität untersucht. (In romanischen Sprachen findet man hier noch keine homogenen Cluster, sehr wohl aber im Schwedischen, wo alle Verben auf *–o, –y, –ä* homogen und regelmäßig sind.)

Zweiter Schritt: Es werden nur noch die Verben betrachtet, die im ersten Schritt nicht in homogenen Clustern enthalten waren. Alle Verben mit dem gleichen zweistelligen Ausgang, d.h. mit dem gleichen letzten und dem gleichen vorletzten Buchstaben im Infinitiv, bilden jeweils ein Cluster. Die so konstruierten Cluster werden auf Homogenität untersucht.

Allgemein: *n*-ter Schritt: Es werden nur noch die Verben betrachtet, die im *(n–1)*-ten Schritt nicht in homogenen Clustern enthalten waren. Alle Verben mit dem gleichen *n*-stelligen Ausgang, d.h. mit den gleichen schließenden *n* Buchstaben im Infinitiv bilden jeweils ein Cluster. Die so konstruierten Cluster werden auf Homogenität untersucht.

Die Zerlegung wird solange fortgeführt, bis nur noch homogene Cluster vorliegen. Diese sind nicht selten nur einelementig. Der Algorithmus terminiert auf jeden Fall, da jeder Infinitiv nur aus endlich vielen Buchstaben besteht. Am Ende liegt jedes Verb in genau einem homogenen Cluster von Verben mit ausgangsgleichem Infinitiv, und zwar in dem mit dem längstmöglichen Ausgang. Das nächsthöhere Cluster (mit einem gemeinsamen schließenden Buchstaben weniger) ist bereits inhomogen.

Die entstehende Struktur kann als Baum (in der Informatik immer von einem oben liegenden Wurzelknoten nach unten wachsend) dargestellt werden, in dem genau die Blätter (untenliegende Blattknoten ohne „Kinder“) homogen sind. Alle Nicht-Blattknoten (Knoten mit „Kindern“) sind inhomogen. Auf der *n*-ten Ebene des Baumes liegen genau die Cluster, deren Verben im Infinitiv *n* schließende Buchstaben gemeinsam haben.

Zeichnungen großer Bäume anzufertigen ist schwierig. Stattdessen bringt man alle Knoten eines Baumes in eine lineare Reihenfolge und kann ihn so textuell beschreiben. Zu diesem Zweck gibt es Baumdurchlaufverfahren, mit denen man alle Knoten eines Baumes in geordneter Weise nacheinander besuchen kann. Man denke sich dazu, dass man den Baum beim Wurzelknoten beginnend gegen den Uhrzeigersinn umfahre. Das Preorder-Verfahren ist intuitiv am einsichtigsten; es notiert jeden Knoten dann, wenn man auf diesem Weg das erste Mal an ihm vorbeikommt. Bei jedem Teilbaum besucht man also zunächst die Wurzel und dann von links nach rechts alle seine Teilbäume (cf. Puttkamer 1990: 202 f.). Alle Baumdurchlaufverfahren nennen die Blattknoten in der gleichen Reihenfolge von links nach rechts, was für die vorliegende Problemstellung den Vorteil bietet, dass die rückläufige Sortierung der Verben bei der Baumlinearisierung erhalten bleibt.

Als Beispiel wird in 5.1 das Ergebnis für die Verben auf *–rir* im Französischen gezeigt. Der Konjugationstyp wird durch die Stammreihe mit Inf.Prs., Ind.Prs.1.Sg., Ind.Prs.1.Pl., Ind.Prs.3.Pl., hist.Pf.1.Sg., Ptzp.Pf., Fut.1.Sg. angegeben. Jede Zeile trägt an der ersten Stelle die Nummer der Ebene des Baumes, die der Länge des gemeinsamen Ausgangs entspricht, an zweiter Stelle den Buchstaben H bzw. I für homogen bzw. inhomogen. Nur homogene Cluster enthalten Verblisten: Mehrelementige homogene Cluster haben vor der Liste ihrer Verben eine Überschriftszeile, einelementige nennen nur das jeweilige Verb. Aus Platzgründen werden Komposita weggelassen und Cluster regelmäßiger Verben, die bei der didaktischen Aufbereitung in 2.2 keine Rolle mehr spielen, komprimiert dargestellt (mit # markiert).

1.3 Bottom-up-Clusteranalyse morphologischer Systeme

Dasselbe Ergebnis wie in 1.2 ließe sich auch mit einem automatisierbaren Bottom-up-Verfahren erzielen, das aber inhaltlich nichts Neues bringt und formal weniger schön zu beschreiben ist als das Top-down-Vorgehen. Ich will mich daher nun einem schlecht automatisierbaren Bottom-up-Clusteranalyse-Verfahren zuwenden, wie es in der Praxis häufig vorkommt, wenn eine Sprache linguistisch noch nicht sehr gut erschlossen ist, d.h. wenn ein rückläufiges Wörterbuch mit detaillierten morphologischen Angaben und eine Verbliste mit Paradigmenzuordnung fehlen. Eine Variante dieses Verfahrens ist in Holl 2001: 201 für das Schwedische skizziert und dort praktiziert worden.

Um diese Methode verwenden zu können, müssen mindestens eine vollständige, nicht klassifizierende (keine Zuordnung zu Musterverben) Liste der unregelmäßigen Verben und ein rückläufiges Wörterbuch existieren. Jene wird rückläufig sortiert und jedes Verb darin mit diesem um ausgangsgleiche ergänzt. Man beginnt mit einem möglichst langen gemeinsamen Ausgang und verkürzt diesen buchstabenweise, bis man das erste ausgangsgleiche regelmäßige Verb gefunden hat. Den Cluster mit dieser Ausgangslänge vervollständigt man um alle hierher gehörigen Verben. Potentiell zu dieser Ergänzung heranzuziehende Einträge in einem rückläufigen Wörterbuch ohne grammatische Angaben müssen – je nach Vertrautheit des Linguisten mit der untersuchten Sprache – anhand eines ein- oder zweisprachigen Wörterbuchs auf ihre Wortart hin überprüft werden. So erhält man eine Verbliste, die auf homogene Cluster untersucht werden kann. Dieses Vorgehen erfordert viel Handarbeit, insbesondere auch bei der Beurteilung, ob die Unregelmäßigkeiten zweier Verben vollständig gleichartig sind, ob also

morphologische Analogie besteht.

Das Verfahren ist nicht formalisierbar, aber die erstellte rückläufige Verbliste mit ihren morphologischen Angaben erfüllt die Voraussetzungen der automatisierbaren Top-down-Analyse in 1.2 und kann damit als deren Input dienen. In dieser Liste fehlen zwar die Cluster ausgangsgleicher regelmäßiger Verben (in 5.1 mit # markiert), denn man hat ja nur die rückläufige Nachbarschaft der unregelmäßigen soweit nötig untersucht. Dies spielt aber keine Rolle, da gerade diese Cluster bei der Weiterbearbeitung in 2. ohnehin unberücksichtigt bleiben.

2. Weiterbearbeitung des analysierten Materials

Zu Beginn der Weiterbearbeitung liegt vor: Ein Baum mit genau den homogenen Clustern als Blattknoten, linearisiert als rückläufig sortierte Liste aller Verben – zumindest aller unregelmäßigen und ihrer rückläufig ähnlichen Umgebung - unter Angabe aller homogenen und inhomogenen Cluster ausgangsgleicher Verben. Dies ist eine weitgehend objektivierbare Grundlage für die beiden im folgenden erläuterten Weiterbearbeitungsmethoden, deren Ergebnisse im Detail durchaus subjektive Züge tragen können.

2.1 Reduktion auf die minimal nötige Information

Diese Methode reduziert das aufzulistende Material drastisch, einerseits durch die Beschränkung auf Stammreihen, andererseits durch einen Zuordnungsalgorithmus (rechtsbündiges longest matching) beliebiger Verben zu den aufgeführten Musterverben, der die Repräsentation homogener Cluster durch jeweils ein Element und die weitgehende Elimination regelmäßiger Verben ermöglicht. Die Resultate dienen als äußerst kompakte Nachschlagewerke. So ist die gesamte morphologische Beschreibung der synthetischen Verbalformen des Französischen in Holl 1988: 220–229 auf wenigen Seiten untergebracht. In 5.2 wird der Ausschnitt für die Verben auf *–rir* gezeigt.

Die Methode führt nicht zu einem absolut eindeutigen Ergebnis, sondern erlaubt geringfügige subjektive Varianten in Bezug auf ihre Optimierung. An dieser Stelle soll der Verweis auf eine Kurzbeschreibung in Holl 2001: 201–204 genügen, da eine solche Kompaktifizierung den in diesem Beitrag als vorrangig erklärten didaktischen Zwecken kaum dienen kann. Die Vorgehensweise soll aber hier gezielt genannt sein, um die Besonderheiten einer didaktischen Aufbereitung im Kontrast zu zeigen.

2.2 Didaktische Aufbereitung

2.2.1 Redundanzen: Im Gegensatz zu 2.1 müssen für den Lernprozess wichtige Redundanzen erhalten bleiben. Zur Vermeidung von Missverständnissen ist es für den Lernenden nützlich, alle unregelmäßigen Verben und alle Verben nicht produktiver Klassen explizit zu kennen und nicht erst bei der Lektüre von Texten zufällig darauf zu stoßen. Daher wird erstens jedes unregelmäßige

Simplex explizit genannt, nicht nur implizit mit einem Repräsentanten eines homogenen Clusters, z.B. *offrir* und *souffrir* im Cluster *–ffrir*. Komposita eines Verbs werden nur dann aufgeführt, wenn sie unterschiedlich flektieren, z.B. *(re)dire* vs. *maudire* (stammerweiternd) vs. *interdire* (regelmäßiger Imperativ). Zweitens finden sich alle Vertreter von nicht produktiven Verbklassen, auch wenn sie als regelmäßig klassifiziert werden, z.B. die Verben auf *–re* vom Typ *vendre*. Homogene Cluster regelmäßiger Verben produktiver Klassen (z.B. *–*rir* mit Stammerweiterung) lässt man wegfallen, da sie keine relevante Information beinhalten; das sind genau diejenigen, die der Algorithmus 1.3 erst gar nicht liefert. Homogene Cluster von Verben auf *–er* mit regelhaften Besonderheiten (z.B. *e / è-*, *é / è-*, *c / ç-*, *g / ge*-Alternanz) reduziert man am besten auf jeweils einen Repräsentanten.

2.2.2 Clusterzusammenlegung: Einelementige trivialerweise homogene Cluster sind für den Sprachlernenden uninteressant. Bei der Untersuchung benachbarter Cluster ergeben sich zwei Möglichkeiten: Gibt es ein rückläufig nahe verwandtes Cluster („langer“ gemeinsamer Ausgang), so empfiehlt sich die Zusammenlegung zu einem inhomogenen Cluster, dessen Kenntnis für den Sprachlernenden hilfreich ist. Sonst muss man es beim einelementigen Cluster belassen. Wie diese Bedingung im Detail auszuwerten ist, bleibt dem geschulten linguistischen Auge vorbehalten und trägt daher immer gewisse subjektive Züge. Homogene und inhomogene Cluster werden in der didaktischen Aufbereitung explizit genannt. Ihre Typisierung ist in 3. beschrieben.

2.2.3 Hervorhebung von Unregelmäßigkeiten: Die Konjugationsbesonderheiten werden bei jedem Verb mit einer Stammreihe angegeben; über sie hinausgehende Ausnahmen finden sich in einer Anmerkung. Bei jeder unregelmäßigen Verbalform werden Art und Ort der Unregelmäßigkeitsstelle typographisch markiert (siehe 4.).

3. Typisierung der homogenen und inhomogenen Cluster

Jedes Cluster wird in seiner Überschrift durch ein bis zwei Buchstaben am Zeilenanfang näher charakterisiert.

H: **H**omogenes Cluster aus verschiedenen Verben (Simplicia), von denen nur ein Repräsentant genannt ist.
Beispiele: Verben auf *–er* mit regelhaften Besonderheiten (z.B. *e / è-*, *é / è-*, *c / ç-*, *g / ge*-Alternanz).

HD: **H**omogenes Cluster aus verschiedenen Verben, deren Simplicia vollständig **d**etailliert sind. Findet seine Anwendung bei analogen unregelmäßigen Verben und bei unproduktiven Konjugationsklassen.
Beispiele: *–frir*, *–ouvrir*, *–indre*, *–ondre*.

H1: Cluster aus einem **einzigen** Verb (Simplex und ggf. Komposita); trivialerweise **h**omogen und vollständig detailliert. Diese Angabe wird nur gelegentlich zur Verdeutlichung verwendet.
Beispiele: *–égrer*, *–eyer*.

HK: **H**omogenes Cluster aus **K**omposita, deren Simplex nicht existiert; nicht detailliert.
Beispiele: *–crire*, *–cevoir*, *–quérir* (Simplex ungebräuchlich).

I: **I**nhomogenes Cluster aus verschiedenen Verben, deren Simplicia nicht vollständig detailliert

sind. Findet sich bei großen Clustern, deren regelmäßige Teilcluster nicht angegeben werden. Beispiele: *–er*, *–cer*, *–ger*, *–ir*, *–rir*.

ID: **I**nhomogenes Cluster aus verschiedenen Verben, deren Simplicia vollständig **d**etailliert sind. Findet seine Anwendung bei unregelmäßigen Verben, bei unproduktiven Konjugationsklassen und bei Clustern, die unregelmäßige und regelmäßige Verben enthalten und so zu falschen Analogieschlüssen verleiten können.
Beispiele: *–érir*, *–ourir*, *–vrir*, *–rendre*, *–oudre*.

IK: **I**nhomogenes Cluster aus einem Simplex und den zugehörigen **K**omposita. Es werden mindestens die vom Simplex abweichend konjugierenden Komposita genannt, zur Vermeidung von Missverständnissen gelegentlich sogar alle.
Beispiele: *–soudre*, *–dire*, *–croître*, *–clure*.

4. Klassifizierung und Markierung von Unregelmäßigkeiten

Aus didaktischen Gründen muss die Anzahl der in den einzelnen Verbalformen der Stammreihen unterschiedenen und unterschiedlich markierten Typen von Unregelmäßigkeiten stark begrenzt werden. Eine zu große Vielfalt würde den Sprachlernenden verwirren. Das bisher in der einschlägigen Literatur gepflegte andere Extrem, nämlich pauschaler Fett- oder Farbdruck unregelmäßiger Formen, beinhaltet zu wenig Information. Es ist also ein didaktisch sinnvoller Mittelweg zu suchen. Dazu passen die geringe Zahl von im Layout gut erkennbaren typographischen Markierungsmöglichkeiten, wenn man die übliche schwarze Druckfarbe wählt und auf den zu teueren Vielfarbendruck verzichtet. Man kann in diesem Falle nämlich nur die Schriftattribute „fett“, „kursiv“ und „unterstrichen“ nutzen.

Die folgende Differenzierung von Unregelmäßigkeiten aus synchroner Sicht ist als ein Vorschlag zu verstehen, der mit den genannten einfachen typographischen Möglichkeiten markierbar ist und auch für den Sprachlernenden überschaubar bleibt. Ich gehe von den folgenden zwei Dimensionen aus: Unregelmäßigkeiten finden sich einerseits auf graphemischer oder phonetischer Ebene oder auf beiden und betreffen andererseits den Stamm (im Vergleich zum Infinitivstamm des jeweiligen Verbs) oder die Endung (im Vergleich zum Musterverb der Konjugationsklasse, im Französischen *aimer, finir, vendre*) oder beide. Die Entscheidung, wie welche Verbalform im einzelnen linguistisch beurteilt wird, ist in mehreren Fällen subjektiv.

Die beiden Dimensionen nenne ich – in Anlehnung an einen Sprachgebrauch in der Phonetik - Unregelmäßigkeitsart und Unregelmäßigkeitsort. Zur Bezeichnung des letzteren genügt die Stelle der Markierung, außer bei stark kontrahierten Verbalformen, bei denen Stamm und Endung nicht trennbar sind und die ich mit einem nachgestellten ◄ hervorhebe. Bei der ersten Dimension bezeichne ich graphemische Besonderheiten durch Unterstreichung, phonetische durch Fettdruck.

Diese Klassifizierung erweist sich als für Sprachlernende transparent, orthogonal, d.h. für ein und die selbe linguistische Situation gibt es genau eine Markierungsmöglichkeit, und konsistent, d.h. ein Schriftattribut bezeichnet immer die gleiche linguistische Situation.

Das Konzept wird nun an einigen Beispielen systematisch mit feinerer Klassifizierung veranschaulicht. Einige theoretisch mögliche Fälle treten nicht auf. Formen der Verbliste werden – wie in Abschnitt 5.3 – in `Courier` gesetzt.

1. Veränderte Graphie und unveränderte Phonie: unterstrichen.

1.1 Unregelmäßigkeitsort: Stamm; graphemische Veränderung im Vergleich zum Infinitivstamm.

1.1.1 Fehlen des folgenden Graphems im Vergleich zum Infinitiv bei unveränderter Phonie: kursiv und unterstrichen.
Beispiele: se*n*s, pa*r*s, so*r*s
im Vergleich zu: *sentir*, *partir*, *sortir*.

1.1.2 Phonetisch motivierte orthographische Besonderheit im Vergleich zum Infinitiv zum Zweck der Erhaltung der Infinitivlautung: fett (alternativ doppelt) unterstrichen.
Beispiele: pla**c**ons, mang**e**ons, vain**qu**ent
im Vergleich zu *placer*, *manger, vaincre*.

1.1.3 Graphemische Besonderheit im Vergleich zum Infinitiv ohne phonetische Motivation und ohne phonetischen Effekt: einfach unterstrichen.
Beispiel: plaît
im Vergleich zu *plaire*.

1.2 Unregelmäßigkeitsort: Endung; graphemische Veränderung im Vergleich zum Musterverb der Konjugationsklasse ohne phonetische Motivation und ohne phonetischen Effekt: einfach unterstrichen.
Beispiele: occlus, acquis◄
im Vergleich zu *vendu*, *fini*.

2. Unveränderte Graphie und veränderte Phonie: fett. Wird nicht markiert, wenn durch Anwendung graphophonemischer Regeln erkennbar, wie im Falle von *plaignons* ([ɛ] nicht nasal im Vgl. zum Infinitiv). Vokalharmonische Effekte werden auch nicht markiert (z.B. [ɛ] → [*e*] in *taisez-vous*).
Beispiel: f**ai**sons
im Vergleich zu *faire*.

3. Veränderte Graphie und veränderte Phonie: fett und unterstrichen.

3.1 Unregelmäßigkeitsort: Stamm; graphemische und phonetische Veränderung im Vergleich zum Infinitivstamm; bei Fehlen des folgenden Graphems zusätzlich kursiv.
Beispiele: acqu**ie**rs, fin**iss**ons, plai**gn**is, v**éc**u; cou**r***r*ai, mou**r***r*ai
im Vergleich zu *acquérir, finir, plaindre, vivre; courir, mourir*.

3.2 Unregelmäßigkeitsort: Endung; graphemische und phonetische Veränderung im Vergleich zum Musterverb der Konjugationsklasse.
Beispiele: offr**e**, cour**u**s, ouv**ert**, di**tes**, suiv**i**
im Vergleich zu *finis, finis, fini, vendez, vendu*.

3.3 Unregelmäßigkeitsort: Stamm und Endung in kontrahierter Verbalform; graphemische und phonetische Veränderung im Vergleich zum Infinitivstamm und im Vergleich zum Musterverb der Konjugationsklasse: nachgestelltes ◄. Mit Schriftattributen markiere ich die Unregelmäßigkeit häufig als eine, die die Endung betrifft.
Beispiele: hist.Pf. und Ptzp.Pf. l**u**s◄, lu◄; b**u**s◄, bu◄; cr**û**s◄, crû◄
im Vergleich zu *lire, boire, croître* und zu *vendis, vendu*.

5. Veranschaulichung der Verfahrensergebnisse anhand der Verben auf *-rir*

5.1 Ergebnis des Clusteranalyse-Algorithmus

3I **-rir**

4H# **-arir**: tarir; *regelmäßig*
4H# **-brir**: assombrir; *regelmäßig*
4H# **-drir**: attendrir, amoindrir; *regelmäßig*

4I **-érir**
5H férir 3.S f**ie**rt - - - fér**u** -
5H chérir chér**i**s chér**iss**ons chér**iss**ent chéris chéri chérirai
5H périr pér**i**s pér**iss**ons pér**iss**ent péris péri périrai
5I **-uérir**
6H guérir guér**i**s guér**iss**ons guér**iss**ent guéris guéri guérirai
6H **-quérir**
quérir [*nur Infinitiv; ungebräuchlich; vernachlässigt*]
acquérir acqu**ie**rs acquérons acqu**iè**rent acquis◄ acquis◄ acqu**er**rai
ebenso: requérir, enquérir, (re)conquérir

4H **-frir**
offrir offr**e** offrons offrent offris off**ert** offrirai
souffrir souffr**e** souffrons souffrent souffris souff**ert** souffrirai

4H# **-grir**: aigrir, maigrir, rabougrir; *regelmäßig*
4H# **-orir**: endolorir; *regelmäßig*
4H# **-rrir**: barrir, équarrir, amerrir, terrir, aguerrir, nourrir, pourrir; *regelmäßig*
4H# **-trir**: flétrir, pétrir, meurtrir; *regelmäßig*

4I **-urir**
5H# **-eurir**: fleurir [*Nebenformen* florissait, florissant *vernachlässigt*], *sonst regelmäßig*
5H# **-hurir**: ahurir; *regelmäßig*
5H# **-mûrir**: mûrir; *regelmäßig*
5I **-ourir**
6H courir cours courons courent cour**u**s cour**u** cou**r**rai
6H mourir m**eu**rs mourons m**eu**rent mour**u**s m**ort**◄ mou**r**rai
5H **-surir**: surir; *regelmäßig*

4I **-vrir**
5I **-uvrir**
6H appauvrir appauvr**i**s -vr**iss**ons -vr**iss**ent -pauvris -pauvri -pauvrirai
6H **-ouvrir**
ouvrir ouvr**e** ouvrons ouvrent ouvris ouv**ert** ouvrirai
couvrir couvr**e** couvrons couvrent couvris couv**ert** couvrirai

5.2 Reduzierte, redundanzfreie Aufbereitung

férir	3.S f**ie**rt	-	-	-	fér**u**	-
acquérir	acqu**ie**rs	acquérons	acqu**iè**rent	acquis◄	acquis◄	acqu**er**rai
offrir	offr**e**	offrons	offrent	offris	off**ert**	offrirai
courir	cours	courons	courent	cour**u**s	cour**u**	cou**r**rai
mourir	m**eu**rs	mourons	m**eu**rent	mour**u**s	m**ort**◄	mou**r**rai
ouvrir	ouvr**e**	ouvrons	ouvrent	ouvris	ouv**ert**	ouvrirai

5.3 Didaktische Aufbereitung

I **-rir**

ID **-érir**

férir	3.S f**ie**rt	-	-	-	fér**u**	-
chérir	chér**i**s	chér**iss**ons	chér**iss**ent	chéris	chéri	chérirai
périr	pér**i**s	pér**iss**ons	pér**iss**ent	péris	péri	périrai

ID **-uérir**

guérir	guér**i**s	guér**iss**ons	guér**iss**ent	guéris	guéri	guérirai

HK **-quérir**

acquérir	acqu**ie**rs	acquérons	acqu**iè**rent	acquis◄	acquis◄	acqu**er**rai

HD **-frir**

offrir	offr**e**	offrons	offrent	offris	off**ert**	offrirai
souffrir	souffr**e**	souffrons	souffrent	souffris	souff**ert**	souffrirai

ID **-ourir**

courir	cours	courons	courent	cour**u**s	cour**u**	cou**r**rai
mourir	m**eu**rs	mourons	m**eu**rent	mour**u**s	m**ort**◄	mou**r**rai

ID **-vrir**

ID **-uvrir**

appauvrir	appauvr**i**s	-vr**iss**ons	-vr**iss**ent	-pauvris	-pauvri	-pauvrirai

HD **-ouvrir**

ouvrir	ouvr**e**	ouvrons	ouvrent	ouvris	ouv**ert**	ouvrirai
couvrir	couvr**e**	couvrons	couvrent	couvris	couv**ert**	couvrirai

6. Zusammenfassung

Es zeigt sich, dass Clusteranalyse-Verfahren sehr gut an die Anforderungen der Analyse flexionsmorphologischer Systeme natürlicher Sprachen angepasst und dort gewinnbringend eingesetzt werden können. Wesentlich ist, dass ein objektivierbares Zwischenergebnis erzielt wird (Baum mit den größtmöglichen ausgangsgleichen, homogenen Clustern als Blättern), auf das Weiterbearbeitungen aufsetzen können, die verschiedenen Zwecken dienen und einen subjektiven Interpretationsanteil beinhalten.

Bibliographie

Alpar, Paul; Niedereichholz, Joachim (2000): *Data Minig im praktischen Einsatz*, Braunschweig/Wiesbaden.

Holl, Alfred (1988): *Romanische Verbalmorphologie und relationentheoretische mathematische Linguistik – Axiomatisierung und algorithmische Anwendung des klassischen Wort- und Paradigma-Modells*, Tübingen [= Linguistische Arbeiten 216].

Holl, Alfred (1999): „Empirische Wirtschaftsinformatik und Erkenntnistheorie“, in: Jörg Becker et al. (ed.): *Wirtschaftsinformatik und Wissenschaftstheorie – Bestandsaufnahme und Perspektiven*, Wiesbaden, 165–207.

Holl, Alfred (2001): „The inflectional morphology of the Swedish verb with respect to reverse order: analogy, pattern verbs and their key forms“, in: *Arkiv för nordisk filologi* 116, 193–220.

Holl, Alfred (2002): „Licht und Schatten von Analogieschlüssen auf der Basis rückläufiger Ähnlichkeit in der Verbalmorphologie romanischer und germanischer Sprachen“, in: Heinemann, Sabine; Bernhard, Gerald; Kattenbusch, Dieter (ed.): *Roma et Romania – Festschrift Prof. Dr. Gerhard Ernst zum 65. Geburtstag*, Tübingen, 152–167.

Juilland, Alphonse (1965): *Dictionnaire inverse de la langue française*, Den Haag.

Langendorf, Dieter (1984): *Le nouveau Bescherelle. L'art de conjuguer*, Frankfurt.

Larousse (1980): *Larousse de la conjugaison*, Paris.

Puttkamer, Ewald von (1990): *Wie funktioniert das? Der Computer*, Mannheim.

Rousseau, Pascale (2002): *PONS Verbtabellen Französisch*, Stuttgart.

Willers, Hermann (1984): *Langenscheidts Verb-Tabellen Französisch*, Berlin.

Zaliznjak, Andrej A. (1977): *Grammatičeskij slovar' russkogo jazyka: slovoizmenenie*, Moskau.