Praxis des kreativen Standardisierens

Dr. Melanie Siegel, Dr. Sabine Lehmann, acrolinx GmbH, Berlin

Es gibt Standards für technische Dokumentation, wie der DocCert Anforderungskatalog des TÜV-Süd oder der QualiAssistent der tekom. Einerseits lassen diese jedoch Spielräume. So wird bei DocCert z.B. nicht festgelegt, ob eine Handlungsanweisung in direkter oder indirekter Ansprache des Lesers erfolgen soll, aber es wird Konsistenz gefordert. Andererseits passen nicht alle Richtlinien eines Standards zu jedem Unternehmen und erfordern Variation. Schließlich gibt es branchenspezifische, ja sogar unternehmensspezifische Richtlinien für die technische Dokumentation. So müssen Regeln nach Zielgruppeneigenschaften und Erwartungen, aber auch nach branchenüblichen Mustern ausgewählt werden. Wichtig dabei ist jedoch eine konsistente Anwendung, so dass es nicht zu Brüchen in der Dokumentation kommt. Dabei helfen automatische Prüfwerkzeuge.

Die Entwicklung eines individuellen Standards „vom grünen Tisch“ führt selten zu zufriedenstellenden Ergebnissen. Bei der automatischen Prüfung stellt man schnell fest, dass die „ausgedachten“ Regeln einer systematischen Anwendung nicht standhalten. Bei der Implementierung solcher Richtlinien stellt man fest, dass sie oft zu wenig konkret formuliert sind, wie z.B. „formulieren Sie Handlungsanweisungen knapp und präzise“.

Wie jedoch kann ein Standard entwickelt werden, der zu einem Unternehmen, seiner Branche und Zielgruppen passt und für die automatische Prüfung implementiert werden kann? Sprachtechnologie hilft effizient bei der Entwicklung individueller Richtlinien. Durch Datenanalyse, Satzcluster und Parametrisierung entsteht ein textspezifischer individueller Standard. Ist damit aber der Gegensatz von Kreativität und Standardisierung aufgehoben?

Systematische Datenanalyse mit sprachtechnologischen Methoden

Die Basis einer systematischen Datenanalyse bildet typisches Textmaterial, also technische Dokumentation des Unternehmens. Dieses Material wird zunächst mit linguistischer Information angereichert, wobei computerlinguistische Methoden angewandt werden. Diese Methoden stammen aus Tokenisierung, Morphologieanalyse und Terminologie.

Oft existieren in einem Unternehmen bereits auch Richtlinien für die technische Redaktion sowie Terminologie oder Wörterlisten. Die Richtlinien werden auf existierende Prüfregeln zur automatischen Prüfung abgebildet. Voraussetzung dafür ist eine große Datenbank von Regeln. Terminologie oder Wörterlisten werden in die Terminologiekomponente importiert.

Auf dieser Basis werden sogenannte „Satzcluster“ automatisch erstellt. Satzcluster sind Gruppen von Sätzen oder Phrasen mit gemeinsamem oder ähnlichem Inhalt, aber unterschiedlicher Struktur. Ein Beispiel für ein Satzcluster, das wir aus technischen Dokumentationen extrahiert haben, ist:

- Gehen Sie dazu wie folgt vor
- Gehen Sie dazu folgendermaßen vor
- So gehen Sie vor
- Bitte gehen Sie wie folgt vor
- Gehen Sie dabei folgendermaßen vor
- .. gehen Sie wie folgt vor

Die Zusammenstellung dieser Satzcluster ist nicht nur ein Vergleich von Zeichenketten, sondern basiert auf der annotierten linguistischen Information und der importierten Information des Unternehmens. So werden Lemmata und Phrasen verglichen, aber auch von

Zahlen abstrahiert und gegensätzliche Terminologie eingebunden. Dazu kommt die Einbeziehung von Antonymen und Synonymen, so dass z. B. Sätze mit den Verben „einschalten“ und „ausschalten“ nicht in einem Cluster stehen, aber Sätze mit den Nomen „PC“ und „Computer“ durchaus.

Die Satzcluster geben wichtige Hinweise darauf, welche Variationen in den Texten des Unternehmens tatsächlich vorhanden sind und auf welche Bereiche sich die Entwicklung von Regeln für die automatische Prüfung konzentrieren muss. In einigen Fällen wird das Ergebnis der Analyse von Satzclustern sein, dass eine Standardformulierung gewählt wird, die dann in neuen Texten vorgeschlagen wird, wenn ähnliche Formulierungen genannt werden. Auf der Basis der importierten Regeln und Terminologie sowie der Grammatik- und Rechtschreibprüfung wird bereits eine Variante als Standardformulierung vorgeschlagen. Der Redakteur wird jetzt die Cluster durchsehen und validieren. Aber auch in den Fällen, in denen eine Standardformulierung nicht sinnvoll ist, bieten die Satzcluster eine gute Möglichkeit, neue Stilrichtlinien zu formulieren oder existierende zu evaluieren.

Die technische Redation wird nach der Analyse der Satzcluster in der Lage sein, präzise formulierte Regeln mit dazu passenden Beispielen zusammenzustellen, die direkt auf automatische Prüfregeln abgebildet werden können.

Systematische Terminologieanalyse mit sprachtechnologischen Methoden

Auf der Basis der linguistisch annotierten Textdaten und der importierten Terminologie wird in einem weiteren Schritt Terminologie extrahiert, analysiert und validiert. Morphologische Information wird genutzt, um die Terme auf Lemmata zurückzuführen, so dass nicht alle deklinierten Formen eines Worts betrachtet werden müssen. Dazu gehört auch eine automatische Kompositaanalyse. Existierende Terminologie wird eingebunden, um diese nicht noch einmal zu extrahieren. Zu der reinen Extraktion kommt auch hier ein Clustering-Prozess. Dabei werden ähnliche Wörter zusammengestellt. Ein Beispiel aus Texten der technischen Dokumentation:

- Schraubverbindung
- Schraubenverbindung
- Schraub-Verbindung

Automatisch angewendete Termbildungsregeln geben Hinweise für die Validierung, wie z. B. „Kompositum aus mehr als drei Komponenten“ oder „Kompositum mit drei Konsonanten“. Für die Validierung selbst ist die technische Redakteurin gefragt, sie bekommt effiziente computerlinguistische Unterstützung. Auf diese Weise entsteht eine konsistente Terminologie, die zum Unternehmen und den im Unternehmen produzierten Texten passt und die direkt in die automatische Prüfung importiert werden kann.

Schluss

Generelle Standards für die technische Dokumentation sind notwendig und allgemein akzeptiert. Sie müssen für die Implementierung in automatischen Prüfprogrammen konkretisiert werden. Weiterhin müssen sie an die Erfordernisse des Unternehmens angepasst werden. Dafür steht eine Datenanalyse der Unternehmensdaten auf Basis von computerlinguistischen Verfahren und unter Einbeziehung der im Unternehmen vorhandenen Quellen zur Verfügung. Die Datenanalyse liefert so wichtige Informationen für die Aufstellung spezifischer Stilregeln und einer zum Unternehmen passenden Terminologie.

Für Rückfragen: sabine.lehmann@acrolinx.de, melanie.siegel@acrolinx.de