Maschinenlesbare Textkorpora im Japanischen: Erstellung und Nutzung[1]

Heiko Narrog

1. Zur Einführung

Textkorpora, die beispielsweise dazu dienen, Belege für das Auftreten und die Verwendung bestimmter Wörter in Texten bestimmter Genres oder Autoren zu finden, sind wichtige Hilfsmittel für alle, die sich mit japanischer Sprache oder Literatur beschäftigen. Während das Durchsuchen umfangreicher Texte in Buchform sehr zeitraubend ist, darf man sich durch den Einsatz computativer Mittel Hilfestellung für eine rationellere Analyse von Texten erwarten.

Dieser Artikel soll eine Orientierung darüber geben, welche maschinenlesbaren Textkorpora, die für Studierende und Forschende interessant sein könnten, derzeit erhältlich sind, wie man selbst maschinenlesbare Texte, im folgenden auch „E-Texte", erstellen kann, und über Möglichkeiten informieren, wie der Computer für die Analyse solcher Texte eingesetzt werden kann. Dabei soll hier gleich Abbitte geleistet werden: Dieser Artikel kann keinesfalls Anspruch auf Vollständigkeit erheben. Dafür ist die Weiterentwicklung von Hard- und Software zu schnell und der Markt, vor allem der Free- und Sharewaremarkt, für den einzelnen zu unübersichtlich. Die getroffene Auswahl ist von den persönlichen Interessen des Autors gesteuert.

Die hier dargestellten Möglichkeiten beschränken sich nicht auf ein gängiges Betriebssystem, sondern beziehen sich auf alle drei bei der Arbeit an der Universität üblichen Systeme, DOS/Windows, UNIX und Macintosh in ihren japanischen Versionen. Nicht besprochen werden Datenbank-Programme. Es gibt eine Vielzahl davon; jedes arbeitet nach eigenen Parametern und verlangt daher eine Spezialisierung des Benutzers. Meist teuer in der Anschaffung (¥ 50 000~) haben die Programme durch ihren Aufbau vorgegebene Begrenzungen, aber viele von ihnen (beson-

[1] Bei meinen eigenen Versuchen zur Erstellung und Nutzung maschinenlesbarer Textkorpora habe ich von den folgenden Personen zahlreiche Hinweise erhalten, denen ich an dieser Stelle meinen herzlichen Dank aussprechen möchte: Suzuki Hideo, Kubouchi Tadao, Nitta Haruo (alle Universität Tōkyō), Nagase Mari (Jōsai Universität), Suzuki Tai (Ochanomizu Universität).

ders relationale Datenbanken) sind für die analytische Arbeit mit japanischen Textkorpora gut geeignet.

Bei von bestimmten Anbietern bereitgestellten E-Texten und Software wird im Artikel die e-mail-Adresse, falls keine vorhanden, die Fax-Nummer angegeben.

2. ERSTELLUNG EINES MASCHINENLESBAREN TEXTKORPUS

2.1 Fertige Texte

Im Mittelpunkt der Darstellung stehen „rohe", d.h. noch nicht zu bestimmten Zwecken ausgewertete Texte, die daher vielfältig verwendbar sind. Für die computerlinguistische Forschung sind aber auch zahlreiche Datensammlungen und Wörter- und Morphemwörterbücher erschienen, die hier nur exemplarisch erwähnt werden.

2.1.1 Texte des Gegenwartsjapanischen

Das Repertoire an größeren im Handel erhältlichen E-Text-Mengen aus dem Gegenwartsjapanischen, die für die Analyse mit dem Computer geeignet sind, beschränkt sich hauptsächlich auf Zeitungsjahrgänge. Der Grund hierfür liegt darin, daß die Urheberrechte von literarischen Werken, die im Druck erschienen sind, bei Verlagen liegen, die sich E-Texten gegenüber ablehnend verhalten. *Data novels* und *Electronic books* (EB) sind zumindest für den technisch weniger Versierten prinzipiell nur zum Lesen und evtl. Kopieren (EB) kleiner Textabschnitte geeignet (zu EB und Literatur auf CD-ROM siehe den vorausgehenden Beitrag von Wolfgang Schlecht).

Zeitungsjahrgänge auf CD-ROM

Die Texte sind nur über das Suchprogramm zugänglich, das in der Regel mit Inhaltsfeldern (Politik, Wirtschaft etc.), inhaltlichen Stichwörtern (Personennamen etc.) und Erscheinungsdatum arbeitet. Die einzelnen Artikel sind bereits mit diesen Stichwörtern markiert, d.h. es findet keine freie Wortsuche im Text statt. Eine begrenzte Anzahl der so gefundenen Texte (z.B. bis zu 10 bei der *Asahi Shinbun* und der *Mainichi Shinbun* pro Suche, und bis zu 3 bei der *Nihon Keizai Shinbun*) kann im Text-Format auf die Festplatte oder Floppy-Disk abgespeichert werden und ist somit frei benutzbar für Datenbank-Programme etc.

Texte über Online-Dienste

Innerhalb Japans haben für private Kommunikation und Informationsaustausch per Computer die kommerziellen Online-Dienste noch eine wesentlich größere Bedeutung als das Internet. Die zwei bei weitem größten Online-Dienste sind NIFTY-Serve und PC-Van, wobei letzterer von NEC eingerichtet wurde und sich vor allem an NEC-Benutzer wendet. Im folgenden werden daher kurz die Möglichkeiten vorgestellt, die NIFTY-Serve bietet.

NIFTY-Serve bietet Mail-Service, thematisch gegliederte Foren und kommerzielle Dienstleistungen. Für die Suche von Text-Material sind insbesondere die Foren und unter den Dienstleistungen die Datenbanken potentiell als Textquelle nutzbar. Datenbanken von allgemeinem Interesse werden u. a. von Zeitungsverlagen (wie Asahi, Mainichi, Yomiuri, Nikkei) und Nachrichtenagenturen (Kyōdō, Jiji, NHK) *online* angeboten. Es gibt dabei zum einen aktuelle Nachrichten, die billig (ca. ¥ 20 ~ ¥ 50 pro Minute) sind, aber nicht als Text kopiert und geladen werden können, und die eigentlichen Zeitungstext-Datenbanken, die einen Großteil der in der jeweiligen Zeitung in den letzten Jahren erschienenen Artikel enthalten. Sie sind etwas teurer (ca. ¥ 80 ~ ¥ 200 pro Minute), aber ihre Daten können vom Benutzer geladen werden. Bei der Benutzung muß man darauf achten, daß Datenbanken Befehle haben, die sich von denen, die sonst im Online-Dienst gültig sind, unterscheiden. Über die Datenbanken kann man sich zwar auf diese Weise authentische Zeitungsartikel laden, aber angesichts der Kosten, die dadurch entstehen, daß zusätzlich zu den Telefongebühren und den Benutzungsgebühren des Online-Dienstes auch die Zusatzgebühren für die Benutzung der Datenbank anfallen, dürfte die Anschaffung eines kompletten Zeitungs-Jahrgangs auf CD-ROM (s. o.) günstiger sein.

Über literarische Foren, so dem *bungaku fōramu* (FBUNGAKU), veröffentlichen Autoren ihre Kurzgeschichten, Romane, Gedichte usw. Es werden auch Online-Kettengedichte oder -Kurzgeschichten geschaffen. Zwar kann man sich hier kostengünstig authentische literarische Texte, die bereits maschinenlesbar sind, kopieren und dann zu seinen eigenen Zwecken weiterverwenden, aber die Texte können schwer von anderen, die keinen Zugang zum Online-Dienst haben, überprüft werden und sind im Gegensatz zu gedruckten und im Druck herausgegebenen Texten nicht von Verlagslektoren korrigiert, so daß sie auch Fehler enthalten können. Die literarische Aktivität ist übrigens noch reger auf dem kleineren ASAHI-net, auf dem es aber längst nicht so viel Freeware (s. u.) gibt wie auf NIFTY-Serve.

NIFTY-Serve verlangt für Normal-Benutzer eine Grundgebühr von ¥ 200 im Monat und ¥ 8 pro Minute. Der Online-Service ist auch im Aus-

land, u. a. über Internet und CompuServe preisgünstig zugänglich. FAX: ++81-3-5471-5890/5891 (ASAHI-net: Grundgebühr ¥ 1000, ¥ 10 pro Minute; FAX: ++81-3-5640-2951).

EDR und andere lexikalische Datenbanken

Im Bereich des Gegenwartsjapanischen finden sich auf diesem Gebiet die größten Datensammlungen für sprachwissenschaftliche Zwecke. Neben den aus ihren Druckausgaben bekannten Werken wie den IPAL-Wörterbüchern für Verben, Adjektive und Nomina vom Jōhō shori shinkō jigyō kyōkai gijutsu sentā (FAX: ++81-3-3437-9421) und der vom Kokuritsu kokugo kenyūsho erarbeiteten und beim Shūei Verlag (FAX: ++81-3-3260-5282) herausgegebenen *Bunrui goihyō* (der japanische Wortschatz in semantisch und nach Wortartkategorien gegliederten Wortfeldern), die auf zwei Floppy Disks ¥ 3500 kostet, dürften die EDR-Wörterbücher, die im UNIX-Format erstellt sind, von größtem Interesse sein. Sie bestehen aus acht Wörterbüchern in fünf Typen, die jeweils für die beiden Sprachen Englisch und Japanisch verfügbar sind: Einsprachige Wörterbücher (Japanisch: 250 000 Wörter, Englisch: 190 000 Wörter), zweisprachige Wörterbücher (E-J: 190 000 Wörter, J-E 230 000 Wörter), Begriffswörterbücher, Kookkurrenz-Wörterbücher und Technik-Wörterbücher (insbesondere Computer-Technik). Interessant als Textkorpus ist dabei vor allem der japanische Satz-Korpus, der 220 000 japanische Sätze umfaßt, deren Quelle angegeben ist, und für die jeweils eine vollständige morphologische, syntaktische (Dependenzstruktur) und semantische (Begriffsbeziehungen) Analyse gegeben wird.

Die CD-ROMs kosten als Set ¥ 9 Mio., pro Wörterbuch ¥ 1,2 Mio., für universitäre Einrichtungen aber nur ¥ 100, 000 pro Wörterbuch, an Privatpersonen erfolgt nach Auskunft des Herstellers kein Verkauf (e-mail: thoth@edr.co.jp).

2.1.2 Historische Texte

Genji monogatari [Die Geschichte des Prinzen Genji]

Die von Nagase Mari angefertigte maschinenlesbare Version der *Nihon koten bungaku zenshū*-Ausgabe (Shōgakkan) des *Genji monogatari* von Akiyama Ken *et al.* ist im Oxford Text Archive (OTA) und im Ōgata keisanki sentā der Universität Tōkyō gegen eine Schutzgebühr sowie Kopier- und Versandkosten erhältlich. Die ftp-Adresse des OTA lautet ota.ox.ac.uk, Login als „anonymous", Password: eigene Mail-Adresse.

Das *Genji monogatari* ist in diesem großen, internationalen literarischen Textarchiv übrigens der einzige japanische Text. Der Text ist im COCOA-

Format erstellt (vgl. 3.1), und eine englische Übersetzung von Edward Seidensticker (Tuttle Verlag) ist zum Vergleich beigefügt. Dabei enthalten die englische und die japanische Version Verweise auf die entsprechende Textstelle in der jeweils anderen Sprache. Die von Nagase Mari an beiden Stellen zur Verfügung gestellte Version ist von 1989 und enthält einige kleinere Fehler, die daraus resultieren, daß der Text mit OCR (s. u.) eingegeben wurde (zum Teil sind Seiten doppelt). Inzwischen arbeitet Nagase an einer Hypertext-Version. Sie soll neben den beiden oben genannten Texten auch die Übersetzung ins Gegenwartsjapanische, die französische Übersetzung von René Schiffer, Anmerkungen und Bildmaterial enthalten und so noch besser für die Forschung als auch den Unterricht einsetzbar werden (siehe NAGASE 1995).

Man'yōshū [Die Sammlung der zehntausend Gedichte]

Die von Yoshimura Makoto erstellte maschinenlesbare Version des *Man'yōshū* beruhte ursprünglich auf der Taschenbuch-Ausgabe von Kadokawa, aber inzwischen (in der Version 2.00) aus urheberrechtlichen Gründen direkt auf der Nishi-Honganji-Abschrift unter Vergleich mit anderen Handschriften. Sie enthält den kompletten Originaltext in *Kanji* (*Man'yōgana*), wobei eine Liste mit *Kanji*, die auf dem Computer nicht eingegeben und gelesen werden können, als getrennter File beigefügt ist, den Text im *kundoku-bun*, das heißt in der normalerweise aus gedruckten Textausgaben gewohnten Form, sowie den Text komplett in *Kana* und einen Index mit unterschiedlichen Schreibungen in den verschiedenen Handschriften und Druckausgaben. Diese *Man'yōshū*-Files sind kostenlos bei der JALLC (Jōhō shori gogaku bungaku kenkyūkai) erhältlich, e-mail: HTE70552@pcvan.or.jp.

Andere Texte beim JALLC

Das JALLC-Archiv bietet auch andere Texte und Materialien kostenlos zur Benutzung durch Studenten und Forscher an, hauptsächlich aber Texte aus der Edo-Zeit wie z. B. das *Sonezaki shinjū* [Doppelselbstmord in Sonezaki] von Chikamatsu Monzaemon in zwei Versionen, die auf Photo-Reproduktionen von Handschriften beuhen, *Oku no hosomichi* [Der schmale Pfad in den hohen Norden] von Matsuo Bashō und Werke von Ueda Akinari. Die Texte wurden jeweils von verschiedenen Personen erstellt, z. T. im Rahmen des Unterrichts an der Universität mit Hilfe von zahlreichen Studenten. Dabei handelt es sich zumeist nur um E-Text-Versionen einer bestimmten Handschrift oder Druckversion (e-mail: s. o.).

Hachidaishū [Die acht kaiserlichen Gedichtsammlungen]

Seit Sommer 1995 sind die als *Hachidaishū* bekannten kaiserlichen Gedichtsammlungen (*Kokin wakashū*, *Kōsen wakashū*, *Shūi wakashū*, *Kōshūi wakashū*, *Kin'yō wakashū*, *Shika wakashū*, *Senzai wakashū* und *Shinkokin wakashū*) als CD-ROM von Iwanami für ¥ 52,000 erhältlich. Dies stellt einen neuen Schritt in der Herausgabe literarischer Werke dar, da der Verlag gleichzeitig mit der Erarbeitung einer Druck-Neuausgabe dieser Werke (*Shin koten bungaku taikei*) die CD-ROM Version erstellt und herausgegeben hat. Die CD-ROM enthält neben den Texten auch ein leistungsfähiges Suchprogramm. Mit ihm kann man die Gedichte nach Nummer, Genre, Verfasser, Stichwörtern, Gedichtsanfängen und -enden, und Orts- und Personennamen absuchen. Zu den Gedichten selbst können Übersetzungen ins Gegenwartsjapanische und Erklärungen abgerufen werden. Neben dieser Suchart, die auf dem bereits nach den o. g. Parametern angefertigten Index beruht, ist auch eine freie Suche im ganzen Text möglich, bei der nach jedem beliebigen Wort oder jeder beliebigen Wortform gesucht werden kann. Die Suchergebnisse (d. h. die ganzen Gedichte) können problemlos im Textformat auf die Festplatte oder Floppy-Disk kopiert werden.

Alternativ gibt es die genannten Gedichtsammlungen plus einiger anderer Sammlungen wie dem *Shin chokusen wakashū* und der privaten Sammlung *Sanjūrokkasen* für ¥ 32,000 im Paket bei Benseisha (Adresse s. o.) mit dem zugehörigen Suchprogramm „Keiko-W", das zusätzlich ¥ 10,000 kostet.

Texte aus dem Benseisha-Archiv

Das Benseisha-Archiv enthält eine Reihe von Erzählungen, Tagebüchern, Gedichtsammlungen etc. von der Heian-Zeit bis zur Edo-Zeit, aus letzterer insbesondere Texte von Saikaku sowie Kanbun-Texte. Sie sind für den Individual-Benutzer zu einem Preis zwischen ¥ 2000 und ¥ 5000 und für Institutionen für jeweils etwas mehr als das vierfache zu haben. Bestellscheine liegen immer der von Benseisha herausgegebenen Zeitschrift *Jinbungaku to jōhō shori* bei. Adresse: Benseisha (BS Data-gakari), Nishi-Shinjuku 4–41–7-#104, Shinjuku-ku, Tōkyō 160. Hier zwei Beispiele:

a. *Hōjōki* [Aufzeichnungen aus einer Schilfklause]: Enthält den Original-Text aus dem *Daifuku kōjibon*, den Original-Text in *Hiragana-* statt in *Katakana*-Fassung sowie einen durch vereinheitlichte historische *Kana*-Schreibung, Punkte, Kommas, Trübungspunkte etc. interpretierten Text. Die Texte haben einheitliche Zeilennummern (Kimura Masanori).

b. *Ise monogatari* [Geschichten aus Ise]: Enthält den Original-Text, beruhend auf den *Tenpukubon*-Büchern sowie einen durch vereinheitlichte historische *Kana*-Schreibung, Punkte, Kommas, Trübungspunkte etc. interpretierten Text. Die Texte haben einheitliche Zeilennummern (Kimura Masanori).

Die Texte weisen je nach Herausgeber in ihrer Form geringfügige Unterschiede auf. Das Ziel ihrer Herausgabe liegt in erster Linie in der Anwendung von verschiedenen Suchprogrammen durch den Benutzer.

2.2 Eigenes Erstellen maschinenlesbarer Textkorpora

2.2.1 Eingabe

Beim eigenen Erstellen einer „maschinenlesbaren Bibliothek" sind v. a. zwei Methoden vorstellbar: 1. Erstellung durch Eingabe von Hand, 2. Eingabe mit OCR-Lesegeräten.

OCR (Optical Character Reader)-Lesegeräte sind für verschiedene OS, insbesondere DOS/Windows, OS/2 und Macintosh inzwischen schon für den Individualbenutzer erschwinglich geworden. Neben dem Lesegerät ist eine entsprechende Software erforderlich, die von zahlreichen Herstellern angeboten wird. Da der Text meist vom Computer erst als Image, d. h. mit großer Datenmenge, aufgenommen wird und das Programm aufwendig ist, ist ein größerer Arbeitsspeicher erforderlich.

Die OCR-Geräte für Japanisch machen auch Lesefehler. Diese müssen dann im nachhinein per Handeingabe verbessert werden. Fehlerquellen sind z. B. sich ähnelnde Schriftzeichen wie die *Hiragana ka* und *ga*. Andererseits können Kommata oder Punkte als Teile von Schriftzeichen gelesen werden oder *Furigana* nicht als solche erkannt werden. Am gefährlichsten sind jedoch Verwechslungen quasi-identischer Schriftzeichen, die bei der Korrektur auch mit dem Auge schwer voneinander zu unterscheiden sind, z. B. *Katakana* und *Hiragana he*, *Katakana* und *Hiragana ri* oder Abend *yuu* und *Katakana ta*. Solche Fehler als Spuren des Einlesens mit OCR sind z. B. auch im *Genji monogatari* (s. o.) enthalten. Für das Lesen stellen sie kein Problem dar, aber für die Suchprogramme. Der Gebrauch von OCR ist also mit Vorsicht zu genießen und verlangt eine gründliche nachträgliche Korrektur der eingelesenen Texte.

Ein OCR-Service wird übrigens auch über NIFTY-Serve (s. o.) von kommerziellen Anbietern unter der Rubrik *kagaku/gijutsu/hon'yaku* (Wissenschaft/Technik/Übersetzungen) im Hauptmenü angeboten. Reines Einlesen mit OCR kostet allerdings ca. ¥ 0,4 pro japanisches Schriftzeichen und

mit Korrektur ¥ 1,2, was für die meisten Privatbenutzer und universitären Einrichtungen zu teuer sein dürfte.

2.2.2 Urheberrechtliche Aspekte

Die größte Unsicherheit beim eigenen Erstellen maschinenlesbarer Textkorpora liegt jedoch im Urheberrecht. Erstellt man die elektronische Version eines im Handel befindlichen literarischen Texts und bringt diese Version vervielfältigt, z. B. in Form von Floppy-Disks oder über Internet, in den Umlauf, so begibt man sich eindeutig auf Kollisionskurs mit dem Urheberrecht. Dies in Verbindung mit der Befürchtung der Verlage, die Verkaufszahlen würden sinken, wenn ihre Bücher unkontrolliert als E-Texte zirkulierten, ist der wichtigste Grund dafür, daß es relativ wenige moderne Texte in maschinenlesbarer Form gibt.

Bei E-Text-Versionen literarischer Texte längst verstorbener Autoren wie Murasaki Shikibu oder Ueda Akinari liegen nämlich trotzdem noch Urheberrechte beim Herausgeber, Verlag und bei der Druckerei (NAGASE 1994: 20). Daher hat auch eine vom Kokubungaku kenkyū shiryōkan erarbeitete CD-ROM-Ausgabe des hundertbändigen *Nihon koten bungaku taikei* von Iwanami bis jetzt nicht erscheinen können. Je nach Verlag bestehen aber auch Unterschiede in der Haltung gegenüber maschinenlesbaren Versionen. Sie können den Verkauf gedruckter Ausgaben auch fördern, da die Studierenden bzw. die Forschenden nur selten mit der Floppy-Disk allein arbeiten werden. Es ist also notwendig, vor der Erstellung einer E-Text-Version, die man weiterverteilen möchte, mit dem Verlag in Verhandlungen zu treten. Bei historischen Texten kann man die Hürde des Urheberrechts auch umgehen, indem man aufgrund von Originalquellen eine eigene Textversion erstellt und dann als elektronisches Medium veröffentlicht.

Der Erstellung maschinenlesbarer Textkorpora für den rein privaten Gebrauch sind keine gesetzlichen Grenzen gesetzt. In einer Grauzone befindet sich der Gebrauch im Unterricht an der Universität. Er stellt zwar grundsätzlich eine Ausnahme des Schutzes der Urheberrechte dar, aber die Kopie vollständiger Texte oder Bücher ist auch hier verboten (vgl. NAGASE 1994: 21).

3. Nutzung maschinenlesbarer Textkorpora

3.1 Micro-OCP

Das bei weitem ausgereifteste und komplexeste Programm für die sprach- und literaturwissenschaftliche Analyse japanischer Texte ist die japanische Version des in Oxford entwickelten Micro-OCP (Oxford Concordance Program). Dieses Programm führt die folgenden drei Grundfunktionen durch: Erstellung von Indexen, Erstellung von Wortlisten und Erstellung von Konkordanzen, und kann daneben auch Wortschatzstatistiken erstellen. Die Konkordanz-Funktion dürfte meist die wichtigste sein, da man mit ihrer Hilfe ein oder mehrere im Text vorkommende Wörter suchen und im Textzusammenhang zusammen mit Zeilennummer und anderen frei bestimmbaren Parametern ausgeben lassen kann.

Von der Micro-OCP werden neben Texten, die keinerlei Referenzinformationen enthalten, auch Texte mit Referenzinformationen in bestimmten Formaten, v. a. aber dem COCOA-Format, bearbeitet. Das COCOA (*COunt and COncordance generation on Atlas*)-Format bedeutet einfach, daß zwei Buchstaben definiert werden, die Referenzinformationen, z. B. Seitenzahl, Zeilenzahl, Name des Autors, Kapitelüberschrift, handelnde/sprechende Person etc., einklammern und daher vom Micro-OCP-Programm nicht als Text gelesen werden. Mit Zeilennummern versieht das Programm den Text automatisch. In der Funktion der zwei Buchstaben zur Klammerung wird dabei oft < > gebraucht (z. B.: <T Kokoro> = der Name des Texts ist „Kokoro"). Die o. g. Referenzinformationen können dann als Parameter bei der Erstellung von Konkordanzen, Indexen etc. dienen. Die Möglichkeiten des Micro-OCP-Programms können daher auch am besten dann genutzt werden, wenn der Text mit solchen Referenzinformationen bearbeitet ist.

Das Programm ist komplex und verlangt daher eine gewisse Spezialisierung des Benutzers. Kleine Fehler bei der Befehlseingabe können zu falschen oder gar keinen Ergebnissen führen. Außerdem arbeitet es nicht hundertprozentig zuverlässig. Das größte Problem ist dabei die Worterkennung in japanischen Texten, die besonders bei der Erstellung von Indexen und Wortlisten relevant ist, da Wörter nicht wie im Englischen oder Deutschen durch Spatien getrennt sind. Dennoch handelt es sich beim Micro-OCP um das bis jetzt ausgereifteste Textanalyse-Programm für das Japanische. Es kostet ¥ 80 000 bei Okita denshi giken (FAX: ++81-3-3395-8608).

3.2 Freeware, Shareware

Freeware und Shareware ist kostenlose bzw. sehr billige Software, die vor allem über Online-Dienste und Freeware-/Shareware-Sammlungen, die im Handel verkauft werden, verteilt wird. Da dieser Markt völlig unübersichtlich ist, existiert auch zahlreiche andere als die hier vorgestellte Software, die für sprach- und literaturwissenschaftliche Zwecke interessant ist. Über Suchfunktionen und Stichwörter in den entsprechenden Datenbanken und -büchereien der Online-Dienste kann man sich die passende Software suchen. Im günstigsten Fall kann man über Foren und Konferenzen der Online-Dienste direkt in Kontakt mit Programmierern treten und sie um die Anfertigung einer Software, die genau für die eigenen Zwecke geeignet ist, bitten.

3.2.1 grep, sed, awk

grep, *sed* und *awk* sind ursprünglich Unterprogramme des Betriebssystems UNIX, haben aber ganz unterschiedliche Anwendungsbereiche. *grep* präsentiert sich als relativ einfaches und auch für Anfänger leicht zu meisterndes Kommando, das der Erkennung von Zeichenfolgen in Texten dient, und die als mit der in der Befehlszeile festgelegten Zeichenfolge als identisch erkannte Buchstabenfolgen in Zeileneinheiten auf den Bildschirm (oder in einen anderen File) ausgibt. So bedeutet der Befehl: „grep president poli.txt", daß das Programm alle Zeilen im File „poli.txt", die das Wort „president" enthalten, auf den Bildschirm ausgeben soll.

sed ist ein Editor, der sich aber von gewöhnlichen Editoren zum einen dadurch unterscheidet, daß die Befehle in einer Befehlszeile eingegeben und dann in einem Male ausgeführt werden, und nicht wie sonst schrittweise mit Hilfe von Dialogfenstern, und zum anderen dadurch, daß das Ergebnis nicht die Änderung des ursprünglichen Files ist, sondern daß ein völlig neuer File geschaffen wird. *sed* wird v. a. zur Änderung von Texten, d. h. zum Austauschen von Buchstabenfolgen und zum Löschen und Anfügen von Textteilen, benutzt. *sed* ist relativ komplex und erfordert eine gründliche Einarbeitung. Am interessantesten wird *sed* dann, wenn man selbst E-Texte erstellt und in größeren Mengen distribuiert. Mit den mit *sed* erstellten Programmen kann man dann z. B. Textversionen aktualisieren.

awk kann wie *grep* als einfache Befehlszeile benutzt werden, ist aber noch viel mehr, nämlich eine Programmiersprache, die in ihren Grundzügen der Programmiersprache C ähnelt. Da die Grundfunktion von *awk* wie bei *grep* das Auffinden von Schriftzeichenfolgen ist, kann es auch in der gleichen Funktion verwendet werden. Es bietet aber wesentlich mehr

Möglichkeiten. So kann *awk* nicht nur Zeilen als Grundeinheiten nehmen, sondern auch Felder (das vertikale Pendant in einem Text zur horizontalen Zeile), kann mit diversen Variablen und Symbolen arbeiten, die z. B. Zeilenanfang und -ende, Wiederholung von Schriftzeichenfolgen etc. bezeichnen, und rechnen.

Die japanische Version von *awk* heißt *jgawk*. Sie ist als Freeware sowohl für DOS wie auch Macintosh erhältlich. *grep*, *awk* und *sed* dürften die in Japan in der japanischen Sprach- und Literaturwissenschaft am meisten genutzten Programme bei der Textanalyse sein (zu *awk* vgl. auch NAKAMURA 1994).

3.2.2 Freeware für den Editor VZ

Für die Textverarbeitung im Japanischen existiert besonders viel Freeware in Form von Macros für Editoren. Editoren unterscheiden sich von anderen Textverarbeitungs- und Wapro (= Wordprozessor)-Programmen wie „Word" und „Ichitarō" vor allem dadurch, daß sie nicht für die druckfertige, d. h. voll formatierte Erstellung von Texten, sondern für die rationelle Eingabe, Änderung und Bearbeitung von Textdaten gedacht sind. Sie bieten auch mehr Möglichkeiten für die Erweiterung durch den Einbau selbsterstellter Zusatzprogramme (Macro) als solche Textverarbeitungs-Programme. Besonders viel Macro-Software auf dem Freeware-Markt gibt es für VZ und MIFES, auf NIFTY-Serve im Software-Forum FGAL unter *Sōgō*. Im folgenden soll exemplarisch Software vorgestellt werden, die für VZ erstellt wurde:

– MITUKE: einfaches Suchprogramm, das diejenigen Zeichenfolgen, die mit der gesuchten Zeichenfolge übereinstimmen, im Text automatisch mit Dreiecken ∇Δ kennzeichnet.

– LM und BUNBOG: mit LM (Line Marker) können Textabschnitte oder Wörter im Text individuell mit Dreiecken oder Klammern markiert und mit BUNBOG gesucht und automatisch markiert werden (entspricht MITUKE). Die markierten Stellen können automatisch numeriert, im Falle von BUNBOG mit der gesuchten Zeichenfolge als Überschrift versehen und nach Datum und Uhrzeit markiert und in einen neuen File (.lst) geschrieben werden. Von diesem File kann direkt in die entsprechende Stelle im ursprünglichen File zurückgesprungen werden.

– TUIBAMI, BAMISORT, BAMIJUMP: TUIBAMI erstellt automatisch eine Wortliste eines Text-Files, von dem dies gewünscht wird. Mit BAMISORT kann man diese Wortliste alternativ nach bestimmten Parametern (Häufigkeit, alphabetische Reihenfolge etc.) sortieren lassen und mit BAMIJUMP von bestimmten Wörtern in der Liste zu der Stelle im Text, an der sie auftauchen, springen. Hier besteht (wie bei allen mir bekannten

Programmen dieser Art) aber das Problem, daß im Augenblick noch keine fehlerfreie automatische Worterkennung möglich ist. Man muß die Wortliste also selbst nachträglich korrigieren.

3.3 Programmieren

Das Thema „Programmieren" kann ich nur kurz anreißen, da ich auf diesem Gebiet keine ausreichende Erfahrung habe. In den letzten Jahren ist die Programmiersprache C im Bereich Sprache, besonders automatische Text-Analyse, zusammen mit der Verbreitung von UNIX sehr stark aufgekommen. Bei MITA (1990: 180ff.) findet sich ein Beispiel für ein mit C erstelltes Suchprogramm, das im Ergebnis im wesentlichen *grep* (s. o.) entspricht. Die Frage ist dabei, ob man mit dem eigenen Programmieren so weit kommt, Programme zu erstellen, die über die bereits erhältlichen hinausgehen. Für Nicht-Informatiker und Nicht-Computer-Linguisten könnte der hierfür erforderliche Zeitaufwand zu groß sein.

Eine weitere Möglichkeit des Programmierens ist die Einarbeitung in die Macro-Sprache eines Editors wie VZ oder MIFES. Dies setzt natürlich eine längerfristige Entscheidung für eine bestimmte Software voraus, da auch das Programmieren mit Macros, will man es auf nutzbringendem Niveau betreiben, einen gewissen Zeitaufwand erfordert. Aufgrund des unterschiedlichen Aufbaus jeder Programmiersprache und Macro-Sprache ist die Entscheidung für eine Sprache ratsam.

4. ZUSAMMENFASSUNG UND AUSBLICK

Die in diesem Artikel gegebenen Informationen können nur einen ersten Überblick liefern. Die Möglichkeiten der Nutzung computativer Mittel für die Erschließung literarischer Texte im Japanischen sind längst noch nicht so weit entwickelt, wie sie es etwa im Fall des Englischen sind. Dies trifft insbesondere auf die Zahl der literarischen E-Texte und auf ihre Qualität zu. Neben den Schwierigkeiten, die das japanische Schriftsystem stellt, sind hierfür auch Probleme des Urheberrechts verantwortlich.

Für den Bereich Klassische Sprache und Literatur hat Nagase Mari mit ihrer Version des *Genji monogatari*, die bald als Hypertext erscheinen wird, Pionierarbeit geleistet. Es ist zu erwarten, daß auch andere klassische Texte in der Folge als Hypertext mit Anmerkungen, Kommentaren, Übersetzungen etc. erscheinen werden, in einer nicht nur für die Forschung, sondern auch für den Unterricht attraktiven Form. Für die moderne Literatur ist dies jedoch nicht abzusehen, solange sich die Denkweise der Verlage

nicht ändert. Hier werden wohl auf längere Sicht nur Zeitungs-Datenbanken *online* und auf CD-ROM als objektive Text-Quelle für die sprachwissenschaftliche Forschung dienen. Neben den Online-Diensten ist in Japan besonders seit 1995 auch Internet stark im Kommen. Dabei bleibt aber noch abzuwarten, was sich auf dem Gebiet der Textkorpora an interessanten Entwicklungen ergibt.

Der beste Weg zur Erschließung von E-Texten zu Forschungszwecken ist das eigene Erstellen von Programmen, die auf die persönlichen Zwecke genau zugeschneidert sind. Die Freeware *awk* kann wie ein einfacher Befehl benutzt werden, aber auch zur Programmierung mit vielfältigen Möglichkeiten dienen, und wird daher häufig zur Textanalyse im Japanischen verwendet. Ein komplettes Programm zur automatischen Textanalyse wird mit der japanischen Version der Micro-OCP angeboten. Seine Stärken kann man am besten dann nutzen, wenn man mit Texten arbeitet, die fortlaufend mit Referenzinformationen markiert sind. Außerdem bietet der Freeware-Markt vielfältige Möglichkeiten, die immer wieder neue Wege eröffnen.

Literaturverzeichnis

Allgemein

Bátori, István S., Lenders, Winfried und Wolfgang Putschke (Hg.) (1989): *Computational Linguistics/Computerlinguistik.* Berlin: Walter de Gruyter (= Handbücher der Sprach- und Kommunikationswissenschaften; 4).

Miyaji, Hiroshi *et al.* (Hg.) (1995): *Nihongo gaku* Vol. 14, 7. *Rinji zōkangō: Pasokon o tsukau nihongo kenkyū* [Sonderausgabe: Japanischforschung mit dem Computer]. Tōkyō: Meiji Shoin.

Toyoshima, Masayuki (1994): Denshika tekisuto no kokusaiteki kyōyū [Internationaler Gebrauch maschinenlesbarer Texte]. In: *Kokugo gaku* Nr. 178, S. 77–85.

Beispiele konkreter Arbeit an maschinenlesbaren Texten

Nagase, Mari (1995): *Genji Monogatari Hypertext no sakusei to kyōiku riyō no tame no kisoteki kenkyū* [Grundlagenforschung für die Erstellung und didaktische Nutzung einer Hypertext-Version des *Genji monogatari*]

(= Heisei rokunendo kagaku kenkyūhi hojokin kenkyū seika hōkokusho). Unveröffentlichter Forschungsbericht.

OGINO, Tsunao und Takehiro SHIONO (1994): Asahi dēta bēsu o riyō shita gengo kenkyū [Sprachwissenschaftliche Forschung mit Hilfe der Asahi Datenbank]. In: *Nihongo gaku* Vol. 15, 5, S. 28–39.

UEMURA, Kazumi (1995): Denshika tekisuto no sakusei to katsuyō – Akutagawa Ryūnosuke sakuhin o rei to shite [Erstellung und Nutzung maschinenlesbarer Texte am Beispiel von Werken Akutagawa Ryūnosukes]. In: *Jinbungaku to jōhō shori* Nr. 9, S. 64–71.

Zum Urheberrecht

NAGASE, Mari (1994): Kenkyū dēta bēsu no chosakuken to ryūtsū kankō [Das Urheberrecht und der Usus bei der Zirkulation von Datenbanken für die Forschung]. In: Benseisha Dēta Sentā (Hg.): *Jōhōka jidai no chosakuken* [Das Urheberrecht im Informationszeitalter], S. 20–26 (= Jinbungaku to jōhō shori; 5).

KIMURA, Takashi (1993): *Konpyūta, Maruchimedia to hōritsu* [Computer, Multimedia und Recht]. Tōkyō: Toraiekkususha.

NIFTY-Serve

SUZUKI, Yasuyuki (1994): *NIFTY-Serve dēta bēsu tettei katsuyō manyuaru* [Handbuch grundlegender Anwendungen für NIFTY-Serve Datenbanken]. Verbesserte Neuauflage. Tōkyō: HBJ Shuppankyoku.

Software und Programmieren

awk, grep, sed

HAYAMA, Hiroshi (1991): *Jitsuyō UNIX* [UNIX Praxis]. Tōkyō: ASCII Shuppankyoku.

NAKAMURA, Kazuo (1994): Kokugo kokubungaku ni okeru hurii sohutouea no katsuyō [Die Nutzung von Freeware für die japanische Sprach- und Literaturwissenschaft]. In: *Jinbungaku to jōhō shori* Nr. 3, S. 85–90.

TOMINAGA, Hiroyuki und Toshio UEMURA (1993): *awk de puroguramingu* [Programmieren mit *awk*]. Tōkyō: Ohmsha.

OCP

NAGASE, Mari und Hiroyuki NISHIMURA (1986): *Bunshō kaiseki nyūmon – OCP e no shōtai* [Einführung in die Textanalyse – eine Einladung zu OCP]. Tōkyō: Ohmsha.

Programmiersprache C

HAYASHI, Haruhiko (1992): *C gengo nyūmon. Sūpā biginā hen, shinia hen (1991), ōyōhen* [Einführung in die Programmiersprache C. Super-Anfänger-Ausgabe, Fortgeschrittenen-Ausgabe, Anwender-Ausgabe]. Tōkyō: Softbank.

MITA, Norihiro (^{2}1990): *Nyūmon C gengo* [Einführung in die Programmiersprache C]. Tōkyō: ASCII Shuppankyoku.