Computerfonts, Mediävistik und Unicode

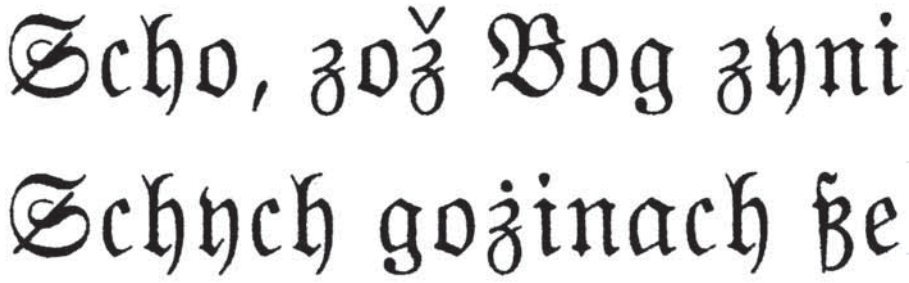


Abb. 2:
Das „gestrichene S“
in einem Druck des 19. Jh.s

Wie seltenes Sprachgut den Weg ins

von Sebastian Kempgen

Was hat Bamberg mit dem „großen Eszett“ und neuen Computertastaturen zu tun? Und wie kommen die Sorben zu den Zeichen, die sie zur Digitalisierung ihrer Bibeln brauchen? Ein grundlegendes Werkzeug, das alle nutzen, die sich mit Sprachen beschäftigen, sind Computerfonts, die die benötigten Zeichen enthalten. Der Artikel erklärt einige Zusammenhänge.

Die „Europäische Charta der Regional- oder Minderheitensprachen“ in der „zwischen Deutschland, Österreich und der Schweiz abgestimmten Fassung mit Fußnoten“ setzt sich dafür ein, dass Regional- und Minderheitensprachen als Ausdruck des kulturellen Reichtums anerkannt werden. Als weitere Ziele und Grundsätze der gemeinsamen Politik legt sie unter anderem fest, ihren Gebrauch in Wort und Schrift sowie das Lehren und Lernen dieser Sprachen sowie das Studium und die Erforschung dieser Sprachen an den Universitäten zu fördern. Außerdem möchte sie anregen alles dafür zu tun, dass Behinderungen in der Ausübung dieser Sprachen unterbleiben.

Auch wenn die Europäische Charta in erster Linie auf die Gegenwart abzielt, so gehören zum Studium und zur Erforschung solcher Regionalsprachen neben deren Geschichte auch alte und ausgestorbene Sprachen, ihre Schriften und ihr Schrifttum. Eine grundlegende Voraussetzung, um dieses kulturelle Erbe heute nutzbar zu machen, ist seine Digitalisierung. Dies bedeutet aber nicht nur, dass Faksimiles alter Texte eingescannt und online bereitgestellt werden. Der Computer muss auch die Buchstaben aller Schriften und Sprachen „kennen“, um überall in diesen und über diese kommunizieren zu können – auf einem Smartphone genauso wie auf einem Tablet-PC oder einem Laptop. Die universelle Zeichenbasis, die eben dieses leistet, heißt Unicode und wird seit zwei Jahrzehnten kontinuierlich ausgebaut. In dieser Zeit ist das erfasste Zeichenrepertoire von anfangs etwas unter 30.000 Zeichen zu jetzt mehr als 100.000 Zeichen (Buchstaben) gewachsen. Mit der Kurzbezeichnung „Unicode“ meint man in der Regel diesen Schlüssel zur eindeutigen Wiedergabe von Zeichen. Dahinter steckt aber eine gleichnamige Organisation, die sich um diesen offenen Standard kümmert, ihn weiter-

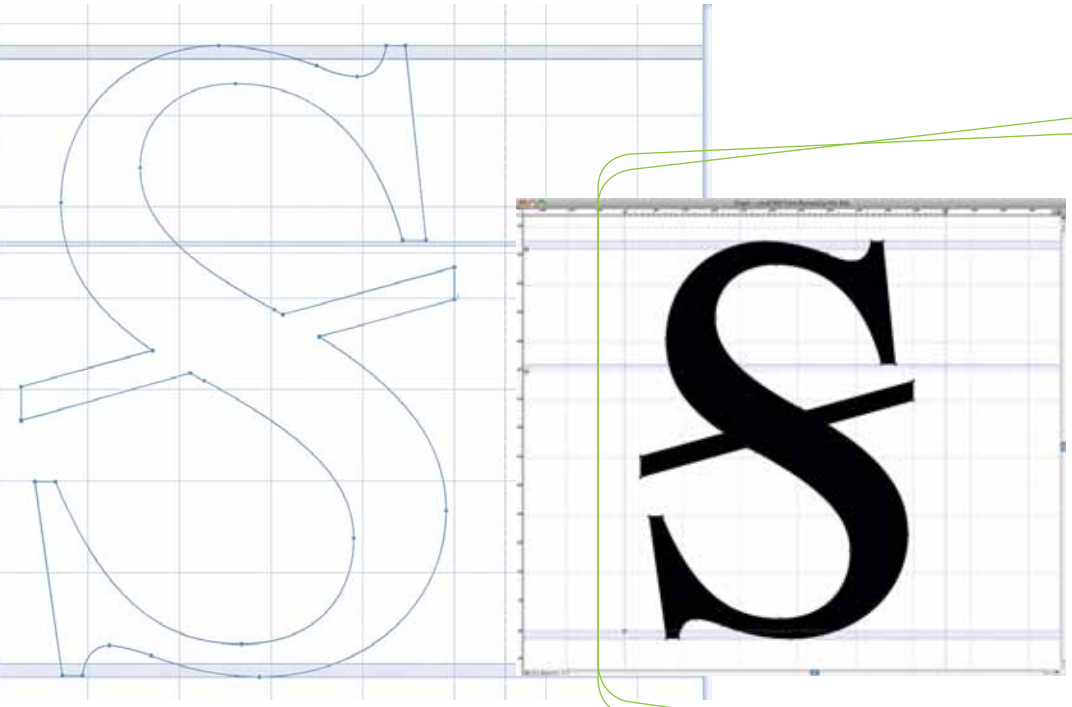


Abb. 4:
Das „gestrichene S“ in einem Font-Editor

digitale Zeitalter findet

entwickelt, seine Spezifikationen veröffentlicht usw. Im Februar 2011 wurde die komplette Dokumentation zu Unicode 6.0.0 veröffentlicht (vgl. www.unicode.org). In diesem Code ist noch reichlich Platzreserve vorhanden, um alle noch nicht erfassten Schriften und Zeichen aufzunehmen: Knapp 1 Mio. Slots stehen insgesamt zur Verfügung.

Wer a sagt, muss auch b sagen

Ein Computer-Nutzer hat normalerweise nicht direkt mit Unicode zu tun, aber er spürt seine Segnungen. Was nämlich auf einem Gerät als „a“ eingetippt wird, das soll auf jedem beliebigen anderen Gerät auch wieder als „a“ angezeigt werden, und nicht als „b“. Das gewährleisten die auf Computern genutzten Schriften, wenn sie sogenannte „Unicode-Fonts“ sind, die sich an diesen Standard halten. Ein „a“ mag dann in jeder Schrift ein wenig anders aussehen, aber es bleibt doch immer ein „a“.

Dass grundlegende Alphabete wie das englische von Anfang an in Unicode berücksichtigt waren, ist klar. Je exotischer die Schrift (aus europäischer Sicht), je älter das Schrifttum oder je spezieller ein Zeichen, desto eher kommt es vor, dass bestimmte Zeichen in Unicode noch nicht vorhanden sind. Deshalb gibt es einen festgelegten Weg, wie neue Zeichen beantragt und aufgenommen werden. Der Vorschlag erfolgt in Form von „proposals“, die einer bestimmten Form und vor allem gewissen Kriterien genügen müssen.

Was ist ein Buchstabe?

Damit beispielsweise ein Zeichen neu in Unicode aufgenommen werden kann, muss man seine frühere oder aktuelle Existenz belegen, zum Beispiel mit Scans aus einschlägigen Handschriften, frühen Drucken, Grammatiken und dergleichen. Ferner ist klarzustellen, dass auch heute noch mindestens zwei Personen oder eine „academic community“ ein Interesse daran haben, dieses Zeichen zu verwenden und nicht Einzelpersonen ausschließlich ihre persönlichen Bedürfnisse artikulieren. Vor allem aber muss ein bestimmtes Zeichen oder ein Buchstabe eine eigene Funktion haben, die es beziehungsweise ihn von anderen, schon vorhandenen Zeichen oder Buchstaben unterscheidet.

Diese Anforderung kann knifflig sein, macht aber Sinn: Unicode ist nicht dafür da, die individuellen handschriftlichen oder grafischen Gestaltungsformen der Buchstaben mit unterschiedlichen

A7A8	Ꞩ	LATIN CAPITAL LETTER S WITH OBLIQUE STROKE
A7A9	ꞩ	LATIN SMALL LETTER S WITH OBLIQUE STROKE • also used in pre-1950 Lower Sorbian orthography → 1E9C ẜ latin small letter long s with diagonal stroke

Abb. 3:
Das „gestrichene S“ in der Unicode-Dokumentation

Codes zu versehen und so zu unterscheiden. Als „Glyphen“ in Unicode werden nur abstrakte Einheiten codiert, nicht deren konkrete Vorkommen oder Realisierungen.

Im Prinzip kann jeder zur Weiterentwicklung von Unicode beitragen; praktischerweise bilden sich Arbeitsgruppen, die sich der fachwissenschaftlichen und technischen Aspekte der zu schreibenden Anträge annehmen. Um unnütze Arbeit zu vermeiden, wird man dabei immer auch ein Auge auf die „Pipeline“ von Unicode werfen, die auflistet, welche anderen Anträge gerade schon gestellt worden sind und in welchem Antragsstadium sie sich befinden.

Nicht alles hat eine Chance – Kooperation gegen Chaos

Nicht jedes Zeichen oder Symbol hat eine Chance, in Unicode aufgenommen zu werden. Doch auch für dieses Problem gibt es eine Lösung, nämlich die sogenannte „Private Use Area“ (PUA), ein Vorrat von mehr als 130.000 Slots, die Schriftdesigner auf eigenes Risiko füllen können: nicht jeder Computer und nicht jede Schrift kennt dann allerdings dieses Zeichen; einige Fonts werden es haben, andere nicht. Möglichkeiten zur Realisierung bestehen beispielsweise in Form von Firmenlogos. Auch die Uni Bamberg hat etwas Entsprechendes in ihrer Geschichte, nämlich das alte zweizeilige „UNI BA“-Logo oder das noch ältere „UB“-Logo. Der Charakter der „Private Use Area“ bringt es mit sich, dass verschiedene Personen den gleichen Slot für unterschiedliche Zwecke nutzen können.

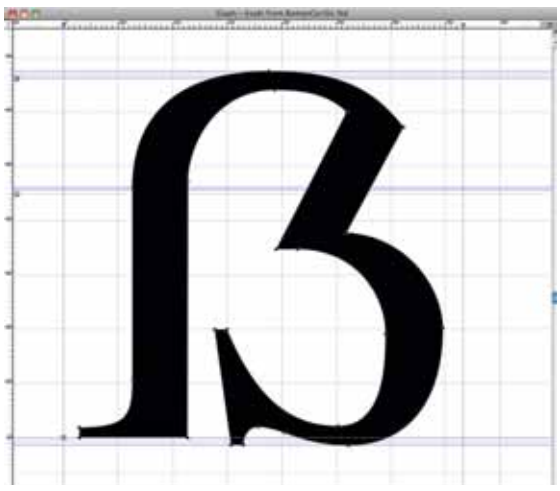


Abb. 5:
Das große Eszett in einem Font-Editor

Um wenigstens im Bereich der Wissenschaft ein Chaos in der Nutzung der PUA zu vermeiden und gemeinsam schneller zu einer Erweiterung des Standards zu kommen, haben sich etliche User zusammengetan. Im Bereich der Sprach- und Literaturwissenschaften gibt es seit 2001 die „Medieval Unicode Font Initiative“ (www.mufi.info, s. Abb. 1), in der ursprünglich vor allem Germanisten, Romanisten, Anglisten und Klassische Philologen aktiv waren. In der Slavistik gibt es seit 1999 eine internationale „Commission for Computer Processing of Manuscripts and Early Printed Books“, die sich der gleichen Problematik annimmt. Sie organisiert Konferenzen zum Thema, tritt auf dem Internationalen Slavistenkongress mit Vorträgen auf oder publiziert ihre Ergebnisse regelmäßig in der Zeitschrift „Scripta & e-Scripta“ (Sofia). Seit Kurzem arbeiten Slavisten und MUFI-Vertreter zusammen, um die jeweilige Nutzung der Private Use Area untereinander zu koordinieren. Beispielsweise enthält der aktuelle Vorschlag der Slavistik 150 Zeichen für die Private Use Area.

Minderheitensprache in Deutschland – das Sorbische

Ein aktuelles Beispiel für eine Erweiterung des Unicode-Standards betrifft unter anderem eine der Minderheitensprachen in Deutschland, nämlich das Sorbische – genau genommen das Obersorbische und das schon fast ausgestorbene Niedersorbische. Das Niedersorbische kannte, vor allem in den Bibeldrucken des 19. Jahrhunderts, ein „gestrichenes S“ (siehe Abb. 2). Das aber fehlte bislang in Unicode,

Computer fonts, medieval literature and Unicode

How a wealth of rare language is finding its way into the digital age

What does Bamberg have to do with a new German computer keyboard's "capital ss" ("großes Eszett"-ß)? And how do the Sorbs access the characters they require for the digitalization of their bibles? Computer fonts incorporating the necessary characters are the vital tools used by anyone seriously engaged in languages. This article explains some key correlations.

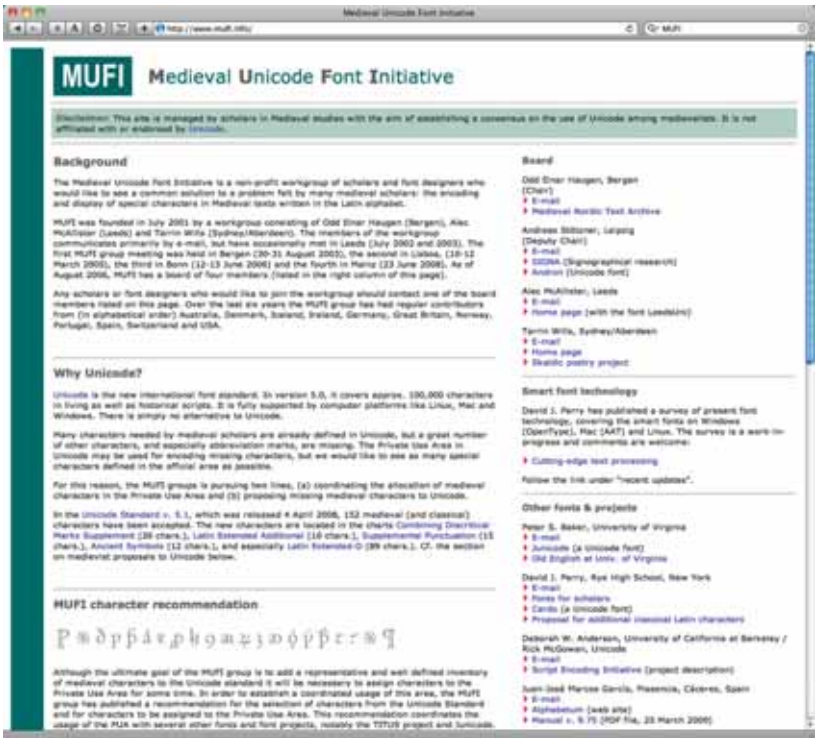
was bedeutete, dass Bücher in dieser Sprache nicht vollständig korrekt im Unicode-Standard digital veröffentlicht werden konnten. Der vor Kurzem verabschiedete Unicode-Standard 6.0. enthält nun unter anderem das „gestrichene S" und zwar als Großbuchstaben wie als Kleinbuchstaben (vgl. Abb. 3). Damit ist jetzt das letzte noch fehlende Buchstabenpaar in Unicode vorhanden, das zur Schreibung der modernen wie der historischen Orthographie des Sorbischen benötigt wurde. Seine „Code-Points" lauten, wie die Abbildung 3 zeigt, A7A8 und A7A9 oder U+A7A8 bzw. U+A7A9. Computer arbeiten intern nur mit diesen Hexadezimal-Codes, der normale Nutzer hingegen sieht die Implementierung eines Zeichens mit diesem Hexadezimalcode in einem Computerfont.

Von der Theorie zur Praxis – Implementierung in Computerfonts

Der schönste Standard aber nützt nichts, wenn er nicht in die Praxis umgesetzt wird. Das bedeutet in diesem Falle: Schriftdesigner müssen die Neuerungen aus den aktuellen Unicode-Versionen in ihre Fonts aufnehmen, denn nur so werden sie auf Computern darstellbar. Die Rolle des Autors als Mitglied der bereits erwähnten „Commission for Computer Processing of Manuscripts and Early Printed Books" ist es, mit spezieller Software Referenzfonts herzustellen, die zur Darstellung der neuen Zeichen genutzt werden können. Dieser Aufgabe dient der seit 1996 im Uni-Netz hängende „Kodeks"-Webserver. Hier gibt es unter anderen Schriften zum Download – jetzt auch mit Unterstützung von Unicode 6.0. Die vom Autor entworfene Schrift Roman-Cyrillic Std wird übrigens auch vom deutschen DIN-Institut in Berlin genutzt, um über ein neues deutsches Tastaturlayout zu informieren: Mit der Aufnahme eines „großen Eszett" in den Unicode-Standard brauchen deutsche Computer-Tastaturen künftig eine andere Beschriftung und womöglich auch ein etwas anderes Tasten-Layout. RomanCyrillic Std enthält auch ein solches „großes Eszett" – und ist dazu noch kostenlos.

http://kodeks.uni-bamberg.de/AKSL/Schrift/RomanCyrillicStd.htm

Abb. 1:
Die MUFI-Webseite



Bibliographische Angaben:
Bibliographical Entry:

Sebastian Kempgen: Computerfonts, Mediävistik und Unicode: Wie seltenes Sprachgut den Weg ins digitale Zeitalter findet. In: *uni.vers Forschung*, Universität Bamberg 2011, 24-27.