Linguistische Einheiten im Schreibprozess. Systematische Untersuchung von Planungs- und Redigiereinheiten

Dr. phil. Cerstin Mahlow

1 Problemstellung

Das Verfassen von schriftlichen Texten ist ein bedeutender Teil unserer Kommunikation. Schreiben erfordert die Koordination zahlreicher kognitiver, linguistischer und technischer Aspekte. Die *Schreibprozessforschung* untersucht die Produktion von Texten, dabei werden jedoch linguistische Einheiten und ihre Grenzen nicht berücksichtigt, wenn es etwa um die Erklärung von Pausen und *bursts* (Schreibeinheiten ohne Pausen) geht. Ebensowenig spielen solche Einheiten und ihre Komplexität zur Erklärung von Textplanungsaktivitäten und Redigieroperationen über das einzelne Wort hinaus eine Rolle.

Geschriebene Texte sind auch der Hauptuntersuchungsgegenstand in der *Linguistik*, der *Computerlinguistik* und den *Philologien*. Obwohl linguistische Einheiten bei der Definition von Textsorten und bei der Analyse von Diskursstrukturen in Texten eine entscheidende Rolle spielen, wird bislang jedoch nicht untersucht, wie Autoren solche Einheiten und Strukturen produzieren.

Die erwarteten Erkenntnisse des Forschungsprojektes erlauben für Text erstmals eine Verbindung von Prozess und Produkt auf linguistischer Ebene. Für die Theoriebildung innerhalb der Linguistik dominiert nicht die Zuordnung Satz-Text-Diskurs, sondern der Schreibprozess selbst bestimmt die Perspektive bei der Klassifikation von textunabhängigen und textspezifischen Einheiten. Für die Praxis liefert eine solche Systematisierung sowohl Grundlagen für die Entwicklung und Förderung von Schreibkompetenz als auch für die Entwicklung von Funktionen in Textbearbeitungsprogrammen. Zudem lässt sich aus der Operationalisierung von Versionen während der Textproduktion ein neuer Zugang zu kritischen Editionen ableiten, um auch hier den Produktionsprozess zu fokussieren.

Publizierte Texte von muttersprachlichen professionellen Autoren enthalten häufig trotz einer expliziten Korrekturphase grammatische Fehler, etwa Sätze ohne finites Verb, doppelte Wörter oder fehlerhafte Wortstellung. Es lässt sich zeigen, dass viele diese Fehler Nebeneffekte von Redigieroperationen sind. Wenn wir das Redigieren mit einem stilistischen oder semantischen Fokus genauer verstehen und mit den involvierten linguistischen Einheiten verbinden können, haben wir eine entscheidende Voraussetzung geschaffen, um Funktionen in Editoren implementieren zu können, die diese Redigierabsicht sehr einfach ausführen lassen und solche Nebeneffekte vermeiden.

2 Stand der Forschung

2.1 Schreibprozessforschung

In den späten 1970er Jahren rückt der *Prozess* des Schreibens in den Fokus. Das erste Modell zum Schreibprozess von Flower und Hayes [1981] wird später von Bereiter und Scardamalia [1987] und Hayes [1996, 2012] weiterentwickelt. Schreibprozessforschung beschäftigt sich anfangs intensiv mit der Beobachtung und Erklärung des Redigierens, wobei erst noch mit Stift und Papier gearbeitet wird [etwa Sommers 1980, Faigley und Witte 1981], später auch mit Computerprogrammen [Flinn 1987]. Linguistische Einheiten werden jedoch nur beiläufig (zum Beispiel als Redigieren auf Wortebene oder auf Satzebene) erwähnt und nicht zur Theoriebildung verwendet.
Durch Keystroke-Logging entstehen bei der Aufzeichnung des Schreibens mit elektronischen Werkzeugen ab den 1990er Jahren große Datenmengen, deren manuelle Inspektion nicht mehr möglich ist. Schreibprozessdaten können mittels Progressionsanalyse [Perrin 2002] als Pfad durch den entstehenden Text visualiert werden [Perrin und Wildi 2009, Ehrensberger-Dow und Perrin 2009]. Diese Repräsentation wird zur Entwicklung von Schreibstrategien verwendet, ist jedoch nicht zur Untersuchung linguistischer Einheiten geeignet. Standardnotation zur Kodierung des sich entwickelnden Textes ist die S-Notation [Severinson Eklundh 1994, Kollberg 1998], die allerdings nur auf Zeichenbasis operiert. Basierend auf S-Notation-Daten erstellte Taxonomien zu Redigieroperationen [Lindgren 2005] berücksichtigen keine linguistischen Einheiten. Baaijen et al. [2012] entwerfen eine komplexe Taxonomie von Pausen und bursts, berücksichtigen jedoch ebenfalls keine linguistischen Einheiten.
Leijten et al. [2012] sind bislang die einzigen, die computerlinguistische Werkzeuge verwenden, um S-Notation-Daten mit basalen linguistischen Informationen anzureichern. Dies erfolgt jedoch nur auf Wortebene, eine Untersuchung auf höherer Ebene ist so noch nicht möglich. Will et al. [2006] fokussieren zwar auf linguistische Einheiten, jedoch lediglich innerhalb eines Wortes. Die Probanden schreiben diktierte Wörter, untersucht wird das Pausenverhalten an Silben- und Morphemgrenzen.

2.2 Angewandte Linguistik

Internetbasierte Kommunikation gewinnt in den letzten Jahren als Forschungsgegenstand an Bedeutung. Aussagen zur Schreibkompetenz von Jugendlichen/Studierenden oder allgemeiner «digital natives» stützen sich jedoch auf die in verschiedenen Medien mit verschiedenen Werkzeugen produzierten *Texte* und die darin gefundenen Phänomene in Abgrenzung zu traditionellen Zeitungskorpora [etwa Dürscheid et al. 2010, Storrer 2014], der Entstehungsprozess wird ausgeblendet.
In Projekten, die aktuelle Alltagstexte untersuchen, wie «sms4science»[1] und «What's up Switzerland»[2], oder Texte von professionellen Schreibern, wie im «Projekt Schreibgebrauch»[3], werden lediglich die produzierten (und im Fall von Zeitungstexten auch professionell redigierten!) Texte aufbereitet und annotiert, der Schreibprozess selbst wird ausgeblendet.
Die Analyse von Diskursstrukturen und Koreferenzketten ist ein etabliertes Feld innerhalb der Linguistik und bezüglich der automatisierten Analyse auch der Computerlinguistik [zum Beispiel Hobbs 1978, Mann und Thompson 1983, Soricut und Marcu 2003], jedoch gibt es bislang keine Untersuchungen, wie diese Strukturen und Referenzketten produziert werden und welche Auswirkungen Textrevisionen auf sie haben.

2.3 Stand der eigenen Forschung

In meinem Dissertationsprojekt habe ich gezeigt, dass die Implementierung von sprachbewussten Funktionen in Textbearbeitungsprogrammen auf der Basis von computerlinguistischen Ressorucen und Erkenntnissen der Schreibprozessforschung prinzipiell möglich ist [Mahlow und Piotrowski 2008, 2009, Mahlow 2011]. Für die Klassifikation von Schreibfehlern, die durch Standardgrammatik- und -rechtschreibprüfung nicht zu erkennen und zu beheben sind, habe ich das Konzept der *action slips* von Norman [1981] auf Fehler in natürlichsprachlichen Texten angewandt und gezeigt, dass viele

[1] http://www.sms4science.ch/ [2] http://www.whatsup-switzerland.ch/ [3] http://www.schreibgebrauch.de/

Abbildung 1: Ausschnitt annotierter Daten mit zwei identifizierten Mehrwortausdrücken.

Word	Pause at end	Pause before	TotalPause2N
Dass	831	110	941
nicht	80	130	190
nur	61	150	211
die	60	3335	3395
das	90	241	331
Ergebnis	221	5247	5468
,	130	70	200
sondern	90	131	221
auch	90	60	150
die	80	8472	8552
gegebene	241	160	401
Leistung	4006	1362	5368
bewertet	411	130	541
wird	490		490
,	141	260	401

der Fehler durch die Verbindung von Prozess und Produkt erklärbar sind [Mahlow 2015b].

Eine erste Auswertung von exemplarischen Schreibprozessdaten zeigt, dass beim Schreiben kurzer argumentativer Essays sehr kurze Pausen zwischen Wörtern – weit unter dem Medianwert – Indikatoren für Mehrwortausdrücke auf verschiedenen Ebenen sind: (a) Diskursphrasen wie *meiner Meinung nach*, *alles in allem*, *oder eben nicht*, (b) Funktionsverbgefüge wie *in Betracht ziehen*, (c) grammatische Konstruktionen wie Modalverbkonstruktionen (*weiter kommen kann*), Passivkonstruktionen (*belohnt werden*), Infinitiv mit zu (*zu dienen*) oder reflexive Verbformen (*sich fragen*) und (d) domänenspezifische Terminologie.

Abbildung 1 zeigt einen Ausschnitt der annotierten Daten einer Schreibsitzung. Im ersten Teil des Satzes schreibt der Autor: *Dass nicht nur die das Ergebnis, sondern auch die gegebene Leistung bewertet wird, ….* Dies enthält das zweiteilige Element *nicht nur …, sondern auch*. Die Zahlen geben Pausenzeit in Millisekunden an; unterhalb des Schwellenwerts der Medianpausenzeit sind sie blau, die Hintergrundfarbe der Zellen kodieren das Spektrum der Pausenlänge (je grüner die Zelle, desto kürzer die Pause). Fett gedruckte Wörter bedeuten, dass die Pause nach diesem Wort und vor dem nächsten Wort (also die Summe der Pausen vor und nach dem Drücken der Leertaste) kürzer als die Medianpause ist. Die Leertaste nach dem Ende eines Mehrwortausdrucks wird jeweils extrem schnell betätigt – die gesamte Einheit ist vollständig als burst produziert worden. Die zweite Einheit schließt auch das Komma ein, der Autor hat offenbar die gesamte Konstruktion inklusive Interpunktion verinnerlicht.

In Mahlow [2015a] habe ich erste Überlegungen zum tatsächlichen inkrementellen Prozessieren der Textproduktion auf syntaktischer Ebene entworfen und Bezüge zwischen dem Versionenbegriff innerhalb der Dokumentverarbeitung und der Digital Humanities hergestellt. Gelingt es, die Textproduktion auf syntaktischer Ebene zu modellieren, lassen sich daraus auch neue Sichtweisen auf die Genese von Texten in der Editorik ableiten.

3 Vorhaben während des Fellowship

Bereits vorhandene Prozessdaten (Deutsch und Griechisch) werden hinsichtlich der Produktion von Mehrwortausdrücken ausgewertet und erste Ergebnisse publiziert. Die in Abschnitt 2.3 gezeigten ersten Beobachtungen sollen an einer genügend grossen Datenmenge verifiziert werden. Während eines Forschungsaufenthalts an der Universität Patras wurden ca. 90 45-minütige Schreibsitzungen mit Inputlog aufgezeichnet. Die Studienteilnehmer haben jeweils ein kurzes argumentatives Essay verfasst. Bislang sind diese Daten noch nicht systematisch ausgewertet worden. In diesen Daten wird die Produktion typischer Diskurselemente ähnlich dem Beispiel in Abbildung 1 analysiert. Die daraus gewonnenen Erkenntnisse sollen in die Konzeption gezielter Schreibexperimente zu kurzen Essays wie auch zum gezielten Überarbeiten solcher Texte in Deutsch genutzt werden.

Als zweiter Aspekt wird evaluiert, welcher Grammatikformalismus aus der Computerlinguistik sich als Grundlage für die syntaktische Analyse von Texten *während*

des Schreibprozesses eignet. Ein geeigneter Formalismus muss zum einen tatsächlich inkrementell – also parallel zur Textproduktion – arbeiten, zum anderen muss er robust sein, um mit ungrammatischen Zwischenständen und Fragmenten umgehen zu können. Insbesondere sollen hierbei *Tree-Adjoining Grammars* (TAG) [beginnend mit Joshi et al. 1975], *Supertagging* [Bangalore und Joshi 1999] und die Arbeiten von Nivre zu inkrementellem Parsing [Nivre 2008] auf ihre Eignung untersucht werden. Lichte und Kallmeyer [2016] konnten zudem zeigen, dass TAG zur Prozessierung von Mehrwortausdrücken exzellent geeignet sind.

4 Methoden

Zur Untersuchung der Mehrwortausdrücke werden zwei Verfahren verwendet:
Zum einen werden potentielle Mehrwortausdrücke entsprechend der in Abbildung 1 gezeigten Annotation ermittelt. Mittels statistischer Verfahren lässt sich berechnen, welche Wortketten besonders schnell produziert werden. Entsprechend meiner Hypothese sind dies Mehrwortausdrücke auf verschiedenen Ebenen. Diese Kandidaten müssen anschliessend manuell gesichtet und klassifiziert werden. Die manuelle Sichtung ist notwendig, da es keine abschliessenden Inventarlisten von Mehrwortausdrücken gibt. Die Klassifizierung soll genutzt werden, um daraus später Muster ableiten und somit automatisch klassifizieren oder nur nach der Produktion bestimmter Mehrwortklassen suchen zu können.
Zum anderen wird in den produzierten Texten automatisch nach potentiellen Mehrwortausdrücken gesucht. Für Deutsch und für Griechisch werde ich hier FipsCo [Seretan und Wehrli 2010] der Universität Genf verwenden können. Auch hier müssen potentielle Kandidaten manuell evaluiert werden.
Anschliessend werden beide Kandidatenlisten automatisch miteinander verglichen, um zwei Fragen zu klären: (a) Sind die in bursts produzierten Mehrwortausdrücken die einzigen, die in den fertigen Texten enthalten sind? (b) Welche Mehrwortausdrücke werden nicht als bursts produziert, welche Erklärungen kann es dafür geben?
Abbildung 2 zeigt exemplarisch die Notwendigkeit echten inkrementellen Parsens. Hierfür wird TAG evaluiert und entsprechende Implementierungen an den vorhandenen deutschen Daten testen und weiterentwickelt.

Abbildung 2: S-Notation Beispiel

```
[Eltern]3|3B[ri]4|4eispielsweise denken Eltern[i]5|5 in Bezug
[su]6|6auf ihr{e}81|82 Kind[er ]7|7{er}82|83 {wohl }8weniger
leistungsorientiert als|8 Lehrträfte.
```

Der Autor beginnt diesen Satz mit *Eltern*, löscht dann dieses Wort sofort und möchte offenbar den Satz ganz anders beginnen, mit *Beispielsweise denken Eltern …* dabei werden Tippfehler sofort bemerkt und korrigiert. Die erste Fassung des Satzes lautet dann: *Beispielsweise denken Eltern in Bezug auf ihr Kind wohl weniger leistungsorientiert als Lehrträfte*. Auffällig ist, dass *wohl* eingefügt wird, bevor der Satz beendet wird (nach *als*). Hier wird die Aussage des Satzes abgeschwächt, noch bevor er vollständig geschrieben ist. Neben dieser semantischen Änderung und ihrem zeitlichen Stattfinden, muss es für eine automatische syntaktische Analyse möglich sein, unvollständige Sätze zu parsen und zudem Elemente – wie hier Partikel – in bereits erzeugte Strukturen einzubinden. TAG scheinen hierfür sehr gut geeignet.
Zu einem viel späteren Zeitpunkt (Index 81) kommt der Autor noch einmal zu diesem Satz und ändert *ihr Kind* in *ihre Kinder* – zu sehen ist hier auch, dass ganz zu Beginn die Form des Substantivs bereits im Plural war, jedoch in Singular geändert wurde (Index 7). Die Gründe für diese Änderung können hier nur vermutet werden, in jedem Fall wird hier eine komplette Nominalphrase unter Beibehaltung der Kongruenz in einem morphosyntaktischen Feature (Numerus) geändert. Auswirkungen auf andere Phrasen bestehen nicht. Es muss für einen Parser also auch möglich sein, bereits erzeugte Strukturen in ihren grammatikalischen Eigenschaften anzupassen und festzustellen, ob und welche «Nebenwirkungen» es gibt.

Literaturverzeichnis

[Baaijen et al. 2012] Veerle M. Baaijen, David Galbraith und Kees de Glopper. Keystroke Analysis. In *Written Communication*, 29(3):246–277, 2012. doi:10.1177/0741088312451108.

[Bangalore und Joshi 1999] Srinivas Bangalore und Aravind K. Joshi. Supertagging: an approach to almost parsing. In *Comput. Linguist.*, 25(2):237–265, 1999.

[Bereiter und Scardamalia 1987] Carl Bereiter und Marlene Scardamalia. *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum, 1987.

[Dürscheid et al. 2010] Christa Dürscheid, Frank Wagner und Sarah Brommer (Hg.) *Wie Jugendliche schreiben: Schreibkompetenz und neue Medien*. Linguistik – Impulse & Tendenzen. De Gruyter, 2010.

[Ehrensberger-Dow und Perrin 2009] Maureen Ehrensberger-Dow und Daniel Perrin. Capturing translation processes to access metalinguistic awareness. In *Across Languages and Cultures*, 10(2):275–288, 2009. doi:10.1556/acr.10.2009.2.6.

[Faigley und Witte 1981] Lester Faigley und Stephen Witte. Analyzing Revision. In *College Composition and Communication*, 32(4):400–414, 1981. doi:10.2307/356602.

[Flinn 1987] Jane Z. Flinn. Case studies of revision aided by keystroke recording and replaying software. In *Computers and Composition*, 5(1):31–44, 1987. doi:10.1016/s8755-4615(87)80013-0.

[Flower und Hayes 1981] Linda S. Flower und John R. Hayes. A Cognitive Process Theory of Writing. In *College Composition and Communication*, 32(4):365–387, 1981. doi:10.2307/356600.

[Hayes 1996] John R. Hayes. A new framework for understanding cognition and affect in writing. In C. Michael Levy und Sarah Ransdell (Hg.) *The Science of Writing. Theories, Methods, Individual Differences and Applications*, 1–27. Hillsdale, NJ, USA: Lawrence Erlbaum, 1996.

[Hayes 2012] John R. Hayes. Modeling and Remodeling Writing. In *Written Communication*, 29(3):369–388, 2012. doi:10.1177/0741088312451260.

[Hobbs 1978] Jerry R. Hobbs. Resolving pronoun references. In *Lingua*, 44(4):311–338, 1978. doi:10.1016/0024-3841(78)90006-2.

[Joshi et al. 1975] Aravind K. Joshi, Leon S. Levy und Masako Takahashi. Tree adjunct grammars. In *Journal of Computer and System Sciences*, 10(1):136–163, 1975. doi:http://dx.doi.org/10.1016/S0022-0000(75)80019-5.

[Kollberg 1998] Py Kollberg. *S-notation – a Computer Based Method for Studying and Representing Text Composition*. Diplomarbeit, Kungliga Tekniska Högskolan Stockholm, 1998.

[Leijten et al. 2012] Mariëlle Leijten, Lieve Macken, Veronique Hoste, Eric Van Horenbeeck und Luuk Van Waes. From Character to Word Level: Enabling the Linguistic Analyses of Inputlog Process Data. In Michael Piotrowski, Cerstin Mahlow und Robert Dale (Hg.) *Proceedings of the Second Workshop on Computational Linguistics and Writing (CL&W 2012): Linguistic and Cognitive Aspects of Document Creation and Document Engineering*, 1–8. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012.

[Lichte und Kallmeyer 2016] Timm Lichte und Laura Kallmeyer. Same syntax, different semantics: A compositional approach to idiomaticity in multi-word expressions. In Christopher Piñón (Hg.) *Empirical Issues in Syntax and Semantics 11*, 111–140. 2016.

[Lindgren 2005] Eva Lindgren. *Writing and revising: Didactic and Methodological Implications of Keystroke Logging*. Dissertation, Umeå Universitet, 2005.

[Mahlow 2011] Cerstin Mahlow. *Linguistisch unterstütztes Redigieren: Konzept und exemplarische Umsetzung basierend auf interaktiven computerlinguistischen Ressourcen*. Dissertation, University of Zurich, 2011.

[Mahlow 2015a] Cerstin Mahlow. A Definition of "Version"for Text Production Data and Natural Language Document Drafts. In *Proceedings of the 3rd International Workshop on (Document) Changes: Modeling, Detection, Storage and Visualization*, DChanges 2015, 27–32. New York, NY, USA: ACM, 2015. doi:10.1145/2881631.2881638.

[Mahlow 2015b] Cerstin Mahlow. Learning from Errors: Systematic Analysis of Complex Writing Errors for Improving Writing Technology. In Núria Gala, Reinhard Rapp und Gemma Bel-Enguix (Hg.) *Language Production, Cognition, and the Lexicon*, Bd. 48 von *Text, Speech and Language Technology*, 419–438. Springer International Publishing, 2015. doi:10.1007/978-3-319-08043-7_24.

[Mahlow und Piotrowski 2008] Cerstin Mahlow und Michael Piotrowski. Linguistic Support for Revising and Editing. In Alexander Gelbukh (Hg.) *Computational Linguistics and Intelligent Text Processing: 9th International Conference, CICLing 2008, Haifa, Israel, February 17–23, 2008. Proceedings*, Bd. 4919 von *Lecture Notes in Computer Science*, 631–642. Berlin, Heidelberg, New York: Springer, 2008. doi:10.1007/978-3-540-78135-6_54.

[Mahlow und Piotrowski 2009] Cerstin Mahlow und Michael Piotrowski. LingURed: Language-Aware Editing Functions Based on NLP Resources. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, Bd. 4, 243–250. Polish Information Processing Society, 2009.

[Mann und Thompson 1983] William C. Mann und Sandra A. Thompson. Relational Propositions in Discourse. Technical Report ISI/RR-83-115, Information Sciences Institute, 1983.

[Nivre 2008] Joakim Nivre. Algorithms for Deterministic Incremental Dependency Parsing. In *Computational Linguistics*, 34(4):513–553, 2008. doi:10.1162/coli.07-056-R1-07-027.

[Norman 1981] Donald A. Norman. Categorization of action slips. In *Psychological Review*, 88:1–15, 1981.

[Perrin 2002] Daniel Perrin. Progression Analysis (PA): Investigating writing strategies in the workplace. In Thierry Olive und C. Michael Levy (Hg.) *Contemporary Tools and Techniques for Studying Writing*, Bd. 10 von *Studies in Writing*, 105–117. Boston, Dordrecht, London: Kluwer, 2002.

[Perrin und Wildi 2009] Daniel Perrin und Marc Wildi. Statistical Modeling of Writing Processes. In Charles Bazerman, Robert Krut, Karen Lunsford, Susan McLeod, Suzie Null, Paul Rogers und Amanda Stansell (Hg.) *Traditions of Writing Research*, 378–393. New York, NY, USA: Routledge, 2009.

[Seretan und Wehrli 2010] Violeta Seretan und Eric Wehrli. Tools for syntactic concordancing. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, 493–500. IEEE, 2010.

[Severinson Eklundh 1994] Kerstin Severinson Eklundh. Linear and nonlinear strategies in computer-based writing. In *Computers and Composition*, 11(3):203–216, 1994. doi:10.1016/8755-4615(94)90013-2.

[Sommers 1980] Nancy Sommers. Revision Strategies of Student Writers and Experienced Adult Writers. In *College Composition and Communication*, 31(4):378–388, 1980. doi:10.2307/356588.

[Soricut und Marcu 2003] Radu Soricut und Daniel Marcu. Sentence Level Discourse Parsing Using Syntactic and Lexical Information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, 149–156. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003. doi:10.3115/1073445.1073475.

[Storrer 2014] Angelika Storrer. Sprachverfall durch internetbasierte Kommunikation? Linguistische Erklärungsansätze - empirische Befunde. In Institut für Deutsche Sprache (Hg.) *Sprachverfall? Dynamik - Wandel - Variation. Jahrbuch des Instituts für Deutsche Sprache 2013*, 171–196. Berlin, New York: De Gruyter, 2014.

[Will et al. 2006] Udo Will, Guido Nottbusch und Rudiger Weingarten. Linguistic units in word typing: Effects of word presentation modes and typing delay. In *Written Language and Literacy*, 9(1):153–176, 2006. doi:10.1075/wll.9.1.10wil.